

Livestock Breeding and Genomics

Peter von Rohr

2020-12-04

Contents

Preface	7
Prerequisites	7
Goals	7
Exam	8
Lecture Notes	8
1 Introduction	9
1.1 Terminology	9
1.2 History	10
1.3 Fundamental Questions	11
1.4 Improvement of Animal Populations	13
1.5 Statistics	14
1.6 Computer Science	14
2 Basics in Quantitative Genetics	17
2.1 Single Locus - Quantitative Trait	17
2.2 Frequencies	18
2.3 Hardy-Weinberg Equilibrium	19
2.4 Value and Mean	19
2.5 Variances	26
2.6 Extension To More Loci	28
2.7 Genetic Models	29
2.8 Appendix: Derivations	30
3 Genetic Evaluations	33
3.1 Introduction	33
3.2 Basic Principle of Predicting Breeding Values	37
3.3 Animal's Own Performance	37
3.4 Progeny Records	43
4 Best Linear Unbiased Prediction (BLUP)	47
4.1 Problem of Correction	47
4.2 Numeric Example	49

4.3	Linear Mixed Effects Model	50
4.4	Sire Model	53
4.5	Animal Model	54
5	Genetic Covariances Between Animals	55
5.1	Similarity Between Individuals	55
5.2	Numerator Relationship Matrix	58
5.3	The Inverse Numerator Relationship Matrix	60
5.4	Structure of A^{-1}	60
5.5	Henderson's Rule To Set Up A^{-1}	61
5.6	Derivation of Henderson's Rules	62
5.7	Computing Inbreeding Coefficients	66
6	Variance and Inbreeding	69
6.1	Inbreeding	69
6.2	Changes of Mean Value	74
6.3	Changes of Variance	75
7	Additional Aspects of BLUP	77
7.1	Accuracy	77
7.2	Confidence Intervals of Predicted Breeding Values	79
7.3	Relevance of Accuracies	80
7.4	Decomposition of Predicted Breeding Value	81
8	Multiple Traits	85
8.1	Multivariate Predictions Of Breeding Values Using BLUP	86
8.2	Multitrait Selection	88
8.3	Review On Selection Index Theory	92
9	Genomic Selection	103
9.1	Background	103
9.2	A Linear Model To Predict Genomic Breeding Values	106
9.3	GBLUP	108
9.4	Practical Problems	108
9.5	How Does GBLUP Work	109
10	Variance Components	111
10.1	Sire Model	111
10.2	Analysis Of Variance (Anova)	111
10.3	Numerical Example	112
10.4	Negative Estimates with Anova	114
10.5	Likelihood-Based Approaches	114
10.6	Summary	116
11	Genome-Wide Association Studies (GWAS)	117
11.1	Single Marker Regression Tests	117
11.2	Genome-Wide Association Experiments Using Haplotypes	120

<i>CONTENTS</i>	5
-----------------	---

11.3 Fitting All Markers Simultaneously	121
A Introduction To Linear Algebra	123
A.1 Glimpse Ahead	123
A.2 Vectors	123
A.3 Matrices	129
A.4 Systems Of Equations	132
A.5 Solving Systems of Linear Equations	134
B Introduction To R and RStudio	135

Preface

Livestock Breeding and Genomics is a masters course taught in the Institute of Agricultural Sciences at ETH Zurich. The aim of this course is to introduce the basic concepts of livestock breeding and genomics to the students.

Prerequisites

There are no explicit prerequisites for this course. All required concepts are explained during this course. Although basic knowledge in the following areas is definitely helpful for this course.

- **Linear Algebra:** Basic properties of and basic operations with vectors and matrices
- **Statistics:** Simple linear models and least squares
- **R:** Basics of how to work with data and how to do simple computations

Goals

The following goals are taken from the teaching system at ETH. The students are able to set up design matrices, the additive genetic relationship matrix and its inverse as well as the mixed model equations to estimate BLUP breeding values. The concept of BLUP breeding values are also extended to genomic selection. All concepts introduced during this course are not only presented during the lecture, but students are expected to work with the concepts on small example problems.

Specific Learning Goals

The students ...

- ... understand the basic concepts of livestock breeding and genomics.
- ... can separate the concepts of breeding and genomics from the techniques that are important in a livestock production enterprise.
- ... are familiar with the basic tools used in livestock breeding and genomics.
- ... know for which problems these tools can be applied.

- ... understand the genetic foundation behind the theory on which livestock breeding and genomics is based on.
- ... are able to extend the genetic theory into the statistical models used in livestock breeding and genomics.
- ... can extract the practical meaning of the results of the statistical analyses.

Exam

There will be a written exam during the lecture of the last week of the semester. The exam is going to be an open-book exam. All material that seems helpful to students will be allowed. The students must solve the exams on their own. The exam is not designed to be solved in teams.

Lecture Notes

These lecture notes contain all the relevant material for the exam. The appendix of the course notes contains two chapters, which introduce the necessary tools to work through the material of this course. Those tools consist of linear algebra and R. These chapters are mostly left to students to work through by themselves. Different students have different levels of skills in the topics of linear algebra and R. Hence not every student needs to go through the presented material at the same depth.

Chapter 1

Introduction

1.1 Terminology

The terms

- **Livestock Breeding** and
- **Animal Breeding**

are used interchangeably in this course. Although, the latter is a more general term, whereas the former focuses on livestock species, i.e. animals which are typically present on a farm such as cattle, pig, goat and sheep. Animal breeding in general could also include pets such as dogs or cats or even zoo animals. But the very interesting topic of breeding such species is outside of the scope of this lecture and is therefore not covered in these course notes.

The term **livestock breeding** is sometimes understood ambiguously. In general, most people do not differentiate between livestock breeding and animal husbandry or animal production. From a scientific point of view a **Livestock Breeder** is a person who owns a number of animals from which he or she selects parent animals and uses a designed mating scheme to achieve a certain goal with the offspring animals. Most commonly known are breeders of pet animals such as dogs, cats or birds which follow individual breeding goals which focus on specific phenotypic appearance or on special behavioral traits. Around the end of the 19th century, livestock breeders have realized that they have to work together in breeding associations to be able to effectively select parent animals from a large breeding population. This allowed them to achieve more robust selection responses in shorter amounts of time, especially for livestock species with long generation intervals such as cattle and horse. This development is nicely documented by the archive of Swiss agricultural history available at <https://www.histoiredurale.ch/afaahr/>. In summary, livestock breeders are primarily interested in selecting parent animals with the best genetic potential

to produce offspring animals that are closer to a breeding goal. From an economic point of view the sale of breeding animals to other farmers makes an important contribution to the economic result of the farm.

In contrast to livestock breeding, livestock production focuses on the aspects of keeping animals on farms to produce goods that can be sold. Nowadays, the goods produced by farm animals are mostly used in human nutrition. Due to the focus on the production aspects, the economic result of the production process is determined by the difference between monetary revenue for the products and the costs that are caused by the production of the goods.

Depending on the livestock species, the separation between breeding and production is more pronounced. In pigs, most farms are specialized into either breeding farms or production farms. Most cattle farms run as mixed operations which means that they are members of breeding organizations but they also run a production business. While the mixture of both operation types (breeding and production) on the same farm is not negative, it is conceptually important to separate breeding and production.

1.2 History

Livestock breeding and Genomics are two scientific areas which have seen some quite dramatic changes in the last few years. As already mentioned in the previous section, livestock breeding started over 100 years ago and is a much older discipline than Genomics.

1.2.1 Livestock Breeding

In principle, livestock breeding did exist for a very long time in a rather unsystematic form. Individual breeders always made choices about which animals they want to select as parents for the next generation of their livestock herds. Accounts that such early activities in livestock breeding happened as early as the Middle Ages are given in (Duerst, 1931) and (Arndorfer et al., 2010). But to the best of my knowledge it was only in the second half of the 20th century that the area of livestock breeding made some ground-breaking progress which spread all over the world. This progress was initiated by the researcher **Charles R. Henderson**. He and his team developed a solid methodology that is still applied up to the current day. The main achievement of Henderson and his team was to find a class of statistical models that are consistent with the theory of quantitative genetics described in (Falconer and Mackay, 1996) which is one reference among many other sources. Furthermore, the research groups lead by Henderson showed how to efficiently compute the results from the statistical models for large datasets.

1.2.2 Genomics

Genomics started with the Human Genome Project. The publication of the first draft of the complete Human genome (Venter et al., 2001) and a publication by (Meuwissen et al., 2001a) that appeared in the same year made it possible to include information of complete genomes into statistical analyses. The process of including information from complete genomes into statistical analyses is what is understood by the term **genomics**.

The introduction of genomics methodologies in the area of livestock breeding caused a shift of paradigm. In large livestock breeding populations associations between certain genetic variants and the expression of desirable forms of phenotypic traits can be estimated using older breeding animals. The obtained estimation results can be used to assess the genetic potential of young animals which do not have any phenotypic observations available. This type of analysis is part of a procedure which is termed **genomic selection** and it allows for selecting breeding animals at a much younger age which shortens the generation interval.

The basic principle of how animals are selected as parents of future generations did not change, but the availability of different types of information and the amount of information that can be used to assess the genetic potential of a selection candidate changed dramatically since the invention of genomic technologies. Despite these rapid developments of new technologies, livestock breeders are still facing the following two fundamental questions.

1.3 Fundamental Questions

In livestock breeding and genomics, we are interested in addressing two fundamental questions that bothered breeders for a very long time. For this course, we put these two fundamental questions into the following form.

1. What is the *best* animal?
2. What can breeders do to obtain the *best* animal?

The term **best** is relative, because there is no **best** animal for all situations and all environments. Animals that show high performances in one environment, may not be able to produce as much in a different environment. One example for that might be Holstein cows in Europe or North America are able to produce a lot of milk, but they have difficulties to survive in Africa. Knowing that the environment plays an important role for livestock animals, we will be assuming that the animals that we are selecting, are more or less adapted to their environment.

Animals are usually described or characterized in terms of appearance or performance or a combination of both. In any case, we will be talking about **traits** where any trait is an observable or measurable characteristic of an animal. Examples of *observable* traits are

- coat color
- size
- muscling
- leg set
- udder conformation and many more.

Observable traits are mostly used to describe the appearance of an animal. In contrast to that, *measurable* traits are mostly used to describe the performance of an animal. Examples of measurable traits are

- body weight
- milk production
- protein and fat yield.

Note, it is important to distinguish between the observed or measured values of a trait which might be **red** coat color or 343 kg of body weight and the traits themselves which are just coat color or body weight. The observed or measured values of a trait are also called **phenotypes**.

1.3.1 Genotypes and Phenotypes

In livestock breeding we are mainly concerned with changing animal populations at the genetic level. The reason why we are interested in changing a population genetically is because parents do not pass their phenotypes to their offspring. Parents pass a random sample of their genes to their offspring. For each offspring every parent does transmit a different sample of their genes. From a genetic point of view, we want to know not only the most desirable phenotype, but also the most desirable **genotypes**. From the central dogma of molecular biology (https://en.wikipedia.org/wiki/Central_dogma_of_molecular_biology), it follows that an animal's genotype provides the genetic background of phenotypes. The relationship between phenotypes (P) and genotypes (G) can be summarized by the following equation (1.1)

$$P = G + E \tag{1.1}$$

where E represents the **environmental effects**. Because we want to change our populations at a genetic level, we are interested in the effect (G) of every genotype. In most cases, we are not able to directly observe or measure G . But we will see later on how we can estimate G based on measurements and observations of P and based on estimates of E . The estimates of G will later be called **breeding values** and those estimates will be used by breeders as information for their tools to improve animal populations. Those tools are being described in the following section.

1.4 Improvement of Animal Populations

The purpose of livestock breeding is to improve animal populations. Once an animal is conceived, the genotype is fixed¹ and cannot be improved anymore. Breeders can improve populations at the genetic level using the following two tools

1. selection
2. mating

1.4.1 Selection

Selection is the process to determine which individuals of a current population become parents of the next generation. The application of selection in a certain population over a certain time changes the animals in that population at the genetic level. The most familiar form of selection is **natural selection** which occurs in natural and wildlife populations. Natural selection is one of the great forces of evolution and it also affects domestic animals. All animals with lethal genetic defects are naturally selected against, i.e., they never become parents.

Although natural selection cannot be ignored for livestock species, what is most important for animal breeders is **artificial selection**. The idea behind artificial selection is simple. For a given trait all animals in a population are ranked according to their breeding value. From this list the animals ranking top are used as parents for the next generation. In most livestock populations, animal breeders are interested to improve their animals with respect to more than just one trait. When considering more than one trait, the question is how to come up with the ranking for the animals that are selected as potential parents. Several strategies to produce such a ranking based on a number of traits. It has been shown that using a weighted mean of the breeding values of all traits which is called **aggregate genotype** to rank all animals is an optimal procedure to be used as selection criterion (Hazel, 1943).

1.4.2 Mating

The second tool we have available as animal breeders is **mating**. In a mating scheme, we decide which of the selected male animals are bred to which selected female animals. There are a number of different rules that can be followed. The application of a given set of rules are summarized as mating system. There are three reasons for using a mating system.

1. producing offspring with extreme breeding values. When parents with extreme breeding values (high or low) are mated, offspring with extreme phenotypes can be expected. This is mostly used when a given trait is to be changed in one direction

¹Here we do not take into account new technologies such as gene editing.

2. make use of complementarity in parental traits. When neither of the parents is optimal, a mix of traits can be desirable. In such a case parental genotypes can be quite different. When parents of different breeds are mated, then this is called **crossbreeding**.
3. obtain positive effects due to heterosis. Hybrid vigor or heterosis in crossbreeding occurs when offspring performance exceeds the performance of the pure-breds.

There might also be other aspects that influence a mating system, e.g. to restrict the level of inbreeding or to consider optimum genetic contribution theory (Meuwissen and Sonesson, 1997).

1.5 Statistics

Several authors such as (Schaeffer, 2013) and (Gianola and Rosa, 2015) have reviewed the development of statistical methods in the area of animal breeding. Both authors mention that statistical methodology plays an important role in animal breeding. Most animal breeders are concerned with estimating or predicting breeding values. This is still done using a set of methods resulted from the theory developed by Charles Henderson and his team ((Henderson, 1953) and (Henderson, 1975)). These methods are known under the name of **BLUP**. BLUP shows some important regularization properties. These properties allow us to estimate or to predict many more unknown parameters than we have observations. In animal breeding, breeding values of all animals in a population can be predicted also for those animals for which we do not have observations. This is particularly important for traits which can only be observed in animals of one sex.

There are more methods with regularization properties. The so-called **Bayesian methods** are one example. Bayesian methods use the so-called Bayes theorem ((Bayes and Price, 1763) and https://en.wikipedia.org/wiki/Bayes%27_theorem) to come up with parameter estimates. Although Bayesian methods are much older than other methods such as BLUP, they were only introduced into practical animals breeding in the early 1990's. Important pioneering papers for the use of Bayesian methods in animal breeding are (Gianola and Foulley, 1982) and (Gianola and Fernando, 1986). The reasons for the late adoption of Bayesian methods are certainly related to development of cheap computing infrastructure. This is described in more detail in subsection 1.6.

1.6 Computer Science

The development of computing power is often summarized by the so-called **Moore's Law** ((Moore, 1965) and https://en.wikipedia.org/wiki/Moore%27s_law). Moore's law is not a law in the sense of mathematics or physics, but it

is a prediction that Gordon Moore² made as early as 1965. He predicted that the number of components that could be placed on a certain integrated circuit would double roughly every 18 months between 1959 and 1975. This prediction was generalized into a statement that the general computing performance could be doubled every 18 months. In retro-spect this was more or less true for the last 50 years. This considerable increase in computing performance had also a dramatic impact in the costs of a certain computation.

When comparing the development of computing performance with the performance of livestock animals, there is an obvious relation between the two. This means the performance increase of livestock animals was in part facilitated by the development of cheap computing power. The two figures 1.1 and 1.2 compare the two developments. The first diagram shows the annual milk production per cow.

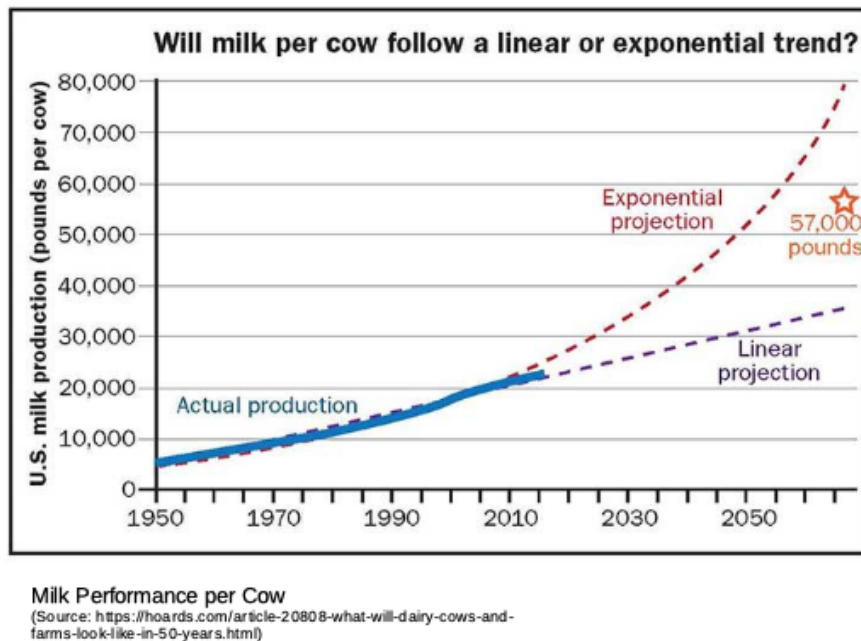


Figure 1.1: Yearly Milk Yield per Cow in the USA

The Figure 1.2 below shows the development of computing power according to Moore's law.

²One of the co-founders and a director of Intel



Chapter 2

Basics in Quantitative Genetics

As already mentioned in section 1.3.1, the central dogma of molecular biology tells us that the genotype is the basics of any phenotypic expression. The genotype of an individual is composed of a number of genes which are also called **loci**. In this section, we start with the simplest possible genetic architecture where the genotype is composed by just one locus. The connection between the genotype and the phenotype is modeled according to equation (1.1). The phenotype is assumed to be a quantitative trait. That means we are not looking at binary or categorical traits. Categorical traits can just take a limited number of different levels. Examples of categorical traits are the horn status in cattle or certain color characteristics. Quantitative traits do not take discrete levels but they show specific distributions.

2.1 Single Locus - Quantitative Trait

In Livestock there are not many examples where a quantitative trait is influenced by just one locus. But this case helps in understanding the foundation of more complex genetic architectures. We start by looking at the following idealized population (Figure 2.1).

2.1.1 Terminology

The different genetic variants that are present at our Locus G are called **alleles**. When looking at all individuals in the population for our locus, we have two different alleles G_1 and G_2 . Hence, we call the locus G to be a **bi-allelic** locus. In any given individual of the population, the two alleles of the locus G together are called the individuals **genotype**. All possible combinations of the two alleles

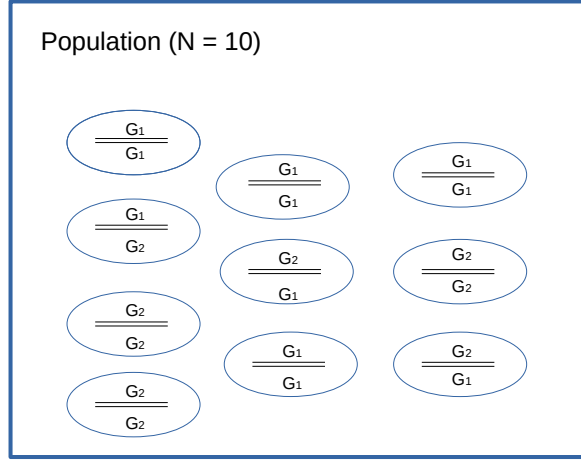


Figure 2.1: Idealized Population With A Single Locus

at the locus G leads to a total number of three genotypes. It is important to mention that the order of the alleles in a given genotype is not important. Hence, G_1G_2 and G_2G_1 are the same genotype. The two genotypes G_1G_1 and G_2G_2 are called **homozygous** and the genotype G_1G_2 is called **heterozygous**.

2.2 Frequencies

To be able to characterize our population with respect to the locus of interest, we are first looking at some frequencies. These are measures of how often a certain allele or genotype does occur in our population. For our example population shown in Figure 2.1, the **genotype frequencies** are

$$\begin{aligned}
 f(G_1G_1) &= \frac{4}{10} = 0.4 \\
 f(G_1G_2) &= \frac{3}{10} = 0.3 \\
 f(G_2G_2) &= \frac{3}{10} = 0.3
 \end{aligned} \tag{2.1}$$

The **allele frequencies** can be determined either by counting or they can be computed from the genotype frequencies.

Table 2.1: Genotype Frequencies under Hardy-Weinberg equilibrium

Alleles	G_1	G_2
G_1	$f(G_1G_1) = p^2$	$f(G_1G_2) = p * q$
G_2	$f(G_1G_2) = p * q$	$f(G_2G_2) = q^2$

$$\begin{aligned}
 f(G_1) &= f(G_1G_1) + \frac{1}{2} * f(G_1G_2) = 0.55 \\
 f(G_2) &= f(G_2G_2) + \frac{1}{2} * f(G_1G_2) = 0.45
 \end{aligned}
 \tag{2.2}$$

2.3 Hardy-Weinberg Equilibrium

The Hardy-Weinberg equilibrium is the central law of how allele frequencies and genotype frequencies are related in an idealized population. Given the allele frequencies

$$\begin{aligned}
 f(G_1) &= p \\
 f(G_2) &= q = 1 - p
 \end{aligned}
 \tag{2.3}$$

During mating, we assume that in an idealized population alleles are combined independently. This leads to the genotype frequencies shown in Table 2.1.

Summing up the heterozygous frequencies leads to

$$\begin{aligned}
 f(G_1G_1) &= p^2 \\
 f(G_1G_2) &= 2pq \\
 f(G_2G_2) &= q^2
 \end{aligned}
 \tag{2.4}$$

Comparing these expected genotype frequencies in a idealized population under the Hardy-Weinberg equilibrium to what we found for the small example population in Figure 2.1, we can clearly say that the small example population is not in Hardy-Weinberg equilibrium.

2.4 Value and Mean

Our goal is still to improve our population at the genetic level. The term improvement implies the need for a quantitative assessment of our trait of interest. Furthermore, we have to be able to associate the genotypes in the population to the quantitative values of our trait.

Table 2.2: Values for all Genotypes

Variable	Genotype	Values
V_{11}	G_1G_1	a
V_{12}	G_1G_2	d
V_{22}	G_2G_2	-a

2.4.1 Genotypic Values

The values V_{ij} to each genotype G_iG_j are assigned as shown in Figure 2.2.

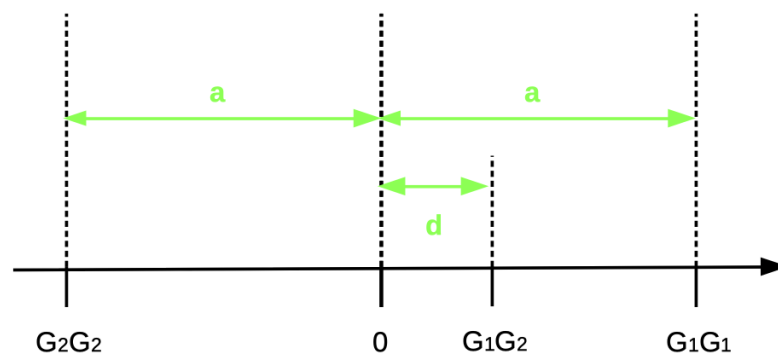


Figure 2.2: Genotypic Values

The origin of the genotypic values is placed in the middle between the two homozygous genotypes G_2G_2 and G_1G_1 . Here we are assuming that G_1 is the favorable allele. This leads to values of $+a$ for genotype G_1G_1 and of $-a$ for genotype G_2G_2 . The value of genotype G_1G_2 is set to d and is called dominance deviation. Table 2.2 summarizes the values for all genotypes.

2.4.2 Population Mean

For the complete population, we can compute the **population mean** (μ) of all values at the locus G . This mean corresponds to the expected value and is computed as

$$\begin{aligned}\mu &= V_{11} * f(G_1G_1) + V_{12} * f(G_1G_2) + V_{22} * f(G_2G_2) \\ &= a * p^2 + d * 2pq + (-a) * q^2 \\ &= (p - q)a + 2pqd\end{aligned}\tag{2.5}$$

The population mean depends on the values a and d and on the allele frequencies p and q . The larger the difference between p and q the more influence the value a has in μ , because for very different p and q the product $2pq$ is very small. On the other hand, if $p = q = 0.5$, then $\mu = 0.5d$. For loci with $d = 0$, the population mean $\mu = (p - q)a$ and hence, if in addition we have $p = q$, then $\mu = 0$.

2.4.3 Breeding Values

The term **breeding value** is defined as shown in Definition 2.1.

Definition 2.1 (Breeding Value). The breeding value of an animal i is defined as two times the difference between the mean value of offsprings of animal i and the population mean.

Applying this definition and using the parameters that we have computed so far leads to the following formulas for the breeding value of an animal with a certain genotype.

2.4.3.1 Breeding value for G_1G_1

Assume that we have a given parent S with a genotype G_1G_1 and we want to compute its breeding value. Let us further suppose that our single parent S is mated to a potentially infinite number of animals from the idealized population, then we can deduce the following mean genotypic value for the offspring of parent S .

	Mates of S	
	$f(G_1) = p$	$f(G_2) = q$
Parent S		
$f(G_1) = 1$	$f(G_1G_1) = p$	$f(G_1G_2) = q$

Because parent S has genotype G_1G_1 , the frequency $f(G_1)$ of a G_1 allele coming from S is 1 and the frequency $f(G_2)$ of a G_2 allele is 0. The expected genetic value (μ_{11}) of the offspring of animal S can be computed as

$$\mu_{11} = p * a + q * d \quad (2.6)$$

Applying definition 2.1, we can compute the breeding value (BV_{11}) for animal S as shown in equation (2.7) while using the results given by equations (2.6) and (2.5).

$$\begin{aligned} BV_{11} &= 2 * (\mu_{11} - \mu) \\ &= 2 (pa + qd - [(p - q)a + 2pqd]) \\ &= 2 (pa + qd - (p - q)a - 2pqd) \\ &= 2 (qd + qa - 2pqd) \\ &= 2 (qa + qd(1 - 2p)) \\ &= 2q (a + d(1 - 2p)) \\ &= 2q (a + (q - p)d) \end{aligned} \quad (2.7)$$

Breeding values for parents with genotypes G_2G_2 and G_1G_2 are derived analogously.

2.4.3.2 Breeding value for G_2G_2

First, we determine the expected genotypic value for offsprings of a parent S with genotype G_2G_2

	Mates of parent S	
	$f(G_1) = p$	$f(G_2) = q$
Parent S		
$f(G_2) = 1$	$f(G_1G_2) = p$	$f(G_2G_2) = q$

The expected genetic value (μ_{22}) of the offspring of animal S can be computed as

$$\mu_{22} = pd - qa \quad (2.8)$$

The breeding value BV_{22} corresponds to

$$\begin{aligned}
BV_{22} &= 2 * (\mu_{22} - \mu) \\
&= 2 (pd - qa - [(p - q)a + 2pqd]) \\
&= 2 (pd - qa - (p - q)a - 2pqd) \\
&= 2 (pd - pa - 2pqd) \\
&= 2 (-pa + p(1 - 2q)d) \\
&= -2p (a + (q - p)d)
\end{aligned} \tag{2.9}$$

2.4.3.3 Breeding value for G_1G_2

The genotype frequencies of the offsprings of a parent S with a genotype G_1G_2 is determined in the following table.

	Mates of parent S	
	$f(G_1) = p$	$f(G_2) = q$
Parent S		
$f(G_1) = 0.5$	$f(G_1G_1) = 0.5p$	$f(G_1G_2) = 0.5q$
$f(G_2) = 0.5$	$f(G_1G_2) = 0.5p$	$f(G_2G_2) = 0.5q$

The expected mean genotypic value of the offsprings of parent S with genotype G_1G_2 is computed as

$$\mu_{12} = 0.5pa + 0.5d - 0.5qa = 0.5 [(p - q)a + d] \tag{2.10}$$

The breeding value BV_{12} corresponds to

$$\begin{aligned}
ZW_{12} &= 2 * (\mu_{12} - \mu) \\
&= 2 (0.5(p - q)a + 0.5d - [(p - q)a + 2pqd]) \\
&= 2 (0.5pa - 0.5qa + 0.5d - pa + qa - 2pqd) \\
&= 2 (0.5(q - p)a + (0.5 - 2pq)d) \\
&= (q - p)a + (1 - 4pq)d \\
&= (q - p)a + (p^2 + 2pq + q^2 - 4pq)d \\
&= (q - p)a + (p^2 - 2pq + q^2)d \\
&= (q - p)a + (q - p)^2d \\
&= (q - p) [a + (q - p)d]
\end{aligned} \tag{2.11}$$

2.4.3.4 Summary of Breeding Values

The term $a + (q - p)d$ appears in all three breeding values. We replace this term by α and summarize the results in the following table.

Genotype	Breeding Value
G_1G_1	$2q\alpha$
G_1G_2	$(q - p)\alpha$
G_2G_2	$-2p\alpha$

2.4.4 Allele Substitution

Comparing the genotype G_2G_2 with the genotype G_1G_2 , one of the differences is in the number of G_1 -alleles. G_2G_2 has zero G_1 -alleles and G_1G_2 has one G_1 -allele. They also have different breeding values. Because the breeding values are to be used to assess the value of a given genotype, any difference between the breeding values BV_{12} and B_{22} can be associated to that additional G_1 allele in G_1G_2 compared to G_2G_2 .

The computation of the difference between the breeding value BV_{12} and B_{22} results in

$$\begin{aligned}
 BV_{12} - BV_{22} &= (q - p)\alpha - (-2p\alpha) \\
 &= (q - p)\alpha + 2p\alpha \\
 &= (q - p + 2p)\alpha \\
 &= (q + p)\alpha \\
 &= \alpha
 \end{aligned} \tag{2.12}$$

The analogous computation can be done by comparing the breeding values BV_{11} and BV_{12} .

$$\begin{aligned}
 BV_{11} - BV_{12} &= 2q\alpha - (q - p)\alpha \\
 &= (2q - (q - p))\alpha \\
 &= \alpha
 \end{aligned} \tag{2.13}$$

The breeding values themselves depend on the allele frequencies. But the differences between breeding values are linear while replacing G_2 alleles by G_1 alleles. This replacement is also called allele-substitution and the term *alpha* is called **allele-substitution effect**.

2.4.5 Dominance Deviation

When looking at the difference between the genotypic value V_{ij} and the breeding value BV_{ij} for each of the three genotypes, we get the following results.

$$\begin{aligned}
V_{11} - BV_{11} &= a - 2q\alpha \\
&= a - 2q[a + (q - p)d] \\
&= a - 2qa - 2q(q - p)d \\
&= a(1 - 2q) - 2q^2d + 2pqd \\
&= [(p - q)a + 2pqd] - 2q^2d \\
&= \mu + D_{11}
\end{aligned} \tag{2.14}$$

$$\begin{aligned}
V_{12} - BV_{12} &= d - (q - p)\alpha \\
&= d - (q - p)[a + (q - p)d] \\
&= [(p - q)a + 2pqd] + 2pqd \\
&= \mu + D_{12}
\end{aligned} \tag{2.15}$$

$$\begin{aligned}
V_{22} - BV_{22} &= -a - (-2p\alpha) \\
&= -a + 2p[a + (q - p)d] \\
&= [(p - q)a + 2pqd] - 2p^2d \\
&= \mu + D_{22}
\end{aligned}$$

The difference all contain the population mean μ plus a certain deviation. This deviation term is called **dominance deviation**.

2.4.6 Summary of Values

The following table summarizes all genotypic values all breeding values and the dominance deviations.

Genotyp	genotypic value	Breeding Value	Dominance Deviation
G_iG_j	V_{ij}	BV_{ij}	D_{ij}
G_1G_1	a	$2q\alpha$	$-2q^2d$
G_1G_2	d	$(q - p)\alpha$	$2pqd$
G_2G_2	$-a$	$-2p\alpha$	$-2p^2d$

The formulas in the above shown table assume that G_1 is the favorable allele with frequency $f(G_1) = p$. The allele frequency of G_2 is $f(G_2) = q$. Since we have a bi-allelic locus $p + q = 1$.

Based on the definition of dominance deviation, the genotypic values V_{ij} can be decomposed into the components population mean (μ), breeding value (BV_{ij}) and dominance deviation (D_{ij}) according to equation (2.16).

$$V_{ij} = \mu + BV_{ij} + D_{ij} \quad (2.16)$$

Taking expected values on both sides of equation (2.16) and knowing that the population mean μ was defined as the expected value of the genotypic values in the population, i.e. $E[V] = \mu$, it follows that the expected values of both the breeding values and the dominance deviations must be 0. More formally, we have

$$\begin{aligned} E[V] &= E[\mu + BV + D] \\ &= E[\mu] + E[BV] + E[D] \\ &= \mu \end{aligned} \quad (2.17)$$

From the last line in equation (2.17), it follows that $E[BV] = E[D] = 0$. This also shows that both breeding values and dominance deviations are defined as deviation from a given mean.

2.5 Variances

The population mean μ and derived from that the breeding values were defined as expected values. Their main purpose is to assess the state of a given population with respect to a certain genetic locus and its effect on a phenotypic trait of interest. One of our primary goals in livestock breeding is to improve the populations at the genetic level through the means of selection and mating. Selection of potential parents that produce offspring that are closer to our breeding goals is only possible, if the selection candidates show a certain level of variation in the traits that we are interested in. In populations where there is no variation which means that all individuals are exactly at the same level, it is not possible to select potential parents for the next generation.

In statistics the measure that is most often used to assess variation in a certain population is called **variance**. For any given discrete random variable X the variance is defined as the second central moment of X which is computed as shown in equation (2.18).

$$Var[X] = \sum_{x_i \in \mathcal{X}} (x_i - \mu_X)^2 * f(x_i) \quad (2.18)$$

where	\mathcal{X} :	set of all possible x -values
	$f(x_i)$	probability that x assumes the value of x_i
	μ_X	expected value $E[X]$ of X

In this section we will be focusing on separating the obtained variances into different components according to their causative sources. Applying the definition of variance given in equation (2.18) to the genotypic values V_{ij} , we obtain the following expression.

$$\begin{aligned}\sigma_G^2 = \text{Var}[V] &= (V_{11} - \mu)^2 * f(G_1G_1) \\ &+ (V_{12} - \mu)^2 * f(G_1G_2) \\ &+ (V_{22} - \mu)^2 * f(G_2G_2)\end{aligned}\quad (2.19)$$

where $\mu = (p - q)a + 2pqd$ the population mean.

Based on the decomposition of the genotypic value V_{ij} given in (2.16), the difference between V_{ij} and μ can be written as the sum of the breeding value and the dominance deviation. Then σ_G^2 can be written as

$$\begin{aligned}\sigma_G^2 = \text{Var}[V] &= (BV_{11} + D_{11})^2 * f(G_1G_1) \\ &+ (BV_{12} + D_{12})^2 * f(G_1G_2) \\ &+ (BV_{22} + D_{22})^2 * f(G_2G_2)\end{aligned}\quad (2.20)$$

Inserting the expressions for the breeding values BV_{ij} and for the dominance deviation D_{ij} found earlier and simplifying the equation leads to the result in (2.21). A more detailed derivation of σ_G^2 is given in the appendix (2.8) of this chapter.

$$\begin{aligned}\sigma_G^2 &= 2pq\alpha^2 + (2pqd)^2 \\ &= \sigma_A^2 + \sigma_D^2\end{aligned}\quad (2.21)$$

The formula in equation (2.21) shows that σ_G^2 consists of two components. The first component σ_A^2 is called the **genetic additive variance** and the second component σ_D^2 is termed **dominance variance**. As shown in equation (2.25) σ_A^2 corresponds to the variance of the breeding values. Because we have already seen that the breeding values are additive in the number of favorable alleles, σ_A^2 is called genetic additive variance. Because σ_D^2 corresponds to the variance of the dominance deviation effects (see equation (2.27)) it is called dominance variance.

2.6 Extension To More Loci

When only a single locus is considered, the genotypic values (V_{ij}) can be decomposed according to equation (2.16) into population mean, breeding value and dominance deviation. When a genotype refers to more than one locus, the genotypic value may contain an additional deviation caused by non-additive combination effects.

2.6.1 Epistatic Interaction

Let V_A be the genotypic value of locus A and V_B denote the genotypic value of a second locus B , then the total aggregate genotypic value V attributed to both loci A and B can be written as

$$V = V_A + V_B + I_{AB} \quad (2.22)$$

where I_{AB} is the deviation from additive combination of these genotypic values. When computing the population mean earlier in this chapter, we assumed that I was zero for all combinations of genotypes. If I is not zero for any combination of genes at different loci, those genes are said to **interact** with each other or to exhibit **epistasis**. The deviation I is called interaction deviation or epistatic deviation. If I is zero, the genes are called to act additively between loci. Hence *additive action* may mean different things. When referring to one locus, it means absence of dominance. When referring to different loci, it means absence of epistasis.

Interaction between loci may occur between pairs or between higher numbers of different loci. The complex nature of higher order interactions, i.e., interactions between higher number of loci does not need to concern us. Because in the aggregate genotypic value V , interaction deviations of all sorts are treated together in an overall interaction deviation I . This leads to the following generalized form of decomposing the overall aggregate genotype V for the case of multiple loci affecting a certain trait of interest.

$$V = \mu + U + D + I \quad (2.23)$$

where U is the sum of the breeding values attributable to the separate loci and D is the sum of all dominance deviations. For our purposes in livestock breeding where we want to assess the genetic potential of a selection candidate to be a parent of offspring forming the next generation, the **breeding value** is the most important quantity. The breeding value is of primary importance because a given parent passes a random sample of its alleles to its offspring. We have seen in section 2.4.4 that breeding values are additive in the number of favorable alleles. Hence the more favorable alleles a given parent passes to its offspring the higher the breeding value of this parent.

On the other hand, the dominance deviation measures the effect of a certain genotype occurring in an individual and the interaction deviation estimates the effects of combining different genotypes at different loci in the genome. But because parents do not pass complete genotypes nor do they pass stretches of DNA with several loci, but only a random collection of its alleles, it is really the breeding value that is of primary importance in assessing the genetic potential of a given selection candidate.

2.6.2 Interaction Variance

If genotypes at different loci show epistatic interaction effects as described in section 2.6.1, the interactions give rise to an additional variance component called V_I , which is the variance of interaction deviations. This new variance component V_I can be further decomposed into sub-components. The first level of sub-components is according to the number of loci that are considered. Two-way interactions involve two loci, three-way interactions consider three loci and in general n -way interactions arise from n different loci. The next level of subdivision is according to whether they include additive effects, dominance deviations or both.

In general it can be said that for practical purposes, interaction effects explain only a very small amount of the overall variation. As already mentioned in section @ref(#epistatic-interaction) for livestock breeding, we are mostly interested in the additive effects. This is also true when looking at the variance components. Hence dominance variance and variances of interaction deviations are not used very often in practical livestock breeding application.

2.7 Genetic Models

In this chapter, we have seen how to model the genetic basis of a quantitative trait when a single locus affects the trait of interest. We call this a single-locus model. When several loci have an effect on a certain trait, then we talk about a **polygenic model**. Letting the number of loci affecting a certain phenotype tend to infinity, the resulting model is called **infinitesimal model**.

From a statistical point of view, the breeding values in an infinitesimal model are considered as a random effect with a known distribution. Due to the central limit theorem, this distribution is assumed to be a normal distribution. The central limit theorem says that the distribution of any sum of a large number of very small effects converges to a normal distribution. For our case where a given trait of interest is thought to be influenced by a large number of genetic loci each having a small effect, the sum of the breeding values of all loci together can be approximated by a normal distribution. Figure (2.3) shows the distribution for a sum of 10, 100 and 1000 components each. The histograms show a better approximation to the normal distribution the larger the number of components considered in the sum.

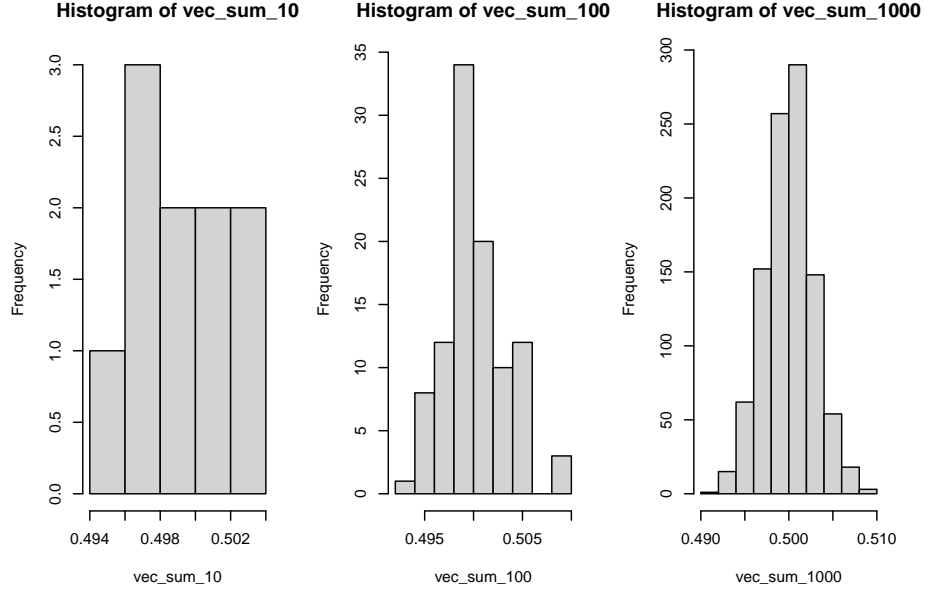


Figure 2.3: Distribution of Sums of Different Numbers of Components

2.7.1 Model Usage In Routine Evaluations

Traditional prediction of breeding values before the introduction of genomic selection is based on the infinitesimal model. When genomic selection was introduced which takes into account the information from a large number of loci, genomic breeding values are estimated using a polygenic model.

2.8 Appendix: Derivations

This section shows how the genetic variance in equation (2.21) is computed.

$$\begin{aligned}
\sigma_G^2 &= (BV_{11} + D_{11})^2 * p^2 \\
&+ (BV_{12} + D_{12})^2 * 2pq \\
&+ (BV_{22} + D_{22})^2 * q^2 \\
&= (2q\alpha - 2q^2d)^2 * p^2 \\
&+ ((q - p)\alpha + 2pqd)^2 * 2pq \\
&+ (-2p\alpha - 2p^2d)^2 * q^2 \\
&= (4q^2\alpha^2 - 8q^3d\alpha + 4q^4d^2) * p^2 \\
&+ (q^2\alpha^2 - 2pq\alpha^2 + p^2\alpha^2 - 4(q - p)pqd\alpha + 4p^2q^2d^2) * 2pq \\
&+ (4p^2\alpha^2 + 8p^3d\alpha + 4p^4\alpha^2) * q^2 \\
&= 4p^2q^2\alpha^2 - 8p^2q^3d\alpha + 4p^2q^4d^2 \\
&+ 2pq^3\alpha^2 - 4p^2q^2\alpha^2 + 2p^3q\alpha^2 \\
&- 8p^3q^2d\alpha + 8p^2q^3d\alpha + 8p^3q^3d^2 \\
&+ 4p^2q^2\alpha^2 + 8p^3q^2d\alpha + 4p^4q^2d^2 \\
&= 4p^2q^2\alpha^2 + 4p^2q^4d^2 \\
&+ 2pq^3\alpha^2 + 2p^3q\alpha^2 \\
&+ 8p^3q^3d^2 \\
&+ 4p^4q^2d^2 \\
&= 2pq\alpha^2 (p^2 + 2pq + q^2) \\
&+ (2pqd)^2 (p^2 + 2pq + q^2) \\
&= 2pq\alpha^2 + (2pqd)^2 \\
&= \sigma_A^2 + \sigma_D^2
\end{aligned} \tag{2.24}$$

From the last two lines of (2.24) it follows that $\sigma_A^2 = 2pq\alpha^2$ and $\sigma_D^2 = (2pqd)^2$. It can be shown that σ_A^2 corresponds to the squared breeding values times the associated genotype frequencies. Because the expected values of the breeding values is zero, σ_A^2 is equivalent to the variance of the breeding values.

$$\begin{aligned}
\sigma_A^2 &= \text{Var}[BV] = (BV_{11} - E[BV])^2 * f(G_1G_1) \\
&\quad + (BV_{12} - E[BV])^2 * f(G_1G_2) \\
&\quad + (BV_{22} - E[BV])^2 * f(G_2G_2) \\
&= BV_{11}^2 * f(G_1G_1) + BV_{12}^2 * f(G_1G_2) + BV_{22}^2 * f(G_2G_2) \\
&= (2q\alpha)^2 * p^2 + ((q-p)\alpha)^2 * 2pq + (-2p\alpha)^2 * q^2 \\
&= 4p^2q^2\alpha^2 + (q^2\alpha^2 - 2pq\alpha^2 + p^2\alpha^2) * 2pq + 4p^2q^2\alpha^2 \\
&= 8p^2q^2\alpha^2 + 2pq^3\alpha^2 - 4p^2q^2\alpha^2 + 2p^3q\alpha^2 \\
&= 4p^2q^2\alpha^2 + 2pq^3\alpha^2 + 2p^3q\alpha^2 \\
&= 2pq\alpha^2 (2pq + q^2 + p^2) \\
&= 2pq\alpha^2
\end{aligned} \tag{2.25}$$

In the above derivation in (2.25) of the variance of the breeding values, we were using the fact that the expected value $E[BV] = 0$. This can be shown more formally as follows

$$\begin{aligned}
E[BV] &= BV_{11} * f(G_1G_1) + BV_{12} * f(G_1G_2) + BV_{22} * f(G_2G_2) \\
&= 2q\alpha * p^2 + (q-p)\alpha * 2pq + (-2p\alpha) * q^2 \\
&= 2p^2q\alpha + 2pq^2\alpha - 2p^2q\alpha - 2pq^2\alpha \\
&= 0
\end{aligned} \tag{2.26}$$

Similarly to (2.25) we can show that σ_D^2 corresponds to the squared dominance deviations times the frequencies of the corresponding genotypes. That is the reason why σ_D^2 is called dominance variance.

$$\begin{aligned}
\sigma_D^2 &= D_{11}^2 * f(G_1G_1) + D_{12}^2 * f(G_1G_2) + D_{22}^2 * f(G_2G_2) \\
&= (-2q^2d)^2 * p^2 + (2pqd)^2 * 2pq + (-2p^2d)^2 * q^2 \\
&= 4p^2q^4d^2 + 8p^3q^3d^2 + 4p^4q^2d^2 \\
&= 4p^2q^2d^2 (q^2 + 2pq + p^2) \\
&= 4p^2q^2d^2
\end{aligned} \tag{2.27}$$

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 setting¹ 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.

¹That 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.

$$y_{ij} = \mu_i + g_i + e_{ij} \quad (3.1)$$

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.

$$y_{ij} = \mu_i + u_i + e_{ij}^* \quad (3.2)$$

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.

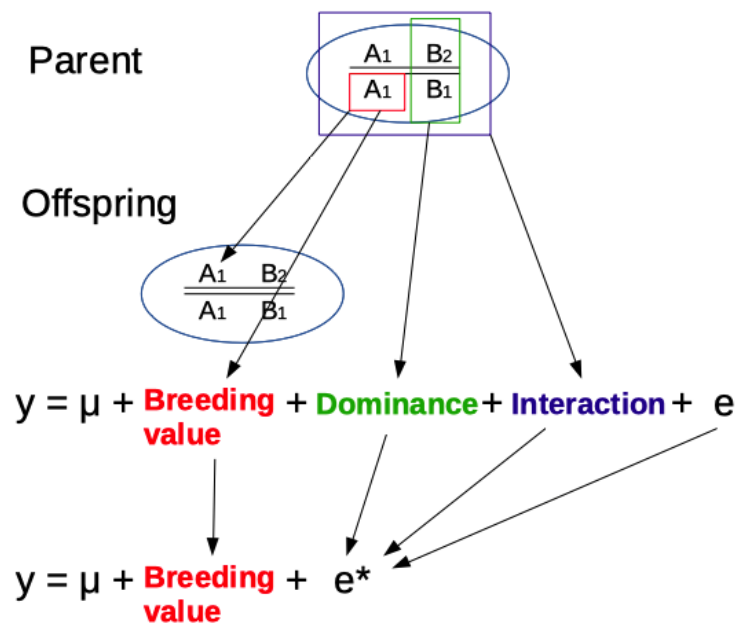


Figure 3.1: Re-arrangement of Terms Representing Genetic Effects

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 $\text{var}(y_{ij})$, $\text{var}(u_i)$ and $\text{var}(e_{ij}^*)$ are all known. Covariances ($\text{cov}(u_i, e_{ij}^*)$) between genetic and environmental effects and covariances ($\text{cov}(e_{ij}^*, e_{kl}^*)$) 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

$$u_i = \frac{1}{2}u_s + \frac{1}{2}u_d + m_i \quad (3.3)$$

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-weight 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 \hat{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 \hat{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

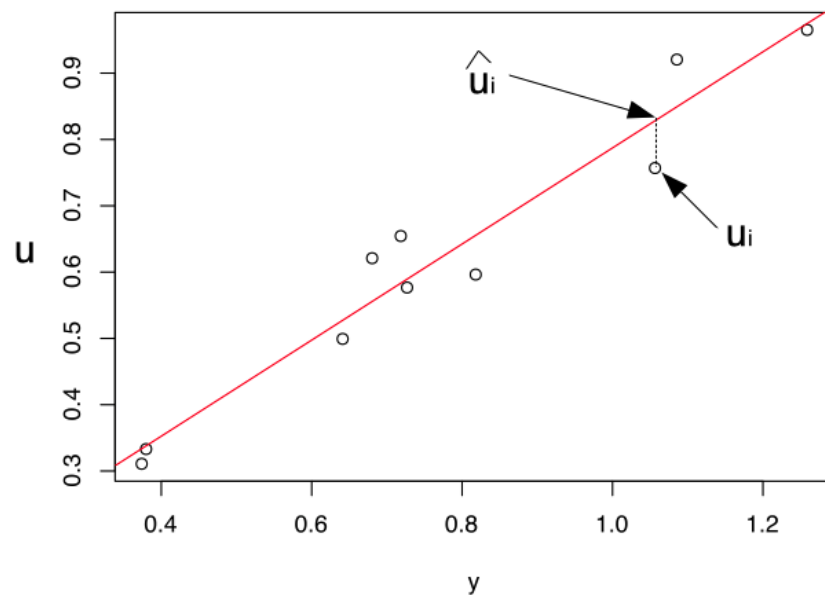


Figure 3.2: Regression of Breeding Values onto Phenotypic Observations

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

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 .

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

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

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

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

$$\begin{aligned}
\text{var}(\hat{u}_i) &= \text{var}(by) = \text{var}(h^2y) \\
&= h^4 \text{var}(y) \\
&= r_{u,y}^2 h^2 \sigma_y^2 \\
&= r_{u,y}^2 \sigma_a^2
\end{aligned} \tag{3.7}$$

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

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

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 ($\text{var}(y)$) can therefore be partitioned as

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

where $\text{var}(u)$ is the genetic additive variance, $\text{var}(pe)$ the variance due to permanent environmental effects and $\text{var}(te)$ 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.

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

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

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

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

$$\begin{aligned} \tilde{y}_i &= \frac{1}{n} \sum_{k=1}^n y_{ik} \\ &= \frac{1}{n} \sum_{k=1}^n (\mu + u_i + pe_i + te_{ik}) \\ &= \mu + u_i + pe_i + \sum_{k=1}^n te_{ik} \end{aligned} \quad (3.12)$$

In this case, we use the mean (\tilde{y}_i) to predict the breeding value (\hat{u}_i)

$$\hat{u}_i = b(\tilde{y}_i - \mu) \quad (3.13)$$

where

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

The single elements are computed as

$$\text{cov}(u, \tilde{y}) = \text{cov}(u, \mu + u + pe + \frac{1}{n} \sum_{k=1}^n te_k) = \text{var}(u) = \sigma_u^2 \quad (3.15)$$

and

$$\text{var}(\tilde{y}) = \text{var}(u) + \text{var}(pe) + \frac{1}{n} \text{var}(te) \quad (3.16)$$

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

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

Inserting this into (3.14) results in

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

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

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

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

$$\hat{u}_i = b * (\bar{y}_i - \mu) \quad (3.20)$$

where

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

and

$$\begin{aligned} \bar{y} &= \frac{1}{n} \sum_{k=1}^n y_k \\ &= \frac{1}{n} \sum_{k=1}^n (\mu + u_k + e_k) \\ &= \mu + \frac{1}{n} \sum_{k=1}^n (u_k + e_k) \\ &= \mu + \frac{1}{n} \sum_{k=1}^n (1/2u_s + 1/2u_{dk} + m_k + e_k) \\ &= \mu + 1/2u_s + \frac{1}{n} \sum_{k=1}^n (1/2u_{dk} + m_k + e_k) \\ &= \mu + 1/2u_s + \frac{1}{n} \sum_{k=1}^n 1/2u_{dk} + \frac{1}{n} \sum_{k=1}^n e_k \end{aligned} \quad (3.22)$$

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

$$\begin{aligned}
\text{cov}(u, \bar{y}) &= \text{cov}\left(u, \frac{1}{2}u_s + \frac{1}{2}\frac{1}{n}\sum_{k=1}^n u_{d,k} + \frac{1}{n}\sum_{k=1}^n m_k + \frac{1}{n}\sum_{k=1}^n e_k\right) \\
&= \text{cov}\left(u, \frac{1}{2}u_s\right) \\
&= \frac{1}{2}\text{cov}(u, u_s) \\
&= \frac{1}{2}\sigma_u^2
\end{aligned} \tag{3.23}$$

where u_s and $u_{d,k}$ denote the breeding values of sire s and dam d of offspring k , 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

$$\text{var}(\bar{y}) = (t + (1 - t)/n)\sigma_y^2 \tag{3.24}$$

where $\sigma_y^2 = \text{var}(u) + \text{var}(e) = \sigma_u^2 + \sigma_e^2$.

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

$$\begin{aligned}
b &= \frac{1/2\sigma_u^2}{(t + (1 - t)/n)\sigma_y^2} \\
&= \frac{1/2h^2\sigma_y^2}{(\frac{1}{4}h^2 + (1 - \frac{1}{4}h^2)/n)\sigma_y^2} \\
&= \frac{2nh^2}{nh^2 + (4 - h^2)} \\
&= \frac{2n}{n + (4 - h^2)/h^2} \\
&= \frac{2n}{n + k}
\end{aligned} \tag{3.25}$$

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

$$\begin{aligned}
r_{u,\bar{y}} &= \frac{\text{cov}(u, \bar{y})}{\sqrt{\text{var}(u)\text{var}(\bar{y})}} \\
&= \frac{1/2h^2\sigma_y^2}{\sqrt{h^2\sigma_y^2(\frac{1}{4}h^2 + (1 - \frac{1}{4}h^2)/n)\sigma_y^2}} \\
&= \frac{1/2h}{\sqrt{\frac{1}{4}h^2 + (1 - \frac{1}{4}h^2)/n}} \\
&= \sqrt{\frac{nh^2}{nh^2 + (4 - h^2)}} \\
&= \sqrt{\frac{n}{n + k}} \tag{3.26}
\end{aligned}$$

The term for $r_{u,\bar{y}}$ in (3.26) approaches 1 as the number of progeny increases, assuming k is constant. The reliability ($r_{u,\bar{y}}^2$) of the predicted breeding value is $n/(n + k)$ and corresponds to $1/2 * b$ computed in (3.25).

Chapter 4

Best Linear Unbiased Prediction (BLUP)

The prediction of breeding values requires to correct the information sources for an appropriate comparison value. So far we have referred to that comparison value as the population mean and we have assumed this correction value to be known. In reality, the computation of these comparison values is a difficult problem. This problem is one of the reasons that nowadays the predictions of all breeding values are based on a method that is called **BLUP**. In this chapter, we first want to have a closer look at the problem of computing these correction factors with which the information sources must be adjusted. After that, the BLUP method will be introduced.

4.1 Problem of Correction

In theory, the population mean is the ideal correction value for all information sources. From our standard model we can derive

$$y = \mu + u + e \quad \rightarrow \quad \bar{y} = \bar{\mu} + \bar{u} + \bar{e} = \mu \quad (4.1)$$

Because, we defined the true breeding value u and the non-identifiable environmental effects e as deviations from a common mean, the average effect of all identifiable environmental components is captured by the population mean μ . But this is only true in an idealized population where all selection candidates are kept in the same environment and where they deliver their performances at the same time. In real world scenarios, this is unrealistic, because e.g. own performance values and progeny performances cannot be delivered at the same time. Furthermore, selection candidates are kept in different herds in different environments. All these factors do have an influence on the performance of

the recorded animals and hence on the predicted breeding values. But good methods for predicting breeding values should be able to correct for such environmental influences. If that is not the case, environmental factors will **bias** the predicted breeding values. To avoid such biases, performance records were subdivided into environmental classes. In dairy cattle such classes were formed based on herds, calving year, calving season and age at first calving. In pigs, performance records might be divided into herds, years and fattening batches. From now on, we call the combination of these environmental effects on the performance records as **identifiable systematic fixed effects**. For the prediction of breeding values, we assume that these fixed effects in a given comparison class have all the same influence on the performance of the animals that are in the same class. Hence if we group all animals who show the same levels of all fixed effects into one comparison class, any biases from the identifiable environment can be avoided.

The more environmental factors can be considered in forming the comparison classes, the better we can correct our performance records for the environmental effects. But when the number of environmental factors increases the number of animals per comparison class decreases. From the statistical point of view, the small number of observations in comparison classes reduce the accuracy with which the environmental fixed effects can be estimated. With smaller comparison groups, the risk that the average breeding value of animals in such a comparison is not zero increases. In case the average breeding value in a comparison group is not zero, predicted breeding values show a deviation which is called **bias**. The occurrence of bias can be shown as follows. Let us assume the average performance of all animals in a comparison group (CG) to be \bar{y}_{CG} :

$$\bar{y}_{CG} = \mu + \bar{u}_{CG} + \bar{e}_{CG} \quad (4.2)$$

In case the average breeding value \bar{u}_{CG} is zero, the population mean μ measures the average identifiable environment effect. If \bar{u}_{CG} is not zero, then the predicted breeding value \hat{u}_i using an older method called selection index, the index value I corresponds to

$$\begin{aligned} I &= b(y_i - (\mu + \bar{u}_{CG})) \\ &= b(y_i - \mu) - b\bar{u}_{CG} \\ &= \hat{u}_i - b\bar{u}_{CG} \end{aligned} \quad (4.3)$$

The first term in the result of (4.3) corresponds to the predicted breeding value where the second term measures the **bias**. This depends on the average breeding values of the animals of the comparison group. If the average breeding value of all animals in the comparison group is zero, then the predicted breeding value from (4.3) is unbiased. Because we have to know the breeding values of the animals in the comparison group to get an unbiased prediction of the

breeding value for a given animal and the breeding values of the animals in the comparison group must also be predicted, this consists of a “chicken-and-egg” problem which cannot be solved.

The solution to this was presented by Charles R. Henderson in several publications ((Henderson, 1973a)) and (Henderson, 1975)). 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 effects model. The properties of the methodology developed by Henderson are similar to those of the selection index method. But the main advantage of Henderson’s methodologies is that phenotypic records do not need to be corrected before breeding values can be predicted. But the effects of the identifiable environmental factors are also a result which come out of the analysis. The methodology developed by Henderson is called **BLUP** and the properties of this methodology are directly incorporated into the name where

- **B** stands for **best** which means that the correlation between the true (u) and the predicted breeding value (\hat{u}) is maximal or the prediction error variance ($\text{var}(u - \hat{u})$) 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 based approaches have found widespread usage in genetic evaluations. They are used 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 Henderson has also developed efficient algorithms to be able to compute predicted breeding values for very large livestock breeding populations. The theoretic foundations, the development of efficient algorithms together with the availability of large computational resources at a very low price have made BLUP to become the de-facto standard methodology for predicting breeding values.

4.2 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 4.1 gives an overview of the dataset.

Table 4.1: 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 (σ_p^2) to be 0.1014 and the heritability (h^2) corresponds to 0.25.

4.3 Linear Mixed Effects Model

A simple linear model contains fixed effects such as *herd* or *sex* of an animal and tries to explain the observations as linear functions of such effects. Because the effects considered in a model cannot account for all influences of a given set of observations, every model must have a random residual component. If a linear model contains besides the residuals any additional random effects, then this model is called a **mixed linear effects model**.

4.3.1 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 as random. Table 4.2 lists a few criteria that might be helpful.

Table 4.2: Classification Factors of Fixed and Random Effects

fixed effect	random effects
classes can be defined exactly	realized value come from an underlying distribution
the value of a class does not have an apriori expected value	each realization is unique
values are exactly estimable	observations are influenced by the variance of the random effect

the expected value of a class effect is of primary interest fixed effects can be corrected for	main interest is on the variance not on the expected value
---	--

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 random effects. Because, we know that breeding values have an expected value of 0 and have a certain variance, they must be modeled as random effects where these properties can be integrated into the model. Furthermore, each animal has a different realization of a breeding value. Exceptions are mono-clonal twins and clones.

From a practical point of view, the software program that is used to analyse the data has also an influence on whether a certain effect is treated as fixed or as random. If a certain effect has very many levels such as herds, then it is sometimes better for the analysis to treat such an effect as random.

4.3.2 Model Specification

In a linear mixed effects model a single observation y_{ijk} is decomposed according to equation (4.4)

$$y_{ijk} = \beta_i + u_j + e_{ijk} \quad (4.4)$$

where β_i stands for the i -th level of a fixed effect, u_j is the j -th realization of the random effect u and e_{ijk} is the residual effect of the k -th observation}. Because, we do not want to model just one observation, but we want to include all observations of a complete population, it is helpful to convert the model in (4.4) into matrix-vector notation. This is shown in equation (4.5)

$$y = X\beta + Zu + e \quad (4.5)$$

where

- y vector of length n of all observations
- β vector of length p of all fixed effects
- X $n \times p$ design matrix linking the fixed effects to the observations
- u vector of length n_u of random effects
- Z $n \times n_u$ design matrix linking random effect to the observations
- e vector of length n of random residual effects.

Furthermore, we assume the following relations for the 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.

$$E(u) = 0 \quad \text{and} \quad E(e) = 0 \quad (4.6)$$

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

$$\text{var}(u) = G \quad \text{and} \quad \text{var}(e) = R \quad (4.7)$$

Under the assumption that $\text{cov}(u, e^T) = 0$, we can compute $\text{var}(y) = Z * \text{var}(u) * Z^T + \text{var}(e) = ZGZ^T + R = V$.

In model (4.5) the vectors β and u are unknown. The solution of the model (4.5) for the unknowns β and u leads to estimates $\hat{\beta}$ for the fixed effects β and for predicted random effects \hat{u} . Unlike with the selection index, with BLUP, we do not have to correct the observations before predicting random effects.

4.3.3 The Solution

An outline of how to derive the BLUP solutions for $\hat{\beta}$ and \hat{u} will be given in an Appendix. The details of this derivation are not important. Therefore, we are presenting here directly the result which are

$$\hat{u} = GZ^TV^{-1}(y - X\hat{\beta}) \quad (4.8)$$

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

$$\hat{\beta} = (X^TV^{-1}X)^{-}X^TV^{-1}y \quad (4.9)$$

The matrix $(X^TV^{-1}X)^{-}$ denotes the generalized inverse of the matrix $(X^TV^{-1}X)$. The generalized inverse K^{-} can be replaced with the simple inverse K^{-1} , whenever the columns of matrix K are linearly independent¹. Analogously to \hat{u} , $\hat{\beta}$ is called the best linear unbiased estimator of the fixed effects β . In short, we can state $\hat{\beta} = BLUE(\beta)$.

4.3.4 Mixed Model Equations

The solutions shown in (4.8) for \hat{u} and in (4.9) for $\hat{\beta}$ 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 and 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{u} and $\hat{\beta}$ are the same when solving the following system of equations simultaneously.

¹For our examples that are shown here, we can always use the simple inverse.

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{bmatrix} \quad (4.10)$$

The above shown equations are called **mixed model equations** (MME). They do no longer contain the inverse V^{-1} and hence these MME are much simpler to solve. The MME contain the inverses R^{-1} and G^{-1} , but we will see with concrete examples that they are much easier to invert. As a consequence, whenever we have to predict breeding values using BLUP, we will use the mixed model equations shown in (4.10).

4.4 Sire Model

The application of the linear mixed effects model from (4.5) to the numerical example in table 4.1. As random effects u we are taking the father s of each animal i with an observation. As fixed effects β we are using the herd effect. When fathers are modeled as random effects, then we call this model a **sire model**. Setting up a sire model for the data in table 4.1 looks as follows

$$\begin{bmatrix} 2.61 \\ 2.31 \\ 2.44 \\ 2.41 \\ 2.51 \\ 2.55 \\ 2.14 \\ 2.61 \\ 2.34 \\ 1.99 \\ 3.1 \\ 2.81 \\ 2.14 \\ 2.41 \\ 2.54 \\ 3.16 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ s_3 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \\ e_9 \\ e_{10} \\ e_{11} \\ e_{12} \\ e_{13} \\ e_{14} \\ e_{15} \\ e_{16} \end{bmatrix}$$

Besides the equation for the sire model we also have to specify the expected values and the variances of all random components. To be able to distinguish the sire model from the general linear mixed effects model, we usually call the random sire effect s and no longer u . The expected values for the random variables were already stated when discussing the general linear mixed effects model in section 4.3.2. Hence

$$E(s) = 0 \quad \text{and} \quad E(e) = 0 \quad \rightarrow \quad E(y) = X\beta \quad (4.11)$$

For the variances there are a few simplifications that we can use in our sire model. The covariance between the random effects s and e are assumed to be 0. The covariances among the single residual effects are also assumed to be 0. Hence, the variance-covariance matrix of the residual effects are $var(e) = I * \sigma_e^2$. The variance of the sire effects s is

$$var(s) = A_s * \sigma_s^2 = G$$

where A_s is the additive genetic relationship matrix between the sires. We will be deriving the matrix A_s in a later chapter. Because our sires are not related, we can say that $A_s = I$ and hence

$$G = I * \frac{\sigma_u^2}{4}$$

Now we are ready to set up the mixed model equations from (4.10) for the sire model. The computation of the numerical solutions from the mixed model equations will be the topic of an exercise.

4.5 Animal Model

The mixed model equations are a universal tool to find BLUPs of random effects and BLUEs of fixed effect simultaneously. On the other hand it is not satisfactory that with the sire model only sires obtain predicted breeding values. All information that is known about the mothers was completely ignored when we specified the sire model. A better approach would be to combine all available information from a given population. This can be done by replacing in the sire model the random sire effects by random animals effects. As a result each animal in the dataset receives a random effect which models its breeding value. This type of model is called an **animal model**. Because the animal model has the breeding values of all animals as random effects, they are often referred to with the variable or the vector a^2 and no longer s as in the sire model. The variance-covariance matrix ($var(a)$) between all animal effects is proportional to the additive genetic relationship matrix A among all animals. We will see in a later chapter how to compute the matrix A .

²This is not the same as the genotypic value in a single locus model.

Chapter 5

Genetic Covariances Between Animals

The prediction of breeding values using BLUP as shown in chapter 4 uses linear mixed effects models where the breeding value of each animal is included as a random effect. Linear mixed effect models in general and specifically Henderson's mixed model equations require us to be able to specify the variance-covariance matrix of all random effects. When using the animal model, the breeding value of each animal is included as a random effect in the linear mixed effects model. As a consequence of that we need to determine the covariance between the true breeding values of all animals. Figure 5.1 tries to display the structure of the required variance-covariance diagrammatically.

The variance-covariance matrix shown at the bottom of Figure 5.1 has the variances of the true breeding values on the diagonal and all the covariances between the true breeding values of all animals as offdiagonal elements. From the formulation of the linear mixed effect model in (4.5), we defined the variance-covariance matrix of the random effects u to be G (see equation (4.7)). When predicting breeding values with the animal model, the variance-covariance matrix of all components in the vector u is defined as

$$\text{var}(u) = G = A * \sigma_u^2 \quad (5.1)$$

where the matrix A is called **numerator relationship matrix**.

5.1 Similarity Between Individuals

At the genetic level there are two different kinds of similarity

1. Identity by descent (IBD)

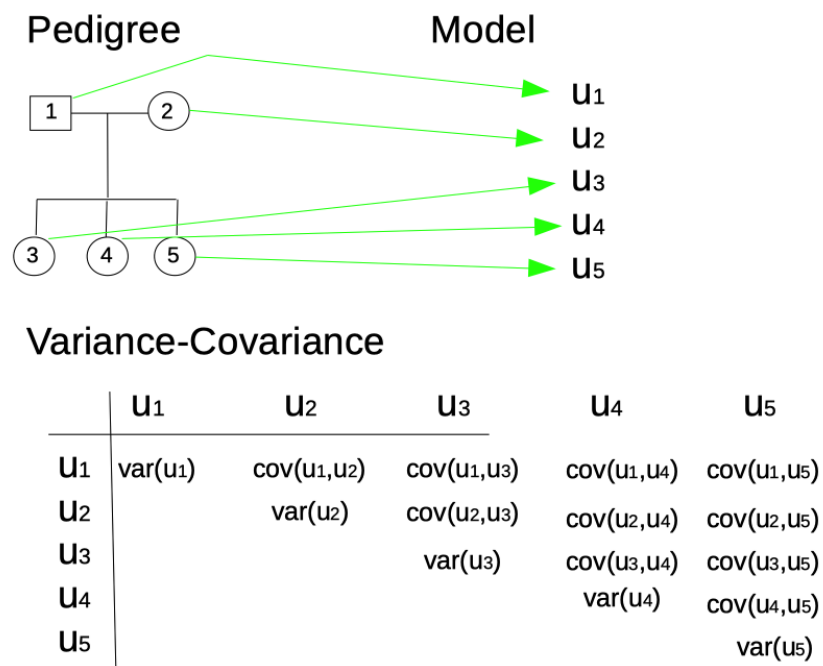


Figure 5.1: Genetic Covariance Between Animals

2. Identity by state

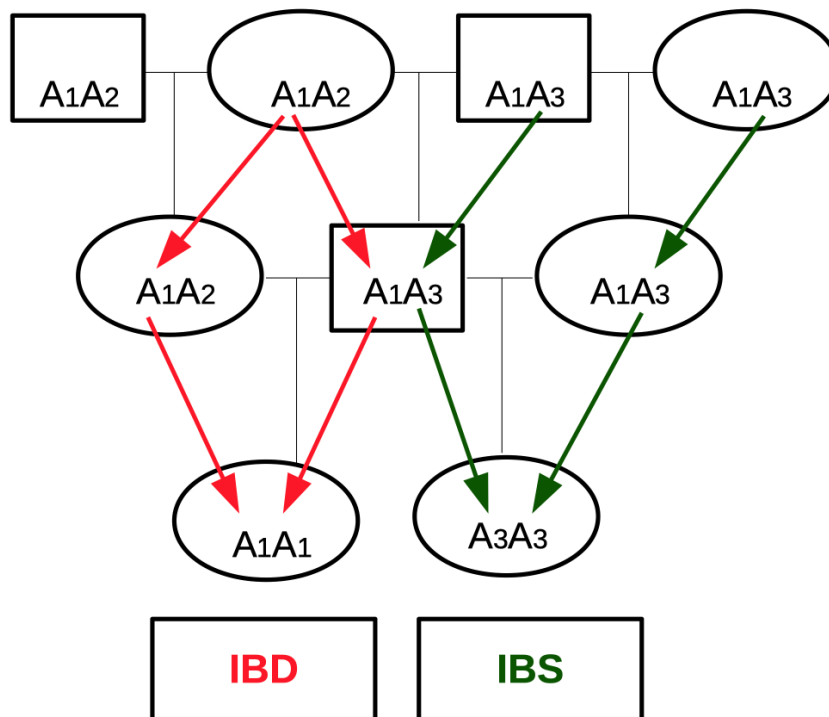


Figure 5.2: Identity by State Versus Identity by Descent

Figure 5.2 illustrates the difference between the two type of identities. The type of graph shown in Figure 5.2 is called a **pedigree** which is commonly used to display the relationship between animals in a population. The rectangle symbols denote male animals and the round symbols stand for female animals. The horizontal connections between female and male animals denote a mating. All animals connected to a vertical line and follow below are progeny of the connected parents.

The notations inside of the symbols stand for the different genotypes of the animals on a given locus. The red arrows denote the path of two A_1 -alleles which are copies of the same ancestral allele. These two copies are called **identical by descent** (IBD). The green arrows show the path of two alleles which are identical by state which do not originate from the same copy of any given ancestral alleles.

5.2 Numerator Relationship Matrix

The probability of identical genes by descent (IBD) occurring in two individuals is termed the co-ancestry or the coefficient of kinship [(Falconer and Mackay, 1996)]. The additive genetic relationship between two individuals is twice their co-ancestry. The matrix that expresses the additive genetic relationship among animals in a population is called the **numerator relationship matrix** A . The matrix A is symmetric and its diagonal elements $(A)_{ii}$ are equal to $1 + F_i$ where F_i is the **coefficient of inbreeding** of animal i . The coefficient of inbreeding F_i indicates whether an animal i is inbred or not. F_i is defined to be half the additive genetic relationship between the parents of i . Hence the diagonal element $(A)_{ii}$ of matrix A corresponds to twice the probability that two gametes taken at random from an animal i will carry IBD-alleles.

The off-diagonal elements $(A)_{ij}$ equals the numerator of the coefficient of relationship between animals i and j . Multiplying the matrix A by the additive genetic variance σ_u^2 leads to the covariance among breeding values. Thus if u_i is the breeding value of animal i then

$$\text{var}(u_i) = (A)_{ii}\sigma_u^2 = (1 + F_i)\sigma_u^2 \quad (5.2)$$

5.2.1 Algorithm To Compute A

The matrix A can be computed using either the

1. path coefficient method or
2. recursive method.

The second method is especially suitable for an implementation by a software program. In what follows the recursive method to compute the components of A is described now. Initially, animals in a pedigree are numbered from 1 to n and ordered such that parents precede their progeny. The following rules are then used to compute the components of A .

- If both parents s and d of animal i are known then
 - the diagonal element $(A)_{ii}$ corresponds to: $(A)_{ii} = 1 + F_i = 1 + \frac{1}{2}(A)_{sd}$ and
 - the off-diagonal element $(A)_{ji}$ is computed as: $(A)_{ji} = \frac{1}{2}((A)_{js} + (A)_{jd})$
 - because A is symmetric $(A)_{ji} = (A)_{ij}$
- If only one parent s is known and assumed unrelated to the mate
 - $(A)_{ii} = 1$
 - $(A)_{ij} = (A)_{ji} = \frac{1}{2}((A)_{js})$
- If both parents are unknown
 - $(A)_{ii} = 1$
 - $(A)_{ij} = (A)_{ji} = 0$

5.2.2 Numeric Example

We are given the following pedigree and we want to compute the matrix A using the recursive method described in 5.2.1.

Table 5.1: Example Pedigree To Compute Additive Genetic Relationship Matrix

Calf	Sire	Dam
3	1	2
4	1	NA
5	4	3
6	5	2

The first step of the computations of A are the numbering and the ordering of all the animals. This is already done in the pedigree shown in Table 5.1. The components of A are computed row-by-row starting with $(A)_{11}$.

$$(A)_{11} = 1 + F_1 = 1 + 0 = 1$$

$$(A)_{12} = 0 = (A)_{21}$$

$$(A)_{13} = \frac{1}{2}((A)_{11} + (A)_{12}) = 0.5 = (A)_{31}$$

$$(A)_{14} = \frac{1}{2}(A)_{11} = 0.5 = (A)_{14}$$

$$(A)_{15} = \frac{1}{2}(A)_{14} + (A)_{13} = 0.5 = (A)_{51}$$

$$(A)_{16} = \frac{1}{2}(A)_{15} + (A)_{12} = 0.25$$

The same computations are also done for all the other components of the matrix A . The final result for the matrix looks as follows

$$A = \begin{bmatrix} 1 & 0 & 0.5 & 0.5 & 0.5 & 0.25 \\ 0 & 1 & 0.5 & 0 & 0.25 & 0.625 \\ 0.5 & 0.5 & 1 & 0.25 & 0.625 & 0.5625 \\ 0.5 & 0 & 0.25 & 1 & 0.625 & 0.3125 \\ 0.5 & 0.25 & 0.625 & 0.625 & 1.125 & 0.6875 \\ 0.25 & 0.625 & 0.5625 & 0.3125 & 0.6875 & 1.125 \end{bmatrix}$$

As a result, we can see from the components of the above shown matrix A that animals 1 and 2 are not related to each other. Furthermore from the diagonal elements of A , it follows that animals 5 and 6 are inbred while animals 1 to 4 are not inbred.

5.3 The Inverse Numerator Relationship Matrix

The general form of Henderson's mixed model equations as shown in (4.10) depend on four matrices

1. Design matrix X for the fixed effects
2. Design matrix Z for the random effects
3. The inverse variance-covariance matrix R^{-1} for the residuals e and
4. The inverse variance-covariance matrix G^{-1} for the random breeding values a .

When using the animal model as described in section @ref(#animalmodel), we have seen in (5.1) that G corresponds to the product of the numerator relationship matrix A and the genetic additive variance σ_u^2 . But the mixed model equations depend on the inverse G^{-1} of G which means

$$G^{-1} = (A * \sigma_u^2)^{-1} = A^{-1} * \frac{1}{\sigma_u^2} \quad (5.3)$$

From (5.3) we can see that in order to be able to set up the mixed model equations for an animal model, we need the inverse numerator relationship matrix A^{-1} . Because in practical routine predictions of breeding values, we have something in the order of 10 million animals that we predict breeding values for. Hence the matrix A has 10 million rows and 10 million columns. A matrix of that size cannot be inverted explicitly with commonly known methods such as described under https://en.wikipedia.org/wiki/Invertible_matrix. This would have been the end of the BLUP animal model, if not Henderson published in [(Henderson, 1976)] a set of simple rules to directly construct the matrix A^{-1} without setting up the numerator relationship matrix.

5.4 Structure of A^{-1}

When looking at a concrete example of an inverse of a numerator relationship matrix as shown below, we can discover some important facts. Let us assume the following pedigree.

Table 5.2: Pedigree Used To Compute Inverse Numerator Relationship Matrix

Calf	Sire	Dam
1	NA	NA
2	NA	NA
3	NA	NA
4	1	2
5	3	2

The numerator relationship matrix A for the pedigree shown in Table ?? is shown in (5.4).

$$A = \begin{bmatrix} 1 & 0 & 0 & 0.5 & 0 \\ 0 & 1 & 0 & 0.5 & 0.5 \\ 0 & 0 & 1 & 0 & 0.5 \\ 0.5 & 0.5 & 0 & 1 & 0.25 \\ 0 & 0.5 & 0.5 & 0.25 & 1 \end{bmatrix} \quad (5.4)$$

The matrix A^{-1} shown below corresponds to the inverse of the matrix A from (5.4).

$$A^{-1} = \begin{bmatrix} 1.5 & 0.5 & 0 & -1 & 0 \\ 0.5 & 2 & 0.5 & -1 & -1 \\ 0 & 0.5 & 1.5 & 0 & -1 \\ -1 & -1 & 0 & 2 & 0 \\ 0 & -1 & -1 & 0 & 2 \end{bmatrix}$$

From the above shown inverse A^{-1} , we recognize that A^{-1} has a simpler structure than A itself. The reason for this is that non-zero elements occur only at matrix elements of A^{-1} corresponding to parents and progeny or to mates. Furthermore off-diagonal elements corresponding to mates are positive and elements corresponding to parents and progeny are negative. These observations were used by Henderson in (Henderson, 1976) to come up with the rules described below.

5.5 Henderson's Rule To Set Up A^{-1}

We denote the row or column index corresponding to an animal of interest as i and the row or column index belonging to the animals father as s and the row or column index corresponding to animal i 's mother as d . The rules differentiate the following three cases

1. both parents of animal i are known
2. only one parent of animal i is known
3. both parents of animal i are unknown

5.5.1 Both Parents Known

- add 2 to the diagonal-element (i, i)
- add -1 to off-diagonal elements (s, i) , (i, s) , (d, i) and (i, d)
- add $\frac{1}{2}$ to elements (s, s) , (d, d) , (s, d) , (d, s)

5.5.2 Only One Parent Known

We assume that sire s is known

- add $\frac{4}{3}$ to diagonal-element (i, i)
- add $-\frac{2}{3}$ to off-diagonal elements $(s, i), (i, s)$
- add $\frac{1}{3}$ to element (s, s)

5.5.3 Both Parents Unknown

- add 1 to diagonal-element (i, i)

The application of Henderson's rules to construct A^{-1} directly will be a problem in one of the coming exercises. When applying the above described rules, it has to be noted well that this simple version of the rules are only valid for a pedigree without inbreeding. We will see in section @ref(#derivationofhendersonrules) how to extend the simple rules such that they can be used for a general pedigree with inbreeding.

5.6 Derivation of Henderson's Rules

Henderson's rules can be related to the so-called LDL-decomposition of the numerator relationship matrix A . The result of this decomposition consists of the equivalence between matrix A and the product of three matrices L , D and L^T . The matrix L is a lower triangular matrix and the matrix D is a diagonal matrix. The reason for why we are doing this decomposition of A is that the matrices L and D can much easier be inverted than the matrix A . The LDL-decomposition is a general procedure in linear algebra and it can be applied to any symmetric and positive-definite matrix not only to numerator relationship matrices. But when the LDL-decomposition is applied to a numerator relationship matrix, the matrices L and D do also have a special genetic meaning. This meaning is revealed in the following derivation.

5.6.1 Decomposition of True Breeding Value and its Variance

The true breeding value (u_i) of animal i can be decomposed into the true breeding values of animal i 's parents s and d and the mendelian sampling term m_i

$$u_i = \frac{1}{2}u_s + \frac{1}{2}u_d + m_i \quad (5.5)$$

Applying the decomposition shown in (5.10) for all animals in the pedigree and combining the decompositions into a matrix-vector notation, we get

$$u = P \cdot u + m \quad (5.6)$$

where

- u vector of true breeding values for all animals
- P matrix linking animals to their parents
- m vector of mendelian sampling terms

Equation (5.6) can be interpreted as regression of the true breeding values onto parental breeding values. In such a regression the random term m is like a residual term. The genetic meaning of m corresponds to the deviation of u_i from the full-sib average of the true breeding values within the nuclear family with parents s and d . The term m is called **mendelian sampling** term. The source of m is the fact that during meiosis, parental alleles are randomly assigned to each progeny. Bulmer [(Bulmer, 1971)] has shown that m_i are independent from true breeding values of parents s and d . Using this result, we take the variance on both sides of (5.10) leading to

$$\begin{aligned} \text{var}(u_i) &= \text{var}\left(\frac{1}{2}u_s + \frac{1}{2}u_d + m_i\right) \\ &= \frac{1}{4}\text{var}(u_s) + \frac{1}{4}\text{var}(u_d) + \frac{1}{2}\text{cov}(u_s, u_d) + \text{var}(m_i) \end{aligned} \quad (5.7)$$

From (5.1) together with the definition of the numerator relationship matrix A , we know that

$$\begin{aligned} \text{var}(u_i) &= (1 + F_i)\sigma_u^2 \\ \text{var}(u_s) &= (1 + F_s)\sigma_u^2 \\ \text{var}(u_d) &= (1 + F_d)\sigma_u^2 \\ \text{cov}(u_s, u_d) &= (A)_{sd}\sigma_u^2 = 2F_i\sigma_u^2 \end{aligned} \quad (5.8)$$

5.6.2 Variance of Mendelian Sampling Terms

Inserting the relations from (5.8) into (5.7) and solving for $\text{var}(m_i)$ gives the following result

$$\begin{aligned} \text{var}(m_i) &= \text{var}(u_i) - \frac{1}{4}\text{var}(u_s) - \frac{1}{4}\text{var}(u_d) - \frac{1}{2}\text{cov}(u_s, u_d) \\ &= (1 + F_i)\sigma_u^2 - \frac{1}{4}(1 + F_s)\sigma_u^2 - \frac{1}{4}(1 + F_d)\sigma_u^2 - \frac{1}{2} * 2 * F_i * \sigma_u^2 \\ &= \left(\frac{1}{2} - \frac{1}{4}(F_s + F_d)\right)\sigma_u^2 \end{aligned} \quad (5.9)$$

In case where only parent s of animal i is known the terms in (5.10) and (5.9) change to

$$u_i = \frac{1}{2}u_s + \frac{1}{2}u_d + m_i \quad (5.10)$$

$$\begin{aligned} u_i &= \frac{1}{2}u_s + m_i \\ \text{var}(m_i) &= \left(1 - \frac{1}{4}(1 + F_s)\right) \sigma_u^2 \\ &= \left(\frac{3}{4} - \frac{1}{4}F_s\right) \sigma_u^2 \end{aligned} \quad (5.11)$$

When both parents are unknown, we get

$$\begin{aligned} u_i &= m_i \\ \text{var}(m_i) &= \sigma_u^2 \end{aligned} \quad (5.12)$$

5.6.3 Decomposition of A

The true breeding values u_s of sire s and u_d of dam d can be decomposed in a similar way as shown for the true breeding value u_i in (5.10).

$$\begin{aligned} u_s &= \frac{1}{2}u_{ss} + \frac{1}{2}u_{ds} + m_s \\ u_d &= \frac{1}{2}u_{sd} + \frac{1}{2}u_{dd} + m_d \end{aligned} \quad (5.13)$$

where

ss sire of s
 ds dam of s
 sd sire of d
 dd dam of d

Using (5.13) together with (5.10) leads to the following expression for u_i

$$\begin{aligned} u_i &= \frac{1}{2}u_s + \frac{1}{2}u_d + m_i \\ &= \frac{1}{2}\left(\frac{1}{2}u_{ss} + \frac{1}{2}u_{ds} + m_s\right) + \frac{1}{2}\left(\frac{1}{2}u_{sd} + \frac{1}{2}u_{dd} + m_d\right) + m_i \\ &= \frac{1}{4}u_{ss} + \frac{1}{4}u_{ds} + \frac{1}{4}u_{sd} + \frac{1}{4}u_{dd} + \frac{1}{2}m_s + \frac{1}{2}m_d + m_i \end{aligned}$$

This type of decomposition can also be done for the grand-parents of animal i and further back until we get to true breeding values of animals with unknown parents. Animals of ancestor generations with unknown parents are called **founder population**. The process of decomposing true breeding values back through all generations of a pedigree is called **recursive decomposition** of animal i 's true breeding value. Because we know from (5.12) that the decomposition of true breeding values u_k for an animal k with unknown parents is $u_k = m_k$, the result of the recursive decomposition of u_i is a sum of mendelian sampling terms linking the ancestors of i back to the founder population.

Ordering all animals in a pedigree according to their age and writing the result of the recursive decomposition of all true breeding values in matrix-vector notation leads to

$$u = L \cdot m \quad (5.14)$$

The matrix L is a lower triangular matrix. The row corresponding to animal i is the average of the rows in L corresponding to parents s and d of i . The matrix L contains the path of the gene flow from the base population to all animals in the population. From equation (5.14), we are computing the variance of all true breeding values which leads to

$$\begin{aligned} \text{var}(u) &= \text{var}(L \cdot m) \\ &= L \cdot \text{var}(m) \cdot L^T \\ &= L \cdot (D\sigma_u^2) \cdot L^T \\ &= L \cdot D \cdot L^T \sigma_u^2 = A\sigma_u^2 \end{aligned}$$

From (??), the LDL-decomposition of A follows directly as $A = LDL^T$. The single components m_i are independent of each other. This also means that $\text{cov}(m_i, m_j) = 0$ for $i \neq j$. Hence the matrix D is a diagonal matrix. In section @ref(#variancemendeliansamplingterm) we have seen that $\text{var}(m_i)$ always contain σ_u^2 as a factor. Hence it is reasonable to express $\text{var}(m)$ as a product of a diagonal matrix D times the genetic additive variance σ_u^2 . The diagonal elements of matrix D are computed as shown in section @ref(#variancemendeliansamplingterm).

The mixed model equations require the inverse numerator relationship matrix A^{-1} . Thanks to the LDL-decomposition of A , we can compute A^{-1} as

$$A^{-1} = (L \cdot D \cdot L^T)^{-1} = (L^T)^{-1} \cdot D^{-1} \cdot L^{-1} \quad (5.15)$$

The inverse D^{-1} is easy to compute, because D is a diagonal matrix. As a consequence of that D^{-1} is also a diagonal matrix where the elements $(D^{-1})_{ii}$

correspond to the inverse of the elements of the original matrix D , i.e. $(D^{-1})_{ii} = 1/(D)_{ii}$. The matrix L^{-1} is also a lower triangular matrix and can easily be computed based on the two relations for the vector m . Based on (5.6), we know

$$m = u - Pu = (I - P)u \quad (5.16)$$

and from (5.14), we get

$$m = L^{-1}u \quad (5.17)$$

Setting both expressions for m in (5.16) and (5.17) equal can be used to obtain L^{-1}

$$m = (I - P)u = L^{-1}u \quad (5.18)$$

Therefore,

$$L^{-1} = I - P \quad (5.19)$$

5.6.4 General Version of Henderson's Rule

Based on the decomposition of A^{-1} shown in (5.15), the general version of Henderson's rule where inbreeding of animals can be accounted for can be summarized as

- Start with a matrix A^{-1} where all elements are set to 0.
- Let d^i be the i -th diagonal element of D^{-1} for animal i , assuming i has parents s and d .
- Then add the following contributions to A^{-1}
 - d^i to the element (i, i)
 - $-d^i/2$ to the elements (s, i) , (i, s) , (d, i) , (i, d)
 - $d^i/4$ to the elements (s, s) , (s, d) , (d, s) , (d, d)

The contributions to rows and columns corresponding to parents s and d are only added, if they are known. Because the elements d^i are dependent on the inbreeding coefficients F_s and F_d of parents s and d , we have to find an efficient way to compute inbreeding coefficients for all animals in a population.

5.7 Computing Inbreeding Coefficients

Inbreeding coefficients can be computed using different methods. From all these methods, we are just showing the one method which was described in [(Quaas, 1976)]. This method is based on the properties of a second decomposition of the

numerator relationship matrix A which is called the **cholesky decomposition**. The cholesky decomposition of a matrix A corresponds to

$$A = RR^T \quad (5.20)$$

where the matrix R is a lower triangular matrix. At this point we have to be clear about this. In practical routine evaluations, we will not explicitly compute this decomposition, because we do not want to construct A explicitly. We are just using the properties of (5.20) to find an efficient way to compute the diagonal elements $(A)_{ii}$ of A without constructing the complete matrix A . The diagonal elements $(A)_{ii}$ are important, because they contain the inbreeding coefficients (F_i) , as $(A)_{ii} = 1 + F_i$. Based on (5.20), $(A)_{ii}$ can be computed from the components of R as

$$(A)_{ii} = \sum_{j=1}^i (R)_{ij}^2 \quad (5.21)$$

This can be shown with a small 3×3 matrix A

$$\begin{bmatrix} (A)_{11} & (A)_{12} & (A)_{13} \\ (A)_{21} & (A)_{22} & (A)_{23} \\ (A)_{31} & (A)_{32} & (A)_{33} \end{bmatrix} = \begin{bmatrix} (R)_{11} & 0 & 0 \\ (R)_{21} & (R)_{22} & 0 \\ (R)_{31} & (R)_{32} & (R)_{33} \end{bmatrix} * \begin{bmatrix} (R)_{11} & (R)_{21} & (R)_{31} \\ 0 & (R)_{22} & (R)_{32} \\ 0 & 0 & (R)_{33} \end{bmatrix}$$

For the above shown example, the diagonal elements $(A)_{ii}$ are computed as

$$\begin{aligned} (A)_{11} &= (R)_{11}^2 \\ (A)_{22} &= (R)_{21}^2 + (R)_{22}^2 \\ (A)_{33} &= (R)_{31}^2 + (R)_{32}^2 + (R)_{33}^2 \end{aligned}$$

This shows that diagonal elements $(A)_{ii}$ can be computed using just all the components of row i in R up to the diagonal. Next, we have to show how to compute the components of R .

5.7.1 Recursive Computation of R

We are setting the two decompositions of A equal which leads to

$$A = R * R^T = L * D * L^T \quad (5.22)$$

Let us write the matrix R as a product of two matrices L and S where L corresponds to the matrix from the LDL-decomposition and insert that product into (5.22)

$$A = R * R^T = L * S * (L * S)^T = L * S * S^T * L^T = L * D * L^T \quad (5.23)$$

From (5.23), we conclude that $D = S \cdot S^T$ where S is also a diagonal matrix with elements $(S)_{ii} = \sqrt{(D)_{ii}}$. For our small example we get

$$\begin{bmatrix} (R)_{11} & 0 & 0 \\ (R)_{21} & (R)_{22} & 0 \\ (R)_{31} & (R)_{32} & (R)_{33} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ (L)_{21} & 1 & 0 \\ (L)_{31} & (L)_{32} & 1 \end{bmatrix} * \begin{bmatrix} (S)_{11} & 0 & 0 \\ 0 & (S)_{22} & 0 \\ 0 & 0 & (S)_{33} \end{bmatrix} \quad (5.24)$$

From (5.24), we can see that the diagonal elements $(R)_{ii}$ are equal to $(S)_{ii}$. Therefore

$$(R)_{ii} = (S)_{ii} = \sqrt{(D)_{ii}} \quad (5.25)$$

Earlier, we have seen that diagonal elements $(D)_{ii}$ of D correspond to

$$(D)_{ii} = \frac{1}{2} - \frac{1}{4} (F_s + F_d) = 1 - 0.25((A)_{ss} + (A)_{dd}) \quad (5.26)$$

and hence

$$(R)_{ii} = \sqrt{1 - 0.25((A)_{ss} + (A)_{dd})} \quad (5.27)$$

The components $(A)_{ss}$ and $(A)_{dd}$ correspond to diagonal elements of A of parents of s and d which were computed earlier.

The off-diagonal elements of R are computed as

$$(R)_{ij} = (L)_{ij} * (S)_{jj} \quad (5.28)$$

One property of the matrix L is that any element $(L)_{ij}$ is equal to the average of elements $(L)_{sj}$ and $(L)_{dj}$, if s and d are parents of animal i . Using this we get

$$\begin{aligned} (R)_{ij} &= (L)_{ij} * (S)_{jj} \\ &= \frac{1}{2}((L)_{sj} + (L)_{dj}) * (S)_{jj} \\ &= \frac{1}{2}((R)_{sj} + (R)_{dj}) \end{aligned} \quad (5.29)$$

With that we have shown how to compute all elements of R recursively. This requires an ordering of all animals according to their age.

Chapter 6

Variance and Inbreeding

Recalling from chapter 5 the variance ($var(u_i)$) of a breeding value u_i of animal i is given by

$$var(u_i) = (1 + F_i)\sigma_u^2 \quad (6.1)$$

where F_i is the inbreeding coefficient of animal i and σ_u^2 corresponds to the additive genetic variance. At first sight this seems difficult to understand why the inbreeding coefficient increases the variance of a breeding value. This chapter aims at explaining the relationship between inbreeding and the genetic variance. The material presented here is based on chapters 3 and 14 of (Falconer and Mackay, 1996).

6.1 Inbreeding

Inbreeding means mating related individuals. The degree of relationship between individuals in a population depends on the size of the population. In a population of bisexual organisms every individual has 2^t ancestors when looking t generations back. Already for small t the number of required individuals in a population becomes quite large in order to provide separate unrelated ancestors. As a consequence of that any pair of individuals must be related to some degree. Furthermore, pairs mating at random are expected to be more related in smaller populations compared to when individuals mate at random in large populations. Therefore properties of small populations can be treated as consequences of inbreeding.

An important consequence of two individuals having a common ancestor is that they may both carry replicates of one of the alleles present in the common ancestor. If the two individuals mate, they may pass on these replicates to their offspring.

6.1.1 Inbreeding in Idealized Population

The coefficient of inbreeding is deduced assuming an idealized population. Starting with the base population and its progeny forming generation 1 the development of the inbreeding coefficient is computed. What is meant by the term **idealized population** is shown in Figure 6.1.

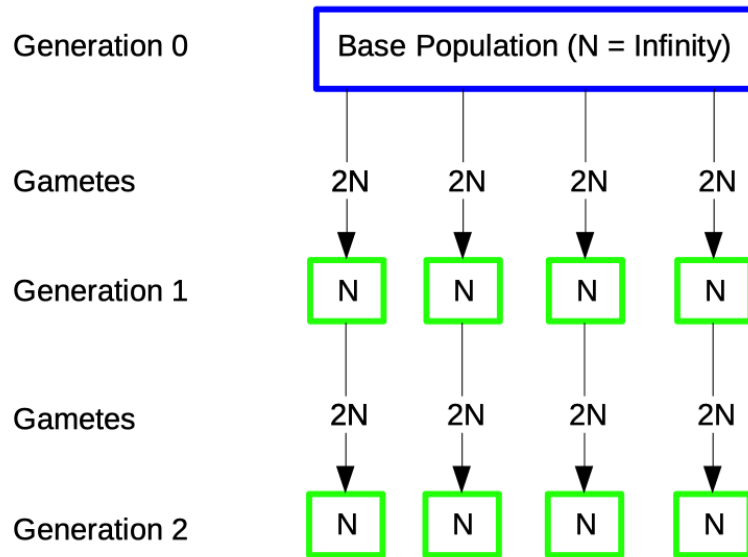


Figure 6.1: Subdivision of a single large population into a number of subpopulations or lines

The computation of the inbreeding coefficient may be visualized by the following situation. Let us assume a hermaphrodite marine organism, capable of self-fertilization shedding eggs and sperm into the sea. There are N individuals, each shedding equal numbers of gametes which mate at random. At a given locus, all the alleles in the base population have to be regarded as non-identical. For that single locus, among the gametes shed by the base population there are $2N$ different sorts in equal number. What is the probability that a pair of gametes taken at random carry identical alleles? This probability corresponds to the **inbreeding coefficient** (F). Any gamete has a chance of $1/(2N)$ to mate with a gamete carrying the same allele. Hence the inbreeding coefficient (F_1) in generation 1 corresponds to

$$F_1 = \frac{1}{2N} \quad (6.2)$$

In generation 2 there are two ways in which identical homozygotes can arise, first from new replication of alleles and second from previous replications. The probability of newly replicated alleles coming together in a new zygote is again $1/(2N)$. The remaining proportion $1 - (1/(2N))$ of zygotes carries alleles that are not identical, but may have been identical from the previous generation. The total probability of identical zygotes in generation 2 is

$$F_2 = \frac{1}{2N} + (1 - \frac{1}{2N}) * F_1 \quad (6.3)$$

The same argument leading to equation (6.3) applies to any subsequent generations. We can therefore write the more general statement

$$F_t = \frac{1}{2N} + (1 - \frac{1}{2N}) * F_{t-1} \quad (6.4)$$

Thus the inbreeding coefficient given in (6.4) consists of two parts: first an increment $(1/(2N))$ attributable to new inbreeding and a remainder that is caused by inbreeding of previous generations. The increment $(1/(2N))$ is assigned to a new variable ΔF , so that

$$\Delta F = \frac{1}{2N} \quad (6.5)$$

With that equation (6.4) can be re-written as

$$F_t = \Delta F + (1 - \Delta F) * F_{t-1} \quad (6.6)$$

Solving (6.6) for ΔF results in

$$\Delta F = \frac{F_t - F_{t-1}}{1 - F_{t-1}} \quad (6.7)$$

The measure of the **rate of inbreeding** given in equation (6.7) provides a convenient way of generalising the concept of inbreeding beyond the simplifications of the idealized population. This generalization provides a means of comparing inbreeding effects of different breeding systems. When expressing inbreeding in terms of ΔF , equation (6.4) is valid for any breeding system and is not restricted to the idealized population where ΔF is set to $1/(2N)$. So far, we have just related the inbreeding coefficient in one generation to the previous generation. It remains to express the inbreeding coefficient in terms of a set of

properties of the base population. This is simplified by defining the symbol P as the complement of the inbreeding coefficient F , hence

$$P = 1 - F \quad (6.8)$$

The quantity symbolized by P is known as the **panmictic index**. Using (6.8) and inserting it into (6.7) leads to

$$\frac{P_t}{P_{t-1}} = 1 - \Delta F \quad (6.9)$$

Hence the rate at which P increases from one generation to the next is reduced to a constant $1 - \Delta F$. Going back t generations to the base population leads to

$$P_t = (1 - \Delta F)^t * P_0 \quad (6.10)$$

In the base population, we assumed no inbreeding, hence $F_0 = 0$ and $P_0 = 1$. Using the result of (6.10) to compute F_t leads to

$$F_t = 1 - (1 - \Delta F)^t \quad (6.11)$$

6.1.2 Variance of Gene Frequency

According the Hardy-Weinberg Equilibrium, gene frequencies are constant over generations. But this is only true, if the base population is not divided into subpopulations or lines. If the base population is split into separate lines as shown in Figure 6.1, the gene frequencies in the single lines start to show variation. The amount of the variation is quantified by the variance of the gene frequencies.

The variance ($\sigma_{\Delta q}^2$) of the change of gene frequency in one generation is first of all the variance of a binomial random variable and can be expressed in terms of the rate of inbreeding, as shown below.

$$\sigma_{\Delta q}^2 = \frac{p_0 q_0}{2N} = p_0 q_0 \Delta F \quad (6.12)$$

An equivalent way of writing (6.13) is in terms of the inbreeding coefficient (F_1) and the variance (σ_q^2) of gene frequencies after one generation. It follows that the relationship is the same after any number of generations, so that after t generations

$$\sigma_q^2 = p_0 q_0 F_t \quad (6.13)$$

Table 6.1: Genotype Frequencies in Population as a Whole

Genotype	Frequency
A_1A_1	$p_0^2 + \sigma_q^2$
A_1A_2	$2p_0q_0 - 2\sigma_q^2$
A_2A_2	$q_0^2 + \sigma_q^2$

6.1.3 Genotype Frequencies

The genotype frequencies in the population as a whole (across all generations) can be deduced from the knowledge of the variance of gene frequencies. If an allele has frequency q in a given line, homozygotes of that allele have frequency q^2 in that line. The frequency of the homozygotes in the complete population over all lines will be the mean value of q^2 across all lines. The mean frequency of homozygotes is written as \bar{q}^2 . The value of \bar{q}^2 is obtained by the knowledge of the variance of gene frequencies. In general the variance of a series of observations is obtained by

From the general formula of obtaining the variance of a set of observations corresponding to

$$\sigma_q^2 = (\bar{q}^2) - \bar{q}^2 \quad (6.14)$$

the mean frequency of homozygotes \bar{q}^2 is obtained as

$$\bar{q}^2 = \sigma_q^2 + \bar{q}^2 \quad (6.15)$$

where \bar{q} is the mean gene frequency which is the same as the original gene frequency q_0 . Thus in the complete population, the frequency of the homozygotes of a particular allele increases and is always in excess of the original frequency by an amount equal to the variance of the gene frequency among the lines. In a two-allele system, the same applies to the other allele. The genotypic frequencies for a locus with two alleles can then be summarized as shown in Table 6.1.

The genotype frequencies shown in Table 6.1 are no longer in Hardy-Weinberg equilibrium. This change in genotype frequencies is the result of a mechanism which is called the **dispersive process**. The dispersive process is active as soon as the idealized base population is subdivided into single lines. The increase of the frequency of the homozygous genotypes is the source of a phenomenon called **inbreeding depression**. This depression refers to the reduced fitness of individuals in populations with increasing levels of inbreeding.

Combining the formulas (6.13) and (6.15) and furthermore dropping the subscript t in F_t leads to

Table 6.2: Genotype Frequencies for a bi-allelic locus, expressed in terms of inbreeding coefficient F

Genotype	Original Frequencies	Changes due to inbreeding
A_1A_1	p_0^2	$+p_0q_0F$
A_1A_2	$2p_0q_0$	$-2p_0q_0F$
A_2A_2	q_0^2	$+p_0q_0F$

Table 6.3: Derivation of Inbreeding Depression

Genotype	Frequency	Value	Product
A_1A_1	$\bar{p}^2 + \bar{p}\bar{q}F$	a	$(\bar{p}^2 + \bar{p}\bar{q}F)a$
A_1A_2	$2\bar{p}\bar{q} - 2\bar{p}\bar{q}F$	d	$(2\bar{p}\bar{q} - 2\bar{p}\bar{q}F)d$
A_2A_2	$\bar{q}^2 + \bar{p}\bar{q}F$	$-a$	$-(\bar{q}^2 + \bar{p}\bar{q}F)a$

$$\bar{q}^2 = \bar{q}_0^2 + \sigma_q^2 = \bar{q}_0^2 + p_0q_0F \quad (6.16)$$

Based on (6.16) Table 6.1 with the genotype frequencies can be re-written as shown in Table 6.2 where genotype frequencies are now expressed in terms of the inbreeding coefficient F .

6.2 Changes of Mean Value

So far, we have explained the consequences of inbreeding on the genotype frequencies. In this section, we have a look at how inbreeding affects the mean values of metric characters. The most important consequence of inbreeding is the reduction of the mean phenotypic value of characters connected to reproduction and fitness. This phenomenon is known as **inbreeding depression**. In saying that a certain trait shows inbreeding depression, we refer to the average change of mean value in a number of lines. The separate lines are commonly found to differ to a greater or lesser extent in the change they show, as indeed, we should expect in consequence of random drift of gene frequencies. The change of mean value that we discuss now refers to changes of the mean value of a number of lines derived from a common base population.

Consider a population, subdivided into a number of lines, with a coefficient of inbreeding F . Table 6.3 shows the genotype frequencies derived earlier, the values of the single genotypes and the contribution to the population mean for each genotype. The allele frequencies \bar{p} and \bar{q} correspond to the frequencies in the whole population.

Summing over the last column in Table 6.3 results in the population mean for the given trait.

$$\begin{aligned}
 M_F &= (\bar{p}^2 + \bar{p}\bar{q}F)a + (2\bar{p}\bar{q} - 2\bar{p}\bar{q}F)d - (\bar{q}^2 + \bar{p}\bar{q}F)a \\
 &= a(\bar{p} - \bar{q}) + 2d\bar{p}\bar{q} - 2d\bar{p}\bar{q}F \\
 &= a(\bar{p} - \bar{q}) + 2d\bar{p}\bar{q}(1 - F) \\
 &= M_0 - 2d\bar{p}\bar{q}F
 \end{aligned} \tag{6.17}$$

where M_0 is the population mean before inbreeding. The change of mean resulting from inbreeding is $2d\bar{p}\bar{q}F$.

6.3 Changes of Variance

The effect of inbreeding on the genetic variance becomes apparent when again imagining the change of gene frequencies in different lines that emerge from a homogeneous base population. Within the different lines, the gene frequencies change to the dispersive process of random drift. This makes that over time some allele frequencies will tend towards 0 and frequencies of other alleles will tend towards 1. This tendency towards the extremes is different in the different lines. As a result in the populations, the similarity within lines increases, but between the lines the differences get bigger.

The subdivision of a population into lines introduces an additional component of variance, the between-line variance component. This means that the total genetic variance is re-distributed into the within-line component and the between line component.

6.3.1 Redistribution of Genetic Variance

For reasons of simplicity, we are currently just looking at genetic loci with purely additive effects. That means the dominance term d for such additive loci is 0. Strictly speaking, the results shown here apply only to traits with no non-additive variance. But still, these results serve as a good approximation to the effect of inbreeding on the genetic variance.

Consider first a single locus. When there is not dominance the genotypic variance in the base population is given by

$$V_G = V_A = 2p_0q_0a^2 \tag{6.18}$$

The variance within one given line is $V_G = 2pqa^2$, where p and q are the allele frequencies in that given line. The mean genetic variance ($V_{\bar{G}}$) within the lines is

Table 6.4: Partitioning of the genic additive variance in a population with lines and a given inbreeding coefficient F

Source	Variance
Between lines	$2FV_G$
Within lines	$(1 - F)V_G$
Total	$(1 + F)V_G$

$$V_{\bar{G}} = 2(\bar{p}q)a^2 \quad (6.19)$$

where $(\bar{p}q)$ is the mean value of pq over all lines. The term $2(\bar{p}q)$ is the overall frequency of heterozygotes in the whole population, which, by Table 6.2, is equal to $2p_0q_0(1 - F)$ where F is the inbreeding coefficient. Therefore

$$\begin{aligned} V_{\bar{G}} &= 2(\bar{p}q)a^2 \\ &= 2p_0q_0(1 - F) \\ &= V_G(1 - F) \end{aligned} \quad (6.20)$$

This remains true when summing the variances over all loci that affect a given trait. The within-line variance corresponds to the original variance times $(1 - F)$. As F approaches 1, the within-line variance tends toward 0.

Now consider the between-line variance. This is the variance of the true means of lines, and would be estimated from an analysis of variance as the between-line component. For a single locus with no dominance, the mean genotypic value of a line with allele frequencies p and q is obtained as

$$M = a(p - q) = a(1 - 2q) \quad (6.21)$$

Now we have to find the variance of M . The term a is a constant, meaning that it is the same in all the lines. Hence the only random term in M is the allele frequency q . Therefore

$$\text{var}(M) = \sigma_M^2 = 4a^2\sigma_q^2 = 4a^2p_0q_0F \quad (6.22)$$

Comparing the results of (6.22) and (6.20) shows that $\sigma_M^2 = 2FV_G$. Putting the two components together leads to the total genetic variance as shown in Table 6.4.

From the last row of Table 6.4, we can see that the total additive genetic variance in a population with inbreeding corresponds to $(1 + F)V_G$ which is exactly what we wanted to show at the beginning of this chapter in equation (6.1).

Chapter 7

Additional Aspects of BLUP

This chapter introduces interesting additional aspects and special properties of BLUP-based predicted breeding values. As we have seen in chapter 4, predicted breeding values which result from solving Henderson's mixed model equations are predictions and these predictions always depend on some assumptions. These assumptions are more or less valid depending on the dataset that is analysed to produce the results. Furthermore, predicted breeding values are a function of recorded data and such data is never perfect. Therefore, we need a measure to quantify how good our predictions are. Such a measure is the **accuracy** of the predicted breeding values.

One of the reasons, BLUP is nowadays the method of choice for predicting breeding values is the fact that in the BLUP animal model all available information is used. This property can be shown by decomposing the predicted breeding values from an animal model.

7.1 Accuracy

The accuracy for a BLUP-based animal model is no longer as easily derived as with the prediction of breeding values based on own-performance or progeny records. The animal model is a linear mixed effects model containing fixed and random effects. Due to the properties of BLUP-based methods, the estimates of the fixed effects and the prediction of the random effects have minimum error variance. For the fixed effects, this error variance can be computed as

$$var(\beta - \hat{\beta}) = var(\hat{\beta})$$

because fixed effects β do not have any variance. For the random effects u the prediction error variance (PEV) does not simplify to the variance of the predicted effects \hat{u} . Random effects by their nature do have a certain variance which is part of the model specification. For a BLUP animal model the variance of the random effects u correspond to $var(u) = A * \sigma_u^2$. Any meaningful prediction \hat{u} of a random effect u should also satisfy that the variance $var(\hat{u})$ predicts $var(u)$ as closely as possible. Following this argument $var(\hat{u})$ cannot correspond to the prediction error variance. The prediction error variance $PEV(\hat{u})$ is computed as

$$PEV(\hat{u}) = var(u - \hat{u}) = var(u) + var(\hat{u}) - 2 * cov(u, \hat{u}) = var(u) - var(\hat{u})$$

Henderson found that $PEV(\hat{u})$ depends on the inverse of the coefficient matrix in the mixed model equations.

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{bmatrix}^{-1} = \begin{bmatrix} C^{11} & C^{12} \\ C^{21} & C^{22} \end{bmatrix}$$

We can state that

$$PEV(\hat{u}) = var(u - \hat{u}) = var(u) - var(\hat{u}) = C^{22} \quad (7.1)$$

For a single animal i , the prediction error variance is $PEV(\hat{u}_i) = C_{ii}^{22}$ where C_{ii}^{22} is the i -th diagonal element in the matrix C^{22} . The accuracy of \hat{u}_i is measured by the squared correlation $r_{u, \hat{u}}^2$ between true and predicted breeding value. This correlation is defined as

$$r_{u, \hat{u}} = \frac{cov(u_i, \hat{u}_i)}{\sqrt{var(u_i) * var(\hat{u}_i)}} = \frac{var(\hat{u}_i)}{\sqrt{var(u_i) * var(\hat{u}_i)}} = \sqrt{\frac{var(\hat{u}_i)}{var(u_i)}} \quad (7.2)$$

Combining equations (7.2) and (7.1) by solving both for $var(\hat{u}_i)$ leads to

$$\begin{aligned} var(\hat{u}_i) &= var(u_i) - C_{ii}^{22} \\ var(\hat{u}_i) &= r_{u, \hat{u}}^2 * var(u_i) \\ PEV(\hat{u}_i) &= C_{ii}^{22} = var(u_i) - r_{u, \hat{u}}^2 * var(u_i) = (1 - r_{u, \hat{u}}^2) * var(u_i) \end{aligned} \quad (7.3)$$

Solving equation (7.3) for $r_{u, \hat{u}}^2$ which is the measure commonly used to assign a certain level of accuracy to the predicted breeding value \hat{u}_i of a given animal i .

$$r_{u, \hat{u}}^2 = 1 - \frac{C_{ii}^{22}}{var(u_i)} = 1 - \frac{PEV(\hat{u}_i)}{var(u_i)} \quad (7.4)$$

From equation (7.4) it becomes clear that the smaller $PEV(\hat{u}_i)$ is the higher the accuracy $r_{u,\hat{u}}^2$ is. In the limit where $PEV(\hat{u}_i)$ tends to 0, the accuracy will tend to 1. Based on the definition of $PEV(\hat{u}_i)$ in (7.1), it can be seen that $PEV(\hat{u}_i)$ tends to 0, if $var(\hat{u}_i)$ tends towards $var(u_i)$. That means the better the variance $var(\hat{u}_i)$ of the predicted breeding values \hat{u}_i approximates the variance $var(u_i)$, the smaller the value for $PEV(\hat{u}_i)$ and the higher the accuracy $r_{u,\hat{u}}^2$ of the predicted breeding value \hat{u}_i will be. On the other hand, if $var(\hat{u}_i)$ tends to 0 which means the prediction of $var(u_i)$ by $var(\hat{u}_i)$ is very poor, $PEV(\hat{u}_i)$ tends to $var(u_i)$ and the accuracy $r_{u,\hat{u}}^2$ tends to its minimum which is 0.

7.2 Confidence Intervals of Predicted Breeding Values

The prediction error variance (PEV) determines the confidence interval of the predicted breeding values. The square root of PEV corresponds to the standard error of prediction (SEP).

$$SEP(\hat{u}_i) = \sqrt{PEV(\hat{u}_i)} = \sqrt{(1 - r_{u,\hat{u}}^2) * var(u_i)}$$

Assuming the predicted breeding values \hat{u} follow a normal distribution and SEP gives a measure of how much the predictions vary. For a given error probability (α) the confidence interval can be derived for probability of $1 - \alpha$. For a given genetic standard deviation σ_u of 12, an error probability of $\alpha = 0.05$ and range of accuracies, the width of the confidence intervals can be computed. The results of these interval widths are shown in Table 7.1.

Table 7.1: Widths of Confidence Intervals for Given Accuracies

Accuracy	Interval Width
0.40	36.44
0.50	33.26
0.60	29.75
0.70	25.76
0.80	21.04
0.90	14.88
0.95	10.52
0.99	4.70

For a given predicted breeding value of 100 and an accuracy of 0.99 the confidence interval is 100 ± 2.35 . The same confidence interval is also shown in Figure 7.1.

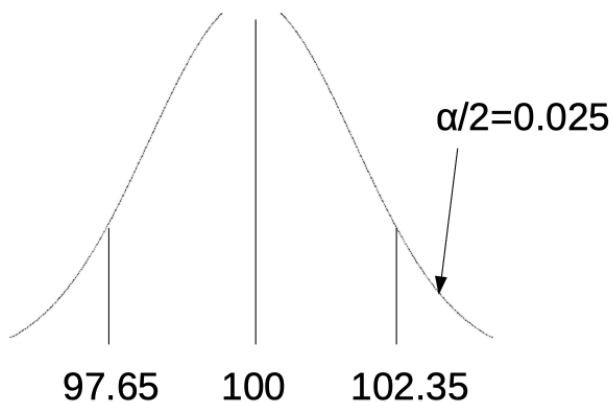


Figure 7.1: Confidence Interval of Predicted Breeding Value

7.3 Relevance of Accuracies

The relevance that is assigned to the accuracies of the predicted breeding values depends on the livestock species and also on the individual breeder. The assessment of the importance of the accuracies is not always easy and is different whether we are looking at a single animal or whether we are looking at a population.

Predicted breeding values are unbiased, hence low accuracies are not considered to be something “bad”. For single animals with predicted breeding values with low accuracies, their predicted breeding value is expected to change more. But the change of the predicted breeding values can be in both directions. Because most breeders want to avoid negative changes, high accuracies are taken to be important.

Concerning the selection response, higher accuracies are better, but these higher accuracies are not for free. The often mean that

- generation intervals increase, because we need to wait for more progeny to deliver a performance record
- more progeny per selection candidate must be tested, hence the number of selection candidates and the selection intensities decrease
- costs for testing animals increase.

For livestock species such as cattle and horses, breeders usually assign too much relevance to accuracies. In general selection response could be increased by lowering the generation interval and increasing the selection intensities and thereby accepting lower levels of accuracies.

7.4 Decomposition of Predicted Breeding Value

The mixed model equations as they are shown in (4.10) can be written in the following abbreviated form

$$M * s = r$$

where

M	coefficient matrix
s	vector of unknowns
r	vector of right-hand sides

The vector s of unknowns in the mixed model equations consists of the vector $\hat{\beta}$ of estimates of fixed effects and the vector \hat{u} of predicted breeding values, which means

$$s = \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix}$$

Because the vector $\hat{\beta}$ has length p , the first p components in s correspond to estimates of fixed effects. The remaining q components of s correspond to the q predicted breeding values of vector \hat{u} . Let us assume that we want to have a closer look at how the predicted breeding value \hat{u}_i of the animal at position i in the vector \hat{u} . The component \hat{u}_i can be found on position $p + i$ in the vector s . As a consequence of that the $(p + i)$ -th line in M contains the coefficients

that are relevant for the computation of the predicted breeding value \hat{u}_i . These coefficients determine what type of information is used to compute \hat{u}_i . In what follows, we describe how these coefficients are determined.

For the decomposition, we are using a simpler model which is shown in (7.5)

$$y_i = \mu + u_i + e_i \quad (7.5)$$

where y_i Observation for animal i
 u_i breeding value of animal i with a variance of $(1 + F_i)\sigma_u^2$
 e_i random residual effect with variance σ_e^2
 μ single fixed effect

The above defined model is used to analyse a dataset in which all animals have an observation. Animal i has parents s and d and n progeny k_j (with $j = 1, \dots, n$) and n mates l_j (with $j = 1, \dots, n$). From this it follows that progeny k_j has parents i and l_j .

For this simple model (7.5) the mixed model equations also have a reduced complexity. Because, we only have one fixed effect which is present in all observations the matrix X has just one column of all ones. Because all animals have an observation, the matrix Z corresponds to the identity matrix.

Taking into account Henderson's rule for setting up A^{-1} directly, the equation for observation y_i which corresponds to the $(i + 1)$ -th¹ equation in our mixed effects model.

$$\begin{aligned} y_i = \hat{\mu} + \left[1 + \alpha\delta^{(i)} + \frac{\alpha}{4} \sum_{j=1}^n \delta^{(k_j)} \right] \hat{u}_i - \frac{\alpha}{2} \delta^{(i)} \hat{u}_s - \frac{\alpha}{2} \delta^{(i)} \hat{u}_d \\ - \frac{\alpha}{2} \sum_{j=1}^n \delta^{(k_j)} \hat{u}_{k_j} + \frac{\alpha}{4} \sum_{j=1}^n \delta^{(k_j)} \hat{u}_{l_j} \end{aligned} \quad (7.6)$$

where α ration between variance components σ_e^2/σ_u^2
 $\delta^{(j)}$ contribution for animal j to A^{-1}

Solving (7.6) for \hat{u}_i leads to

$$\hat{u}_i = \frac{1}{1 + \alpha\delta^{(i)} + \frac{\alpha}{4} \sum_{j=1}^n \delta^{(k_j)}} \left[y_i - \hat{\mu} + \frac{\alpha}{2} \left\{ \delta^{(i)}(\hat{u}_s + \hat{u}_d) + \sum_{j=1}^n \delta^{(k_j)} \left(\hat{u}_{k_j} - \frac{1}{2} \hat{u}_{l_j} \right) \right\} \right] \quad (7.7)$$

¹For the general case, this would be $(p + i)$ -th equation. In the simple example, we have $p = 1$.

From the decomposition in (7.7), we can see that the predicted breeding value \hat{u}_i consists of the following components

- Predicted breeding values \hat{u}_s and \hat{u}_d of parents s and d of i
- Own performance y_i of i
- Predicted breeding values \hat{u}_{k_j} and \hat{u}_{l_j} of progeny k_j and mates l_j

An explicit example of a decomposition in (7.7) will be used as an exercise problem.

Chapter 8

Multiple Traits

So far we have seen how to predict breeding values using the BLUP animal model. These breeding values were always only for one trait. From a statistical point of view, experts are calling such analyses **univariate** evaluations. In reality, livestock breeders want to improve their animals in a population with respect to several traits simultaneously. If there are genetic relationships (measured by genetic correlations) between traits, univariate predictions of breeding values do not make optimal use of the available information. This effect is stronger, if certain traits can only be observed in animals of one sex. According to (Mrode, 2005) who cites (Schaeffer, 1984) and (Thompson and Meyer, 1986), the increased accuracy of the evaluations is one of the main advantage of multivariate BLUP analyses. Predictions of breeding values for several traits in a single evaluation is called **multivariate** prediction of breeding values. Such multivariate analyses can be implemented in different ways such as

- combining different corrected information sources in a multivariate selection index (no longer used).
- multivariate prediction of breeding values using BLUP animal model (method of choice).
- combining predicted breeding values from univariate analyses.

Before the introduction of the BLUP animal model, breeding values were estimated using a method that is called **selection index** method. A brief review about selection index theory is given in section 8.3. While selection indices are no longer used to estimate breeding values, selection index theory is still used to predict the aggregate genotype, as will be shown later in this chapter. Before that, we start with an introduction to multivariate BLUP methods.

8.1 Multivariate Predictions Of Breeding Values Using BLUP

The prediction of breeding values using a multivariate BLUP model is the optimal prediction procedure. It has advantages, if

- some traits have lower heritability than others
- environmental correlations exist between traits measured on the same animal
- some traits are available only a subset of all animals
- some traits were used for a first round of selection

In principle, a multivariate analysis can be thought of as several univariate analyses which are stacked one on top of the other. Let us assume that we have two traits. For the first trait, we define the model

$$y_1 = X_1\beta_1 + Z_1u_1 + e_1$$

Similarly for the second trait, we define the model

$$y_2 = X_2\beta_2 + Z_2u_2 + e_2$$

If we group the data by traits, then we can write the multivariate model as

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} Z_1 & 0 \\ 0 & Z_2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$$

The genetic variance-covariance matrix G_0 for the two traits has the following structure.

$$G_0 = \begin{bmatrix} \sigma_{g_1}^2 & \sigma_{g_1, g_2} \\ \sigma_{g_1, g_2} & \sigma_{g_2}^2 \end{bmatrix} = \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix}$$

The inverse G_0^{-1} of G_0 can be formulated as

$$G_0^{-1} = \begin{bmatrix} g^{11} & g^{12} \\ g^{21} & g^{22} \end{bmatrix}$$

For the random residual effects, the variance-covariance matrix R_0 for the two traits can be written as

$$R_0 = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix}$$

8.1. MULTIVARIATE PREDICTIONS OF BREEDING VALUES USING BLUP87

Also the inverse R_0^{-1} can be written as

$$R_0^{-1} = \begin{bmatrix} r^{11} & r^{12} \\ r^{21} & r^{22} \end{bmatrix}$$

The variance-covariance matrix for the complete set of true breeding values and for all random residuals can be written as

$$\text{var}(u) = \text{var} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} g_{11}A & g_{12}A \\ g_{21}A & g_{22}A \end{bmatrix} = G_0 \otimes A = G$$

where A is the numerator relationship matrix and \otimes denotes the so-called **Kroneckerproduct** of two matrices. The variance-covariance matrix R for the random residuals is given by

$$R = \text{var}(e) = \text{var} \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = \begin{bmatrix} r_{11}I_n & r_{12}I_n \\ r_{21}I_n & r_{22}I_n \end{bmatrix} = R_0 \otimes I_n$$

where n corresponds to the number of animals in the pedigree. The covariances between all elements of a and e are 0. This is denoted by combining both random vectors into a single vector and writing down the variance-covariance matrix of the combined vector as

$$\text{var} \begin{bmatrix} u \\ e \end{bmatrix} = \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix} = \text{var} \begin{bmatrix} u_1 \\ u_2 \\ e_1 \\ e_2 \end{bmatrix} = \begin{bmatrix} g_{11}A & g_{12}A & 0 & 0 \\ g_{21}A & g_{22}A & 0 & 0 \\ 0 & 0 & r_{11}I_n & r_{12}I_n \\ 0 & 0 & r_{21}I_n & r_{22}I_n \end{bmatrix}$$

The solutions to get estimates of fixed effects and to get predictions for breeding values are obtained from the solutions of mixed model equations. These have exactly the same structure as in the univariate case.

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

where

$$y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, X = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}, \hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}, Z = \begin{bmatrix} Z_1 & 0 \\ 0 & Z_2 \end{bmatrix}, \hat{u} = \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \end{bmatrix}$$

Based on the specification of the variance-covariance matrices described earlier, we get $R^{-1} = R_0^{-1} \otimes I_n$ and $G^{-1} = G_0^{-1} \otimes A^{-1}$.

8.2 Multitrait Selection

Now that we have predicted breeding values for a collection of traits available, the question is how to integrate these predicted breeding values into a consistent selection criterion. Selection index theory provides a tool to optimally combine different sources of information in order to approximate the aggregate genotype H . In section 8.2.3, we will return to this topic once again. Although the selection index provides an ideal framework to combine estimated breeding values into an optimal selection criterion, we are going to describe to other selection procedures that are commonly used in practical livestock breeding. The two procedures are

1. Tandem-Selection and
2. Selection based on independent selection thresholds.

The description of these procedures aims at showing the negative consequences that results from the use of these sub-optimal selection strategies.

8.2.1 Tandem-Selection

The term **Tandem-Selection** stands for the strategy of always improving the population with respect to just one trait. Once the breeding goal for this trait is reached the population is improved with respect to a different trait. This sequence of single-trait improvements is continued until the breeding goal for all traits is reached.

The problem with Tandem-Selection is that while improving the population for a given trait, the population can only realize correlated selection responses for all other traits. These correlated selection responses might be very small or can even be negative which causes the time that it takes to reach the breeding goal for all relevant traits to be very long.

8.2.2 Selection Based On Independent Selection Thresholds

This method was used before the selection index was discovered. This selection procedure is very easy to apply. First selection thresholds are defined for all traits. In the next steps, all animals which are above the thresholds for all traits are selected as parents of the next generation. With this method, selection responses for all traits can be obtained in the early generations after the implementation of this selection strategy.

We are using the following example to show how selection based on independent selection thresholds is used. For reasons of simplicity, we restrict ourselves on two traits. But the results can be generalized without any problems.

8.2.2.1 Example: Selection On Independent Thresholds

In a dairy cattle population, breeders want to improve milk yield and protein content. We assume the following selection thresholds for the two traits of interest during first lactation

- milk yield: 6900 kg
- protein content: 3.5 %

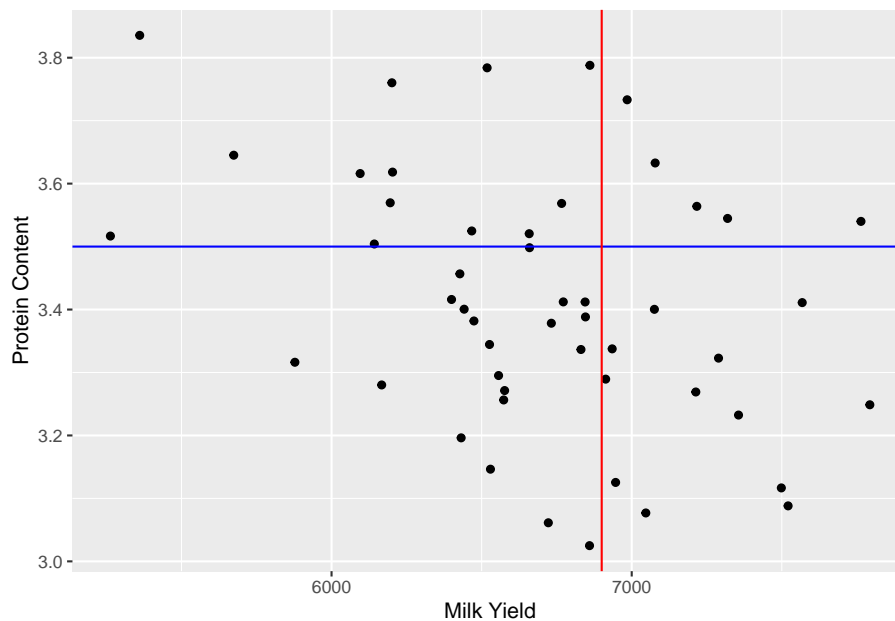


Figure 8.1: Milk Yield and Protein Content For Dairy Cows

Figure 8.1 shows the performance data for a herd of dairy cows. The selection thresholds (red line for milk yield and blue line for protein content) divide the diagram into four quadrants. None of the cows in the lower right quadrant does meet any of the selection criterion imposed by the thresholds. The cows in the upper left quadrant fulfill the requirements for protein content and the cows in the lower right quadrant fulfill the requirements for milk yield. Only the cows in the upper right quadrant fulfill the requirements for both traits.

The disadvantage of this selection strategy becomes apparent with the cows in the upper left and in the lower right quadrant. They are culled and thereby not considered as parents of the next generation even though, they have good performances in one of the traits. From a statistical genetics point of view there are three problems associated with a selection strategy that is based on independent selection thresholds

1. livestock breeders tend to put the thresholds for all traits in the range of positive predicted breeding values. This leads to an exclusion of very many animals and a dramatic reduction in genetic variability
2. genetic relationships between traits are completely ignored. These relationships must be considered when defining selection thresholds. Otherwise the expected genetic gain will not be as expected.
3. differences in the economic relevance of the different traits are completely ignored. Putting the threshold in all traits into the range of positive predicted breeding values leads to a high emphasis on traits with a high heritability. Traits with lower heritability will have only very small selection responses.

8.2.3 Selection Index

In section 8.2, we have already briefly described how we can use selection index theory to approximate the aggregate genotype H in an optimal way. Just as a reminder, the aggregate genotype H combines all economically relevant traits into a single value using a linear function of the true breeding values u and taking the economic values w as weighting factors. Given that H corresponds to the linear function

$$H = w^T u$$

and we want to approximate H by an Index I which is a linear function of all predicted breeding values \hat{u} , we can write

$$I = b^T \hat{u}$$

where b is a vector of unknown index weights. The vector b is determined using the optimality condition of minimum prediction error variance which results in

$$b = P^{-1} G w \tag{8.1}$$

where P is the variance-covariance matrix between all information sources and G is the covariance matrix between the information sources and the traits in the aggregate genotype. In case where the traits in the aggregate genotype H and in the index I are the same, the matrices P and G are defined as

$$P = \text{var}(\hat{u})$$

and

$$G = \text{cov}(u, \hat{u})$$

For predicted breeding values using BLUP, it can be shown that $cov(u, \hat{u}) = var(\hat{u})$ and therefore $P = G$. Using that equality in equation (8.1), we get

$$b = w$$

which means that the vector b of index weights corresponds to the vector of economic values w .

The use of the selection index theory to combine predicted breeding values as information sources to approximate the aggregate genotype has the following advantages

- genetic relationships (correlations) between traits in the aggregate genotype are considered correctly
- relationships between information sources in the index are considered correctly
- information from auxiliary traits can be used
- differences in economic relevance of different traits are considered correctly
- expected selection responses can be estimated and thereby quantified

Despite all these advantages, index selection alone is very rarely used in practical livestock breeding. The reason for this is that every population has a few traits that are difficult to associate with an economic value or for some traits it is difficult to come up with genetic parameters. As a consequence of that, in practical livestock breeding we will always find a mix of index selection and a variety of independent selection thresholds.

8.3 Review On Selection Index Theory

Before the introduction of the BLUP animal model ((Henderson, 1973b) and (Henderson, 1975)), breeding values were estimated using **Selection Index Theory** ((Hazel, 1943) and (Hazel and Lush, 1942)). Both methods - selection index and BLUP - are based on the same genetic model. The main difference between the two methods consists in the way how they correct for identifiable systematic environmental effects. We start with a treatment of selection index theory.

8.3.1 Introduction

In principle, prediction of breeding values aims at assessing the genetic potential of a selection candidate that is due to additive gene effects based on all available information, such that the correlation between true and predicted breeding value is maximal. Because, we want to do this for a large number of selection candidates, we can formulate our aim in a more general way. For a given population, we want to predict breeding values for all animals in the population using all available information, such that the correlation between true and predicted breeding values are maximized. An alternative objective for the prediction to the maximization of the correlation between true and predicted breeding values is the minimization of the mean squared error of the prediction. The description of the aims of our procedure to predict breeding values shows that we are dealing with two different concepts of breeding value.

1. **True breeding value** which corresponds to the sum of all additive gene-effects
2. **Predicted breeding value** which is a function of the phenotypic observations (y) that is determined by statistical methods. As a prediction it is always associated with a certain error which we want to be minimal.

The prediction of breeding values has three different objectives.

1. Selection candidates are ranked according to the predicted breeding values. Hence, it provides a criterion for selecting parents out of a pool of selection candidates
2. Predicted breeding values are used to assess the response to selection and is important for planning a breeding program
3. Predicted breeding values are one criterion that affect the price of breeding animals and the price of semen.

The definition 2.1 of the term **breeding value** has several problems when it comes to its potential usefulness for prediction.

- It is impossible to generate an infinite number of progeny before having a reliable prediction of the breeding value
- Due to the above mentioned objectives, we want to have a prediction of the breeding value available as early as possible.

- The predicted breeding value should be as accurate as possible

To address these issues, the above mentioned methods were developed. We start with the method of the selection index.

8.3.2 Selection Index Method

The selection index is a method to predict the breeding value of an animal (i) by using all available information on the animal and on its relatives. The result of the selection index method is an assignment of a numerical value (I) to each animal. All animals in the population can then be ranked according to their index value. The ranking according to the index value can be used as selection criterion. In principle the index I is defined as linear combination of all available information. This can be written as

$$I = \hat{u}_i = b_1 y_1 + b_2 y_2 + \dots + b_n y_n = b^T y \quad (8.2)$$

where b is a vector of index weights and y is a vector of information sources. Here we assume that all values in y are corrected for appropriate mean levels. The resulting index value I in (8.2) is used as the predicted breeding value \hat{u}_i . From a statistical point of view equation (8.2) corresponds to a multiple linear regression. The vector of index weights b are understood as partial regression coefficients.

8.3.3 Aggregate Genotype

In most practical livestock breeding scenarios, we want to improve a population at the genetic level with respect to more than one trait or characteristic, simultaneously. This requires a procedure that enables us to combine the breeding values of several trait into one selection criterion. This criterion is called the **aggregate genotype** H . It is defined as

$$H = w_1 u_1 + w_2 u_2 + \dots + w_m u_m = w^T u \quad (8.3)$$

where u corresponds to the vector of true breeding values and w is a vector of economic values. The economic value w_k for a given trait k is defined as the marginal change in profit caused by a small change in the population mean (μ_k) of the trait k . At this point, we are not describing how the economic values w_k are derived, but we consider them to be known. For the construction of the selection index, we are using the general form of the aggregate genotype H . Once the selection index is constructed, we can go back to the simple scenario of considering just one trait which reduces the aggregate genotype H to the true breeding value u of the single trait.

8.3.4 Theory of Index Construction

The term *index construction* stands for the computation of the vector of index weights b for a given set of information sources and a given aggregate genotype. Independently from the available information sources, the following parameters must be known

- heritabilities and phenotypic standard deviations for the traits in the aggregate genotype and for the traits in the index.
- phenotypic correlations between the traits in the index
- genetic correlations between the traits in the index and the traits in the aggregate genotype
- genetic correlations between the traits in the aggregate genotype
- economic values for the traits in the aggregate genotype

The objective of the index construction is to maximize the correlation r_{HI} between the index I and the aggregate genotype H . Because the index I corresponds to a multiple linear regression, the mean squared error between aggregate genotype and index is to be minimized. From this it follows that

$$E(H - I)^2 \rightarrow \min \quad (8.4)$$

The solution to the index construction objective in equation (8.4) leads to the so-called index normal equations which have the following form.

$$Pb = Gw \quad (8.5)$$

where P is the variance-covariance matrix between all information sources in the index, G is the genetic variance-covariance matrix between the traits in the aggregate genotype and in the index and w is a vector of known economic values. Solving for the vector of unknown index weights b leads to

$$b = P^{-1}Gw \quad (8.6)$$

The accuracy of the index is assessed by the correlation r_{HI} between the index I and the aggregate genotype H . The higher this correlation, the better the approximation of H by I . The correlation r_{HI} can be computed as shown in (8.7). The terms for $cov(H, I)$, σ_H and σ_I are taken from (8.24) and for b we insert the solution taken from (8.6).

$$\begin{aligned}
r_{HI} &= \frac{\text{cov}(H, I)}{\sigma_H \sigma_I} \\
&= \frac{w^T * G^T * b}{\sqrt{(w^T * C * w) * (b^T * P * b)}} \\
&= \frac{w^T * G^T * P^{-1} * G * w}{\sqrt{(w^T * C * w) * ((P^{-1} * G * w)^T * P * P^{-1} * G * w)}} \\
&= \frac{w^T * G^T * P^{-1} * G * w}{\sqrt{(w^T * C * w) * (w^T * G^T * P^{-1} * P * P^{-1} * G * w)}} \\
&= \frac{w^T * G^T * P^{-1} * G * w}{\sqrt{(w^T * C * w) * (w^T * G^T * P^{-1} * G * w)}} \\
&= \sqrt{\frac{w^T * G^T * P^{-1} * G * w}{w^T * C * w}} \\
&= \frac{\sigma_I}{\sigma_H}
\end{aligned} \tag{8.7}$$

The response to selection R which results from applying a selection scheme according to the index I per generation is computed as

$$\begin{aligned}
R &= i * r_{HI} * \sigma_H \\
&= i * \frac{\sigma_I}{\sigma_H} * \sigma_H \\
&= i * \sigma_I
\end{aligned} \tag{8.8}$$

where i is the selection intensity.

8.3.5 Example of Index with Own Performance

The simplest case of an index I is the one where the aggregate genotype H consists of one trait and the index I contains a single own performance record of the same trait. This is equivalent to using the index I to predicting the breeding value u of an animal based on own phenotypic own performance record y . Hence we can set

$$H = u \quad \text{and} \quad I = by^*$$

During the index construction, we have assumed the information in the index to be corrected for the appropriate population mean μ . For our example here, we can set $y^* = y - \mu$. To determine the unknown index weight b which is on our example just a single number, we have to specify P , G and w . Because,

we are looking at just one trait, the vector of economic values w is set to one. The matrix P was defined to be the variance-covariance matrix between the traits in the index. As the index I contains just one phenotypic record, then P corresponds to the phenotypic variance σ_y^2 of our trait of interest. The matrix G was defined to be the genetic variance-covariance matrix between the traits in the aggregate genotype and the traits in the index. In our example we have just one trait which is the same in H and in I , hence G corresponds to the additive genetic variance σ_u^2 . In summary, we have found that

$$\begin{aligned} P &= \sigma_y^2 \\ G &= \sigma_u^2 \\ w &= 1 \end{aligned} \tag{8.9}$$

Inserting the terms of (8.9) into equation (8.6) to compute the index weight b results in

$$\begin{aligned} b &= P^{-1} * G * w \\ &= \sigma_y^{-2} * \sigma_u^2 * 1 \\ &= \frac{\sigma_u^2}{\sigma_y^2} = h^2 \end{aligned} \tag{8.10}$$

Using the index weight b found in (8.10) to compute the index I , we get

$$\begin{aligned} I &= by^* \\ &= h^2(y - \mu) \\ &= \hat{u}_i \end{aligned} \tag{8.11}$$

The index value I that we obtained in (8.11) corresponds to the predicted breeding value for a given trait of an animal i based on an own performance phenotypic record of animal i in the respective trait. Comparing the predicted breeding value obtained in (8.11) using selection index theory to the result obtained from the regression approach in (3.5) shows that they are identical.

The accuracy r_{HI} of the predicted breeding value (\hat{u}_i) using selection index theory is computed as shown in (8.7)

$$\begin{aligned}
r_{HI} &= \frac{\sigma_I}{\sigma_H} \\
&= \frac{b\sigma_y}{\sigma_u} \\
&= \frac{h^2\sigma_y}{\sigma_u} \\
&= h
\end{aligned} \tag{8.12}$$

Similarly to the predicted breeding value, the accuracy r_{HI} that results from selection index theory is identical to what was found using the regression approach.

8.3.6 Example with Progeny Records

The prediction of breeding values for a given animal i based on progeny records is very common in livestock breeding. Examples are dairy cattle where bulls are evaluated based on lactation records of daughters. Similarly for beef cattle or pigs where sires are evaluated based on carcass performance of their progeny. For a very long time this has been the standard method to predict breeding values to select parents in a breeding program. First we assume that the progeny of animal i are all half-sibs. Before, we can use the performance records of the progeny to predict breeding values for the parents, we have to correct them with the appropriate mean performance. After the correction the progeny performance values are averaged for a given parent. These mean performance values for a given parent i are called \bar{y}_i and are used to predict the breeding values. Hence our index I for a given animal i is defined as

$$I = b\bar{y}_i \tag{8.13}$$

Because, we are only looking at a single trait, the aggregate genotype H corresponds to the single true breeding value u of this trait and the economic weight w is 1. Now we are ready to set up the index normal equations. In general these equations have the form

$$Pb = Gw \tag{8.14}$$

where P corresponds to the variance-covariance matrix of the information sources in the index. Our index I as defined in (8.13) contains just one source of information, namely the average \bar{y}_i of the progeny performance values of animal i . In general the phenotypic variance of the mean \bar{y} of n progeny performance values corresponds to

$$\sigma_y^2 = \frac{1 + (n-1)t}{n} \sigma_y^2 \quad (8.15)$$

For our case with the progeny records, t takes the value of $\frac{1}{4}h^2$. For more details on how to compute σ_y^2 , see section 8.3.8. Hence the matrix P reduces to a single number

$$P = \sigma_y^2 = \frac{1 + (n-1)h^2/4}{n} \sigma_y^2 \quad (8.16)$$

The matrix G in (8.14) is the genetic covariance matrix between the traits in H and the information sources in I . In our current example $G = \text{cov}(u_i, \bar{y}_i) = \frac{1}{2}\sigma_u^2$. For more details on how to compute G , see section 8.3.8.2. Now that we have all the components of (8.14), we can insert them and solve for b .

$$\begin{aligned} \frac{1 + (n-1)h^2/4}{n} \sigma_y^2 * b &= \frac{1}{2} \sigma_u^2 \\ b &= \frac{2nh^2}{4 + (n-1)h^2} \\ &= \frac{2n}{n+k} \end{aligned} \quad (8.17)$$

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

With this the predicted breeding value \hat{u}_i for animal i based on the average progeny performance values using the index approach corresponds to

$$\hat{u}_i = I = b * (\bar{y}_i - \mu) = \frac{2n}{n+k} * (\bar{y}_i - \mu) \quad (8.18)$$

The accuracy for the predicted breeding value in (8.18) is

$$r_{HI} = \sqrt{\frac{n}{n+k}} \quad (8.19)$$

8.3.7 Appendix: Derivation of Index Normal Equations

In this section we want to show how to derive the index normal equations from the objective criterion in the index construction procedure. The objective criterion was formulated in equation (8.4) as

$$\Psi = E(H - I)^2 \rightarrow \min \quad (8.20)$$

The derivation starts by inserting the definitions of H and I into (8.20).

$$\begin{aligned}\Psi &= E(H - I)^2 = E(H^2 - 2 * H * I + I^2) \\ &= E(H^2) - 2 * E(H * I) + E(I^2)\end{aligned}\quad (8.21)$$

Both the expected value $E(H)$ of the aggregate genotype H and the expected value $E(I)$ of the index are both 0. This can be seen by the following expansion

$$E(H) = E(w^T a) = w^T * E(u) = w^T * 0 = 0 \quad (8.22)$$

because the breeding values u are defined as deviations, their expected value $E(u)$ is always 0. Similarly for the index I , we mentioned that the components in the vector y denoting the information sources that enter the index I are corrected by suitable population means. Due to this correction, we can state that $E(y) = 0$ and thereby $E(I) = 0$. Using these results on the expected values of H and I , we can further develop (8.21)

$$\begin{aligned}\Psi &= \text{var}(H) - 2 * \text{cov}(H, I) + \text{var}(I) \\ &= \text{var}(w^T u) - 2 * \text{cov}(w^T u, b^T y) + \text{var}(b^T y) \\ &= w^T \text{var}(u) w - 2 * w^T \text{cov}(u, y^T) b + b^T \text{var}(y) b \\ &= w^T C w - 2 * w^T G^T b + b^T P b\end{aligned}\quad (8.23)$$

where C is the variance-covariance matrix of the true breeding values of the traits in the aggregated genotype, G^T is the genetic variance-covariance matrix between the traits in the aggregate genotype and the traits in the index and P is the phenotypic variance-covariance matrix between the traits in the index. Hence we can state

$$\begin{aligned}\text{var}(H) &= w^T * C * w \\ \text{cov}(H, I) &= w^T * G^T * b \\ \text{var}(I) &= b^T * P * b\end{aligned}\quad (8.24)$$

In the objective criterion in (8.20), we stated that Ψ should be minimized. This is done by computing the derivative of Ψ with respect to the vector b . The solution vector b that sets that derivative to 0 corresponds to the solution that we are looking for. The derivative of Ψ with respect to the vector b is also called the gradient and can be computed as

$$\frac{\partial \Psi}{\partial b} = 0 - 2 * w^T * G^T + 2b^T P \quad (8.25)$$

Setting (8.25) to 0 leads to

$$\begin{aligned} 0 &= -2 * w^T * G^T + 2b^T P \\ w^T G^T &= b^T P \\ Pb &= Gw \end{aligned} \tag{8.26}$$

The last line in (8.26) follows by transposing both sides of the second last line and because P is symmetric, $P^T = P$. As a result we obtain the index normal equations which can be solved for the unknown vector b by pre-multiplying both sides with the inversion matrix P^{-1} of P .

$$b = P^{-1}Gw \tag{8.27}$$

Because P is a variance-covariance matrix, it is guaranteed to be positive definite and its inverse P^{-1} does exist.

8.3.8 Appendix: Derivation of the Index Components for the Example of the Mean Progeny Performance

8.3.8.1 Variance of Mean Progeny Performance

The mean performance values of a group of progeny for a given parent has the following structure

$$\bar{y}_i = \frac{1}{n} \sum_{k=1}^n y_{i,k} \tag{8.28}$$

where y_k is the corrected performance value of progeny k of animal i . Each y_k can be decomposed into

$$\begin{aligned} y_{i,k} &= u_k + e_k \\ &= \frac{1}{2}u_i + \frac{1}{2}u_{d,k} + m_k + e_k \end{aligned} \tag{8.29}$$

The variance (σ_y^2) of a single phenotypic observation ($y_{i,k}$) of progeny k of parent i can be computed as

$$\begin{aligned}
\sigma_y^2 = \text{var}(y_{i,k}) &= \text{var}\left(\frac{1}{2}u_i + \frac{1}{2}u_{d,k} + m_k + e_k\right) \\
&= \text{var}\left(\frac{1}{2}u_i\right) + \text{var}\left(\frac{1}{2}u_{d,k}\right) + \text{var}(m_k) + \text{var}(e_k) \\
&= \frac{1}{4}\text{var}(u) + \frac{1}{4}\text{var}(u_{d,k}) + \text{var}(m_k) + \text{var}(e_k) \\
&= \frac{1}{4}\sigma_u^2 + \frac{1}{4}\text{var}(u_{d,k}) + \text{var}(m_k) + \text{var}(e_k) \tag{8.30}
\end{aligned}$$

In (8.30) we have assumed that all the pairwise covariances between the terms are 0. We define the intra-class correlation t which is the part of the total variance which is attributed to the permanent effect in the single performance records.

$$t = \frac{1/4\sigma_u^2}{\sigma_y^2} = \frac{1}{4}h^2 \tag{8.31}$$

Inserting the decomposition of (8.29) into (8.28) leads to

$$\begin{aligned}
\bar{y}_i &= \frac{1}{n} \sum_{k=1}^n y_{i,k} \\
&= \frac{1}{n} \sum_{k=1}^n \left(\frac{1}{2}u_i + \frac{1}{2}u_{d,k} + m_k + e_k \right) \\
&= \frac{1}{2}u_i + \frac{1}{n} \sum_{k=1}^n \frac{1}{2}u_{d,k} + \frac{1}{n} \sum_{k=1}^n m_k + \frac{1}{n} \sum_{k=1}^n e_k \tag{8.32}
\end{aligned}$$

Taking the variance on both sides of (8.32) leads to our final result the variance ($\sigma_{\bar{y}}^2$) of the mean progeny performance.

$$\begin{aligned}
\sigma_y^2 &= \text{var}(\bar{y}_i) = \text{var}\left(\frac{1}{2}u_i + \frac{1}{n} \sum_{k=1}^n \frac{1}{2}u_{d,k} + \frac{1}{n} \sum_{k=1}^n m_k + \frac{1}{n} \sum_{k=1}^n e_k\right) \\
&= \text{var}\left(\frac{1}{2}u_i\right) + \text{var}\left(\frac{1}{n} \sum_{k=1}^n \frac{1}{2}u_{d,k}\right) + \text{var}\left(\frac{1}{n} \sum_{k=1}^n m_k\right) + \text{var}\left(\frac{1}{n} \sum_{k=1}^n e_k\right) \\
&= \frac{1}{4}\sigma_u^2 + \frac{1}{4n}\text{var}(u_{d,k}) + \frac{1}{n}\text{var}(m_k) + \frac{1}{n}\text{var}(e_k) \\
&= \frac{1}{4}\sigma_u^2 + \frac{1}{n} \left(\frac{1}{4}\text{var}(u_{d,k}) + \text{var}(m_k) + \text{var}(e_k) \right) \\
&= t * \sigma_y^2 + \frac{1}{n}(1-t) * \sigma_y^2 \\
&= \frac{n * t + 1 - t}{n} * \sigma_y^2 \\
&= \frac{1 + (n-1)t}{n} * \sigma_y^2
\end{aligned} \tag{8.33}$$

Because, we saw earlier that $t = h^2/4$, we can insert that into (8.33) which brings us to the final result

$$\sigma_y^2 = \frac{1 + (n-1)h^2/4}{n} * \sigma_y^2 \tag{8.34}$$

8.3.8.2 Covariance between True Breeding Value and Mean Progeny Performance

The set-up of the index normal equations requires the matrix G which corresponds to the genetic covariance between the trait in the aggregate genotype and the information sources in the index. For the example with the mean progeny performance values, the matrix G is defined as

$$\begin{aligned}
G &= \text{cov}(u_i, \bar{y}_i) = \text{cov}\left(u_i, \frac{1}{n} \sum_{k=1}^n y_{i,k}\right) \\
&= \text{cov}\left(u_i, \frac{1}{2}u_i + \frac{1}{n} \sum_{k=1}^n \left[\frac{1}{2}u_{d,k} + m_k + e_k\right]\right) \\
&= \text{cov}\left(u_i, \frac{1}{2}u_i\right) \\
&= \frac{1}{2}\sigma_u^2
\end{aligned} \tag{8.35}$$

In (8.35), we have used that the covariance between u_i and all other components of $y_{i,k}$, except u_i is 0.

Chapter 9

Genomic Selection

Similarly to BLUP, the principles of **Genomic Selection** (GS) was proposed quite a while before its introduction in 2008. The first ideas of GS were presented by (Meuwissen et al., 2001b). They showed that information from genotypes of very many loci evenly spread over the complete genome can successfully be used for the purposes of livestock breeding. Because the information of the genotypes is spread over the complete genome it is often referred to as **genomic information** and from the use of this information for selection purposes the term of genomic selection was invented. The early results on GS were not considered until the paper by (Schaeffer, 2006) showed that in a cattle breeding program the introduction of GS could lead to savings in about 90% of the total costs, provided that the accuracies computed by (Meuwissen et al., 2001b) can really be achieved. After the publication of (Schaeffer, 2006) many livestock breeding organisation started to introduce procedures of GS.

9.1 Background

The single location in the genome that are considered in GS are called **markers**. When looking at the complete set of markers consisting the genomic information in a population, the so-called **Single Nucleotide Polymorphisms** (SNP) have been shown to be the most useful types of markers. These SNP correspond to differences of single bases at a given position in the genome. Based on empirical analyses of very many SNP-loci, almost all SNP just take two different states. Furthermore it is important that these SNPs are more or less evenly spread over the complete genome. Some SNPs are in coding regions and some may be placed in regions of unknown functionality. Figure 9.1 shows the distribution of SNP over the genome.

The loci that are relevant for a quantitative traits are called **Quantitative Trait Loci** (QTL). Any given SNP-Marker can only be informative for a given

Distribution of SNP-Loci

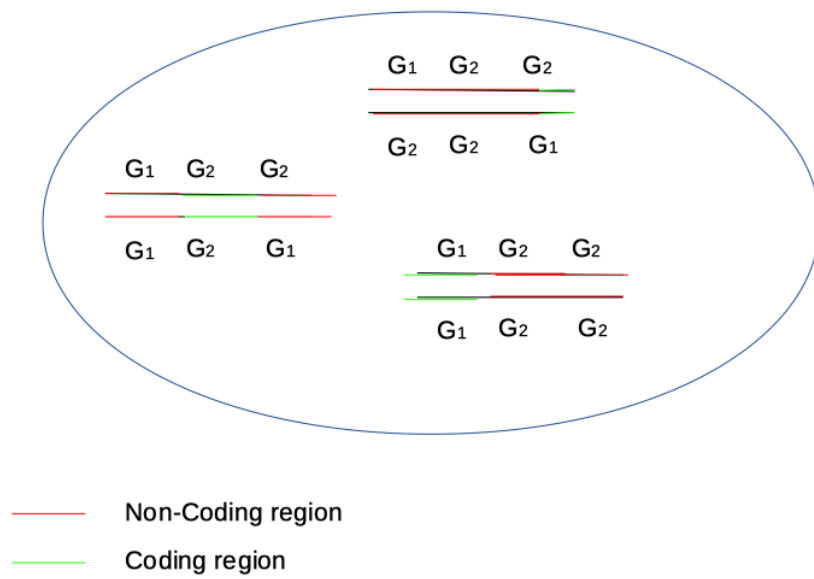


Figure 9.1: Distribution of SNP-Loci Across A Genome

QTL, if a certain **linkage disequilibrium** between the QTL and the marker locus exists. The idea behind this linkage disequilibrium is that a certain positive QTL-allele evolved in a certain genetic neighborhood of a number of SNP loci. As a result of that the positive QTL-allele is very often inherited with the same SNP-allele. Over the generations, recombination between the QTL and the neighboring SNP-loci can happen and thereby weaken the association between the positive QTL-allele and the given SNP-allele. This recombination effect is smaller when the QTL and the SNP-loci are physically closer together on the chromosome. The non-random association between QTL and SNP-markers is called linkage disequilibrium.

The marker locus is called M and the QTL is called Q , then the LD can be measured by

$$D = p(M_1Q_1) * p(M_2Q_2) - p(M_1Q_2) * p(M_2Q_1) \quad (9.1)$$

where $p(M_xQ_y)$ corresponds to the frequency of the combination of marker allele M_x and QTL allele Q_y . Very often the LD measure shown in (9.1) is re-scaled to the interval between 0 and 1 which leads to

$$r^2 = \frac{D^2}{p(M_1) * p(M_2) * p(Q_1) * p(Q_2)} \quad (9.2)$$

In (9.2) r^2 describes the proportion of the variance at the QTL which is explained by the marker M . Hence the LD must be high such that the marker can explain a large part of the variance at the QTL. For the length of most livestock species, about 50'000 SNP markers are required to get a sufficient coverage of the complete genome.

Nowadays the term **genomic selection** is often used ambiguously. What most people mean when they are talking about GS should better be called **genomic prediction** of breeding values. This prediction can be done in different ways which are listed below

- Two-step procedure: Effects of SNPs are predicted using single locus models in a reference population which corresponds of mainly male breeding animals with transformed predicted traditional BLUP-breeding values with an reliability above a certain threshold. Alternatively, it is also possible to use statistics of daughter yields as observations for the prediction of marker effects. Predictions of genomic breeding values for all animals in the population with genomic information are computed by summing up all previously estimated SNP-effects. This procedure is currently applied in the Swiss dairy cattle populations.
- Single-step procedures try to predict genomic breeding values and traditional breeding values in a single evaluation.

9.2 A Linear Model To Predict Genomic Breeding Values

A linear model to estimate SNP-effects based on the data from the reference population in the two-step procedure can be defined as follows

$$y = X\beta + Mg + e \quad (9.3)$$

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

The observations y used in (9.3) are in most evaluations not phenotypes but traditionally predicted breeding values with an reliability above a certain threshold. As a consequence of that the variance-covariance matrix (R) of the residuals e is not just an identity matrix (I) times a residual variance component (σ_e^2) but R is a diagonal matrix with elements $(R)_{ii} = \frac{1}{B_m} - 1$ where B_m is the reliability of the traditionally predicted breeding value from an animal from the reference population, corrected for the parental contributions. In effect, B_m corresponds to the reliability of the mendelian sampling term.

The mixed-model equations resulting from models given in (9.3) have the following structure

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} M \\ M^T R^{-1} X & M^T R^{-1} M + I * \lambda \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} X^T R^{-1} y \\ M^T R^{-1} y \end{bmatrix} \quad (9.4)$$

where

$$\lambda = \frac{\sigma_e^2}{\sigma_a^2} \sum_{i=1}^m 2 * p_i * (1 - p_i) \quad (9.5)$$

In (9.5) σ_a^2 is the total genetic variance and p_i is the frequency of the SNP-allele that is associated with the positive QTL-allele.

The solutions for \hat{g} from (9.4) correspond to the SNP-genotype effects. The predicted breeding value \hat{a} for any selection candidate with genomic information is then computed as

$$\hat{a} = \sum_{i=1}^m M_i \hat{g}_i \quad (9.6)$$

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

9.2.1 Matrix M

The elements in matrix M can be encoded in different ways. The results from the genotyping laboratory sends a code representing the nucleotide that can be found at a given position. For the use in the linear model we have to use a different encoding. Let us assume that at a given SNP-position, the bases G or C are observed and G corresponds to the allele with the positive effect on our trait of interest. Based on the two observed alleles, the possible genotypes are GG , GC or CC . One possible code for this SNP in the matrix M might be the number of G -Alleles which corresponds to 2, 1 and 0. Alternatively, it is also possible to use the codes 1, 0 and -1 instead which corresponds to the factors with which a is multiplied to get the genotypic values in the single locus model.

Multiplying the matrix M with its transpose M^T results in a $n \times n$ square matrix MM^T . On the diagonal of this matrix we get counts of how many alleles in each individual have a positive effect. The off-diagonal elements count how many individual share the same alleles across all SNP-positions. In contrast to the additive genetic relationship matrix A , the counts here are based on identity by state and not on identity by descent.

The problem with matrix MM^T is its dependence on the number SNP-markers. Therefore the matrix MM^T is proportional to the relationship A but it does not correspond to A directly. As a solution to that problem (VanRaden, 2008) proposed to re-scale such that allele frequencies on a given locus are expressed as to times the deviation from 0.5. This re-scaling is done with an $n \times m$ matrix P where each of the m columns corresponds to a SNP-Locus. Elements in column i of matrix P have all the same value corresponding to $2p_i - 0.5$ where p_i corresponds to the frequency of the SNP-allele associated to the positive QTL-allele at locus i .

The difference between matrices M and P is assigned to a new matrix Z

$$Z = M - P$$

Finally the matrix ZZ^T must be scaled with the sum of $2p_i(1 - p_i)$ over all SNP-loci to get to the genomic relationship matrix G .

$$G = \frac{ZZ^T}{\sum_{i=1}^m 2p_i(1 - p_i)} \quad (9.7)$$

The matrix G has similar properties as the numerator relationship matrix A . The genomic inbreeding coefficient F_j is defined as $F_j = (G)_{jj} - 1$. The genomic relationship a_{ij} between two individuals i and j corresponds to the element in matrix G divided by the square root of the diagonal elements

$$a_{ij} = \frac{G_{ij}}{\sqrt{G_{ii}G_{jj}}}$$

9.3 GBLUP

The term **GBLUP** stands for genomic BLUP and is the most widely used single-step procedure. In GBLUP genomic breed values are directly predicted without the prediction of marker effects. This can be done by including the genomic breeding values u which corresponds to the sum of all SNP-allele effects directly as a random effect in the model.

$$y = X\beta + Wu + e \quad (9.8)$$

where W is the design matrix linking genomic breeding values to observations. The mixed model equations are defined as

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} W \\ W^T R^{-1} X & W^T R^{-1} W + G^{-1} * \lambda \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X^T R^{-1} y \\ W^T R^{-1} y \end{bmatrix} \quad (9.9)$$

where G is defined as in (9.7) and λ is the same as in equation (9.5). Several authors have shown that both procedures (two-step and single step) are equivalent. From (9.8) we can see that the GBLUP model looks very similar to the animal model, except that the covariances between random effects in the animal model are based on the numerator relationship matrix and in GBLUP they are modeled via the genomic relationship matrix G . This means in the animal model the covariance between random breeding values is based on the concept of common ancestry and identity by descent. This is replaced in GBLUP by the concept of sharing the same alleles based on identity by state which is assumed to be the cause of the covariance between random genomic breeding values.

The predicted genomic breeding values \hat{u} coming out of (9.9) are referred to as **direct genomic breeding values** (DGV).

9.4 Practical Problems

The model equations (9.8) look very straight-forward, but the practical implementation can be quite complicated. The reason for these problems is the fact that compared to the total size of a population only a small fraction of all animals are genotyped and hence contribute to the genomic evaluation. On the other hand DGV do not contain all information that occur in conventional breeding values.

Because all non-genotyped offsprings of parents are ignored by GBLUP, this loss of information is even more dramatic. For the two step-procedure as long as the

reference population has a reasonable size and is not too heterogeneous, this is not a problem, we can still come up with reasonable estimates of SNP-effects. Due to the in-balanced availability of genotypic information, a procedure to combine DGV with traditional predicted breeding values was adopted. This procedure starts with predicting DGV and combining them with traditionally predicted breeding values from parents which are termed as parent averages (PA). This procedure of combining predicted breeding values from different sources is called **blending**. The problem with blending one has to be aware of is that there is a covariance between DGV and PA which must be accounted for.

A further problem is that there are different techniques to generate genotyping results. The different results also have different densities which means that they give different numbers of SNP-loci per genome. The different techniques also vary in price which is the reason that genotyping results from different technologies must be combined. Combining genotyping results with different densities of SNP-markers per genome is done with a process that is called **imputing**. This basically comes done to inferring missing SNP-genotypes on marker panels with less density based on results from denser marker panels.

9.5 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 than the traditional breeding value based only on ancestral information.

The BVM model given in (??) is a mixed linear effects model. The solution for the unknown parameters can be obtained by solving the mixed model equations shown in (9.10). 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.

$$\begin{bmatrix} X^T X & X^T Z & 0 \\ Z^T X & Z^T Z + G^{(11)} & G^{(12)} \\ 0 & G^{(21)} & G^{(22)} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{g}_1 \\ \hat{g}_2 \end{bmatrix} = \begin{bmatrix} X^T y \\ Z^T y \\ 0 \end{bmatrix} \quad (9.10)$$

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 (9.10) 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.

$$\hat{g}_2 = -(G^{22})^{-1} G^{21} \hat{g}_1 \quad (9.11)$$

Equation (9.11) is referred to as genomic regression of predicted breeding values of animals without observation on the predicted genomic breeding values of animals with observations.

Chapter 10

Variance Components

The prediction of breeding values using a BLUP animal model required the **variance components** σ_e^2 for the residual variance and σ_u^2 for the genetic additive variance to be known. For the sire model, σ_u^2 is replaced by the sire variance component σ_s^2 . In real world livestock breeding evaluations, these variance components are not known and hence must be estimated from the data. The data analysis procedure that estimates the variance components from data is called **variance components estimation**.

10.1 Sire Model

The sire model is used to motivate the introduction of the topic of variance components estimation. The sire model is given by

$$y = X\beta + Z_s s + e \quad (10.1)$$

with $var(e) = R$, $var(s) = A_s \sigma_s^2$ and $var(y) = Z_s A_s Z_s^T \sigma_s^2 + R$. The matrix A_s is the numerator relationship for sires, the sire variance component σ_s^2 corresponds to $0.25 * \sigma_u^2$ and R can often be simplified to $R = I * \sigma_e^2$. The interest in this chapter is how to estimate σ_s^2 and σ_e^2 .

In the simple case the vector β is reduced to just one scalar fixed effects parameter. This reduced X to a matrix with one column with all elements equal to 1. Assuming that we have q unrelated sires the relationship matrix A_s for the sires corresponds to the identity matrix I .

10.2 Analysis Of Variance (Anova)

As a first approach we can use an analysis of variance by fitting

1. a model with an overall effect $\beta = \mu$ and
2. a model with sire effects.

These two models give an analysis of variance of the following structure

Source	Degrees of Freedom (df)	Sums of Squares (SSQ)
Overall (μ)	$Rank(X) = 1$	$y^T X (X^T X)^{-1} X^T y = F$
Sires (s)	$Rank(Z_s) - Rank(X) = q - 1$	$y^T Z_s (Z_s^T Z_s)^{-1} Z_s^T y - y^T X (X^T X)^{-1} X^T y = S$
Residual (e)	$n - Rank(Z_s) = n - q$	$y^T y - y^T Z_s (Z_s^T Z_s)^{-1} Z_s^T y = T$
Total	n	$y^T y$

The sums of squares (SSQ) can also be expanded into sums of scalar quantities which might be easier to understand. For our sire model we get

$$F = y^T X (X^T X)^{-1} X^T y = \frac{1}{n} \left[\sum_{i=1}^n y_i \right]^2$$

where n corresponds to the number of observations in the dataset.

$$S = y^T Z_s (Z_s^T Z_s)^{-1} Z_s^T y - y^T X (X^T X)^{-1} X^T y = \sum_{i=1}^q \frac{1}{n_i} \left[\sum_{j=1}^{n_i} y_{ij} \right]^2 - F$$

where n_i corresponds to the number of observations for sire i .

$$T = y^T y - y^T Z_s (Z_s^T Z_s)^{-1} Z_s^T y = \sum_{i=1}^n y_i^2 - S - F$$

In principle effects β and s are treated as fixed effects in the above anova. If estimates of σ_e^2 and σ_s^2 are required the observed sums of squares S and T can be equated to their expected values $E(T) = (n - q)\sigma_e^2$ and $E(S) = (q - 1)\sigma_e^2 + tr(Z_s M Z_s)\sigma_s^2$ where $M = I - X(X^T X)^{-1} X^T$ and $tr(M)$ stands for the trace of matrix M which corresponds to the sum of the diagonal elements of matrix M .

10.3 Numerical Example

We want to show the estimation of variance components with a very small data set. The data that will be used is shown in the table below. The observations consist of pre-weaning weight gains of beef cattle.

Table 10.1: Small Example Dataset for Variance Components Estimation Using a Sire Model

Animal	Sire	WWG
4	2	2.9
5	1	4.0
6	3	3.5
7	2	3.5

The model used is a simplified sire model where all the fixed effect are captured by a common mean μ . Then there is the sire effect s as a random effect and the random residual effect. Hence for any given observation y_{ij} for animal i of sire j , we can write

$$y_{ij} = \mu + s_j + e_i$$

with μ the common mean, s_j the random effect of sire j ($j = 1, 2, 3$) and e_i corresponds to the random residual of observation i ($i = 1, \dots, 4$). In matrix notation this model was already given in (10.1). The design matrix X is a matrix with one column and with elements all equal to 1. The design matrix Z_s links observations to sire effects.

$$X = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, Z_s = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

An analysis of variance can be constructed as

Source	Degrees of Freedom (df)	Sums of Squares (SSQ)
Overall (μ)	$Rank(X) = 1$	$F = 48.3025$
Sires (s)	$Rank(Z_s) - Rank(X) = q - 1$	$S = 0.4275$
Residual (e)	$n - Rank(Z_s) = n - q$	$T = 0.18$

With

$$M = \begin{bmatrix} 0.75 & -0.25 & -0.25 & -0.25 \\ -0.25 & 0.75 & -0.25 & -0.25 \\ -0.25 & -0.25 & 0.75 & -0.25 \\ -0.25 & -0.25 & -0.25 & 0.75 \end{bmatrix} \text{ and } Z_s^T M Z_s = \begin{bmatrix} 0.75 & -0.5 & -0.25 \\ -0.5 & 1 & -0.5 \\ -0.25 & -0.5 & 0.75 \end{bmatrix}$$

we get the following estimates

$$\hat{\sigma}_e^2 = T = 0.18$$

$$\hat{\sigma}_s^2 = \frac{S - (q - 1)\hat{\sigma}_e^2}{\text{tr}(Z_s^T M Z_s)} = \frac{0.4275 - 2 * 0.18}{2.5} = 0.027$$

The same computations based on an anova can be done in R very easily. Assume that our dataset is in a dataframe which is called `tbl_num_ex_chp12_aov`. We are doing the anova using the function `aov()` to get the sums of squares.

```
aov_num_ex_chp12 <- aov(formula = WWG ~ Sire, data = tbl_num_ex_chp12_aov)
summary(aov_num_ex_chp12)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Sire          2  0.4275   0.2137   1.187  0.544
## Residuals     1  0.1800   0.1800
```

The results from above are obtained for $\hat{\sigma}_e^2 = 0.18$ as the value under the column `Mean Sq` in the row `Residuals`. Because in our computations above, we have considered the estimation of the overall effect which is not done in the function `aov()` in R.

10.4 Negative Estimates with Anova

One of the problems that frequently occurs when using anova to estimate variance components is that some estimates might be negative. Negative estimates are outside of the permissible range for the parameter and hence are not valid estimates. As a consequence of that alternative methods have been proposed to estimate variance components.

10.5 Likelihood-Based Approaches

The maximum likelihood (ML) approach was developed and popularized by R. A. Fisher. ML is a general approach for parameter estimation and is not only used for estimating variance components. Let us assume that our observed traits are continuous and real-valued quantities. In ML we assume that these quantities follow a certain density. This density is a function of the observed values and of unknown parameters that we want to estimate.

10.5.1 Density of Observations

Given a vector y of observations. As already mentioned, the vector y follows a certain density. As an example such a density might be a multivariate normal distribution. For a given vector y of length n , the underlying n -dimensional multivariate normal distribution has the following form

$$f_Y(y) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \exp \left\{ -\frac{1}{2} (y - \mu)^T \Sigma^{-1} (y - \mu) \right\}$$

with μ expected value of y
 Σ variance-covariance matrix of y
 $\det()$ determinant

10.5.2 Likelihood Function

As already mentioned the density is a function of the observed data y and of some unknown parameters. For the multivariate normal distribution these parameters are μ and Σ . Before observing any data, we can interpret the density $f(y|\mu, \Sigma)$ as a function of y for some fixed values of μ and Σ . But once the data has been observed, y is fixed and the parameters μ and Σ are unknown and must be estimated from the data. For the task of parameter estimation, it makes more sense to view $f(y|\mu, \Sigma)$ as a function of μ and Σ . We can write this function a little different

$$L(\mu, \Sigma) = f(y|\mu, \Sigma)$$

The function $L(\mu, \Sigma)$ is called the **Likelihood** function.

10.5.3 Maximum Likelihood

For a given dataset we choose an appropriate density which is suitable for our observations. As already mentioned, due to the Central Limit Theorem, the normal distribution is often used as a density for observations. Once, we have chosen the density, it contains unknown parameters which we have to estimate from the data. Loosely speaking, our goal is to determine the parameters such that the observed data is modeled as good as possible. This requirement is translated into a mathematical framework by the maximization of the likelihood. Hence for a given dataset our parameter estimates are determined such that the likelihood is maximized. For our multi-variate normal distribution, this can be transformed into the following equations

$$\hat{\mu} = \operatorname{argmax}_{\mu} L(\mu, \Sigma)$$

and

$$\hat{\Sigma} = \operatorname{argmax}_{\Sigma} L(\mu, \Sigma)$$

10.6 Summary

The topic of variance component estimation is a huge area. We have just covered two possible approaches to get estimates of variance components. There are many more of them. The coverage of these methods is outside of the scope of this course.

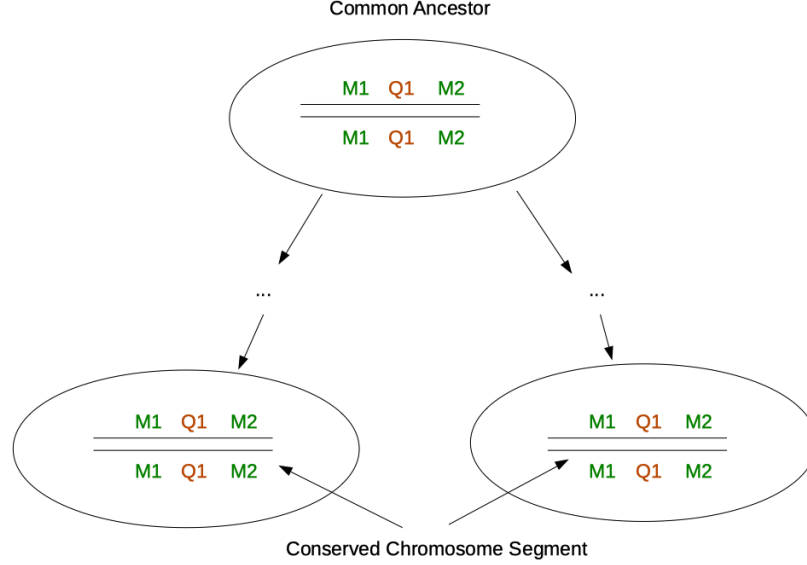
Chapter 11

Genome-Wide Association Studies (GWAS)

This chapter is based on chapter 6 of (Gondro et al., 2013). As such it provides a summary of some of the statistical methods used for genome-wide association studies (GWAS).

11.1 Single Marker Regression Tests

GWAS use linkage disequilibrium which correspond to associations of markers to causative mutations of quantitative trait loci. These associations are only expected to hold at the population level. They arise from small chromosomal segments that are inherited from a common ancestor. These chromosome segments which trace back to a common ancestor without any intervening recombination will carry identical marker alleles or marker haplotypes. If there is a QTL somewhere inside of such marker segments, they will also carry the same QTL allele. There are a number of statistical methods that use these associations to find locations of interesting QTL. A simple method is the single marker regression test.



In a random mating population without population substructures, the association between a marker and a QTL that is relevant for the expression of a phenotypic value of an economically important trait can be tested with a single marker regression as

$$y = Wb + Xg + e \quad (11.1)$$

where y is a vector of phenotypes, b is a vector of fixed effects, g is the marker effect and e is a vector of random error terms. These error terms are all identically and independently distributed with $e_{ij} \sim \mathcal{N}(0, \sigma_e^2)$ where σ_e^2 corresponds to the error variance. The design matrix W links observations to fixed effects and the matrix X allocates records to the marker effect.

In this model the marker effect is treated as fixed and the model is additive which means that two copies of the same allele have twice the effect of a single marker allele and zero alleles have no effect at all. The underlying assumption is that a given marker will only affect the phenotypic observation of a trait if it is linked to an unobservable QTL.

The null hypothesis (H_0) is that the marker does not have an effect on the trait while the alternative hypothesis (H_A) is that the marker does have an effect on the trait. The null hypothesis is rejected if the test statistic F satisfies the condition $F > F_{\alpha, \nu_1, \nu_2}$ where F_{α, ν_1, ν_2} is the value of the F -distribution at significance level α and ν_1 and ν_2 degrees of freedom.

11.1.1 Example

Consider the following example dataset.

Table 11.1: Phenotypic and genotypic data for ten animals and one marker locus

Animal	Phenotype	SNP Allele 1	SNP Allele 2
1	2.03	1	1
2	3.54	1	2
3	3.83	1	2
4	4.87	2	2
5	3.41	1	2
6	2.34	1	1
7	2.65	1	1
8	3.76	1	2
9	3.69	1	2
10	3.69	1	2

We need a design matrix X to allocate both the mean and SNP alleles to phenotypes. In this case we will use an X matrix with number of rows equal to the number of observations and one column for the SNP effect. We will set the effect of the “1” allele to 0 which means that allele “2” is the allele with the positive effect on the phenotype. So the SNP effect column is the number of copies of the “2” allele. We assume a common mean μ as the only fixed effect. Hence the matrices X and W have the following structure.

$$W = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

$$X = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \\ 2 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

The general mean and the SNP effect can be estimated as

$$\begin{bmatrix} \hat{\mu} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} W^T W & W^T X \\ X^T W & X^T X \end{bmatrix}^{-1} \begin{bmatrix} W^T y \\ X^T y \end{bmatrix} = \begin{bmatrix} 2.35 \\ 1.28 \end{bmatrix}$$

The F -value can be computed as

$$F = \frac{(n-1)(\hat{g}X^T y - 1/ny^T y)}{y^T y - \hat{g}X^T y - \hat{u}1_n^T y} = 4.56$$

The tabulated value for $F_{0.05,1,9} = 5.12$ for a significance level $\alpha = 0.05$ and $\nu_1 = 1$ and $\nu_2 = 9$ degrees of freedom. Hence for this small dataset the null hypothesis of the SNP having no effect on the trait cannot be rejected.

11.2 Genome-Wide Association Experiments Using Haplotypes

Instead of using single markers, haplotypes of markers could be used in genome-wide associations. In this context, the term “haplotype” stands for a group of consecutive markers on the same chromosome. The effect of haplotypes in windows across the genome would be tested for their association with phenotype. The justification for using haplotypes is that marker haplotypes may be in greater linkage disequilibrium with the QTL alleles than single markers. If this is true, then the r^2 between the QTL and the haplotypes is increased, thereby increasing the power of the experiment.

¹Note r^2 is defined as $r^2 = \frac{(f(A1B1)f(A2B2) - f(A1B2)f(A2B1))^2}{(f(A1)f(A2)f(B1)f(B2))}$ and measures how closely the two loci A and B are linked.

11.3 Fitting All Markers Simultaneously

There are two disadvantages of the approaches described above that fit single SNPs, haplotypes, or single genome regions in the analysis. One of these is the multiple testing problem, that is many thousands of tests are run, so the significance level must be very stringent to take this into account. Further, the setting of a significance threshold combined with the testing of so many marker effects means that the markers most likely to exceed the threshold are those with favorable error terms, so that the significant markers have overestimated effects. The second disadvantage, particularly of the single SNP approach, is that a region containing the true mutation can be hard to define, as a large number of SNP can be in LD with the QTL, such that significant SNP span a wide region. This is particularly problematic in livestock (and likely some plant species), as low, but non zero, LD extends for Mb. While a partial solution to this second problem is to jointly fit SNP in multiple or conditional regression, an even better solution to both these issues is to fit all SNP simultaneously. This involves fitting the same models that have been proposed for genomic prediction.

This can be achieved by fitting the SNPs as random effects (e.g., derived from a distribution), with different prior assumptions on the distribution of possible SNP effects (e.g., a Bayesian approach). The model is:

$$y = 1^T \mu + Xg + e$$

where g is now a vector of random SNP-effects. Because the above equation consists of a linear mixed-effect model, the solutions can be obtained by the well-known mixed-model equations.

Appendix A

Introduction To Linear Algebra

Linear Algebra is a large area from which we only need the following three topics

1. Vectors
2. Matrices and
3. Systems of linear equations.

A.1 Glimpse Ahead

The central topic of this course is the prediction of breeding values. Most approaches to predict breeding values require the solution of large systems of linear equations. These systems of equations are written down using vectors and matrices. Hence the three mentioned topics are important to understand at a level that they can be used as tools for the prediction of breeding values.

A.2 Vectors

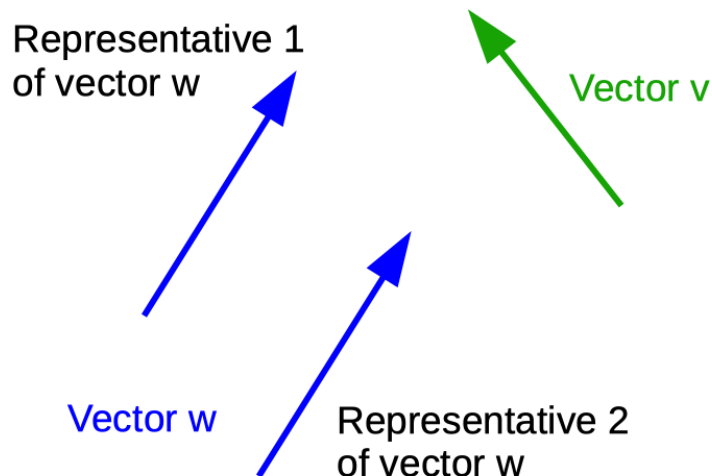
The material of this section is largely based on the video tutorial (https://youtu.be/fNk_zzaMoSs) from (3blue1brown, 2016). We try to give a summarized transcript of the video. The vector is the fundamental building block of linear algebra. There are three different but related concepts about what vectors are. We call them

1. the physics perspective
2. the computer science perspective and
3. the mathematics perspective.

The mathematics perspective tries to provide a very general concept, saying that anything can be a vector as long as, one can add two vectors or a vector can be multiplied by a factor and the result of both operations is a vector again. For what we want to use vectors for in the context of livestock breeding and genomics, the mathematics perspective is not so useful, hence we ignore it from now on.

A.2.1 Physics Perspective



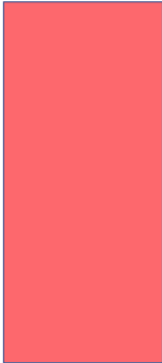
The physics perspective is that vectors are arrows with a certain **length** and a **direction** they are pointing to. As long as length and direction are the same, the arrows can be moved around and they are still the same vector. Different arrows with the same length and the same direction are called **representatives** of the same vector. Vectors that are in a flat plane are called two-dimensional. Those who are sitting in the same Euclidean space that we are all living in, are called three-dimensional.



A.2.2 Computer Science Perspective

In the computer science perspective vectors are ordered list of numbers. Later we will see that vectors can also contain more general objects like strings. As an example, we assume that we are analyzing carcasses and the only thing

we know about a carcass is its slaughter-weight (SW) and its price (P). The different carcasses can then be represented by a pair of numbers the first being the slaughter-weight and the second being the price. It is important to note here, that the order of the number matters. In terms of vectors, here each carcass is represented by a two-dimensional vector.

	Carcass 1	Carcass 2	Carcass 3
			
Weight	290 kg	265 kg	320 kg
Price	1943 Fr	1829 Fr	1984 Fr

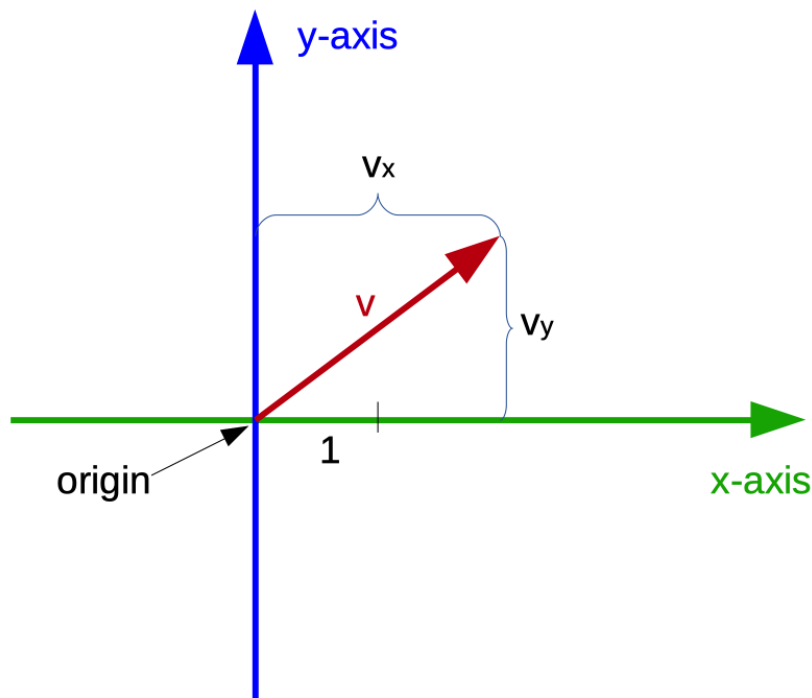
A.2.3 Geometric Context

Some basic properties of vectors are introduced using the geometric context, that a vector is an arrow located in a certain coordinate system with its tail sitting at the origin of the coordinate system. This is a little bit different from the physics perspective (see A.2.1) where the arrow can sit anywhere in space. In linear algebra it is almost always the case that vectors are rooted at the origin. Once we understand the properties of vectors in the context of arrows in space, we can then translate these properties to the list-of-numbers point of view (see A.2.2) considering the coordinates of the vectors.

A.2.4 Coordinate System

It is important to introduce the coordinate system, because this will be the basis of the correspondence between the two perspectives of linear algebra. For the moment, we focus on two dimensions. The horizontal line is called the x-axis and

the vertical line is called the y-axis. The place where the two lines intersect is called the origin. An arbitrary length is chosen to represent 1. The coordinates of a vector is a pair of numbers that give instructions for how to get from the tail of that vector at the origin to its tip. The first number tells you how far to walk along the x-axis (positive numbers indicating rightward motion, negative numbers indicating leftward motion) and the second number tell you how far to walk parallel to the y-axis (positive numbers indicating upward motion, negative numbers indicating downward motion).



A.2.5 Vector Operations

The vectors by themselves can be pretty interesting objects, but they get really useful when considering some operations that we can perform on them. Here we consider three basic operations.

1. addition
2. multiplication by a scalar number and
3. dot product

A.2.5.1 Addition

Let us assume, we have two vectors v and w . To add these two vectors, move the second one such that its tail sits at the tip of the first one. Then draw a

new vector from the tail of the first one to the tip of the second one. The new vector corresponds to the sum of the two vectors (Figure A.1).

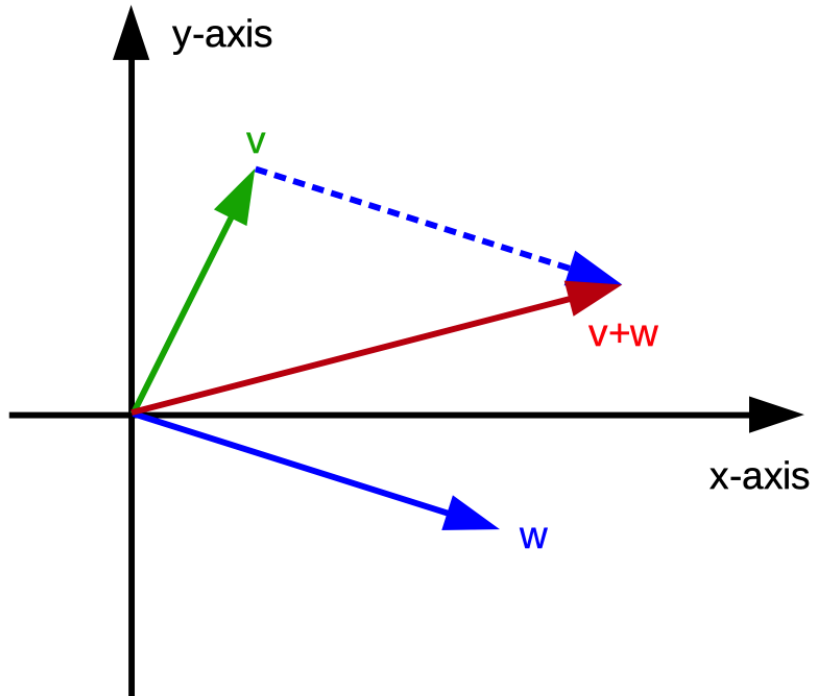


Figure A.1: Addition of two vectors

Numerically, vector addition corresponds to summing up each of the coordinates individually. Hence if we have two vectors v and w with their coordinates given as

$$v = \begin{bmatrix} v_x \\ v_y \end{bmatrix}, w = \begin{bmatrix} w_x \\ w_y \end{bmatrix}$$

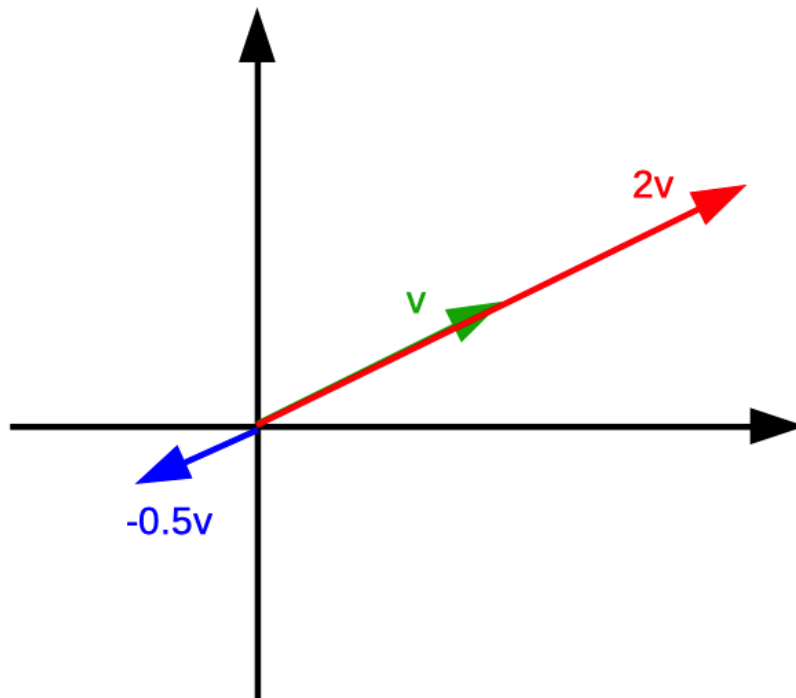
then the sum $v + w$ has coordinates

$$v + w = \begin{bmatrix} v_x + w_x \\ v_y + w_y \end{bmatrix}$$

A.2.5.2 Multiplication by a Scalar Number

This operation is best understood by looking at a few examples. If we take the number 2 and multiply it by a certain vector v , this means that we stretch out

the vector v such that it is 2 times as long as the original vector. Multiplication of a vector with positive numbers does not change the direction of the vector. Multiplying a vector v with a negative number like -0.5 then the direction gets flipped around and then squished by 0.5.



The operation of multiplying a vector by a given number, like 2 or -0.5 is also called **scaling** and that is the reason why in linear algebra the numbers like 2 and -0.5 are called **scalar** numbers or just scalars. Numerically, stretching a vector by a given number like 2, corresponds to multiplying each of the coordinate components by that factor 2. For a vector v with coordinate components v_x and v_y , the vector $2v$ has coordinates $2v_x$ and $2v_y$

$$v = \begin{bmatrix} v_x \\ v_y \end{bmatrix}, \quad 2v = \begin{bmatrix} 2v_x \\ 2v_y \end{bmatrix}$$

A.2.5.3 Dot Product

The dot product is explained in a different video that can be seen on <https://youtu.be/LyGKycYT2v0>. Numerically, if you have two vectors of the same dimension, meaning two lists of numbers of the same length, e.g. v and w then their dot product $v \cdot w$ can be computed by pairing up all of the coordinates, multiplying these pairs together and adding the result. So the vectors

$$v = \begin{bmatrix} v_x \\ v_y \end{bmatrix} \quad \text{and} \quad w = \begin{bmatrix} w_x \\ w_y \end{bmatrix}$$

their dot product $v \cdot w$ then is computed as

$$v \cdot w = v_x * w_x + v_y * w_y$$

A.3 Matrices

The introduction to the topic of matrices is available from <https://youtu.be/kYB8IZa5AuE> and <https://youtu.be/XkY2DOUCWMU>. An $m \times n$ matrix is a table-like object of $m * n$ numbers arranged in m rows and n columns. In general the $m \times n$ matrix A has the following structure.

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

The $m * n$ numbers inside of the square brackets are called elements of the matrix. The element of matrix A that is in row i and in column j is called a_{ij} or $(A)_{ij}$. As an example

$$A = \begin{bmatrix} 2 & 3 & 1 \\ 5 & 1 & 2 \end{bmatrix}$$

is a 2×3 matrix. In the first row the second element corresponds to $(A)_{12} = a_{12} = 3$. An $n \times n$ matrix (i.e. a matrix with equal numbers of rows and columns) is called a **quadratic** matrix. Two matrices A and B are called **equal**, if they have the same number of rows and columns and if the corresponding elements are the same, i.e.

$$(A)_{ij} = (B)_{ij} \text{ for all } i \text{ and } j$$

A.3.1 Special Matrices

The following matrices are special and are used in special cases.

- **Nullmatrix:** The $m \times n$ matrix 0 is called Nullmatrix, if each element is equal to zero.
- **Upper Triangular Matrix:** The square matrix R is called upper triangular matrix, if $(R)_{ij} = 0$ for $i > j$.
- **Lower Triangular Matrix:** The square matrix L is called lower triangular matrix, if $(L)_{ij} = 0$ for $i < j$.

- **Diagonal Matrix:** The square matrix D is called diagonal matrix, if $(D)_{ij} = 0$ for $i \neq j$.
- **Identity Matrix:** The diagonal matrix I is called identity matrix, if all diagonal elements $(I)_{ii} = 1$.
- **Column Vector:** A $m \times 1$ matrix is often called a column vector.
- **Row Vector:** A $1 \times n$ matrix is often called a row vector.

A.3.2 Matrix Operations

The following operations with matrices are defined.

A.3.2.1 Addition

For two $m \times n$ matrices A and B , their sum $A + B$ is again a $m \times n$ matrix with each element corresponding to the sum of the corresponding elements from A and B . Hence, we can write

$$(A + B)_{ij} = (A)_{ij} + (B)_{ij} \text{ for all } i \text{ and } j$$

A.3.2.2 Multiplication with a Number

A $m \times n$ matrix A is multiplied by a number α by multiplying every element $(A)_{ij}$ of A with α . The result $\alpha * A$ is computed as $(\alpha * A)_{ij} = \alpha * (A)_{ij}$ for all i and j .

A.3.2.3 Multiplication of two Matrices

Given a $m \times n$ matrix A and a $n \times p$ matrix B , their matrix product AB is a $m \times p$ matrix with

$$(AB)_{ij} = \sum_{k=1}^n (A)_{ik} * (B)_{kj} = (A)_{i1} * (B)_{1j} + (A)_{i2} * (B)_{2j} + \dots + (A)_{in} * (B)_{nj}$$

A.3.2.4 Laws of Matrix Operations

- **Commutativity:** For two $m \times n$ matrices A and B the addition is commutative, i.e. $A + B = B + A$.
- **Associativity of addition:** For $m \times n$ matrices A , B and C , the addition is associative, i.e., $A + (B + C) = (A + B) + C$
- **Associativity of multiplication:** For a $m \times n$ matrix A , a $n \times p$ matrix B and a $p \times q$ matrix C , the multiplication is associative, i.e., $A(BC) = (AB)C$
- **Distributivity:** For $m \times n$ matrices A and B and $n \times p$ matrices C and D , the distributive law holds, i.e., $(A + B)C = AC + BC$ and $A(C + D) = AC + AD$

A.3.2.5 Matrix Transpose

Given a $m \times n$ matrix A , then the $n \times m$ matrix A^T is called its **transpose**, if $(A^T)_{ij} = A_{ji}$. The matrix A is called **symmetric**, if $A = A^T$. For every matrix A the transpose of the transpose is the matrix itself, i.e., $(A^T)^T = A$. For any $m \times n$ matrices A and B , the transpose $(A + B)^T$ of their sum $(A + B)$ is computed as

$$(A + B)^T = A^T + B^T$$

For every $m \times n$ matrix A and every $n \times p$ matrix B , it holds that

$$(AB)^T = B^T A^T$$

A.3.2.6 Inverse of a Matrix

In this section, we are looking at square matrices. The **inverse** X of a square matrix A is defined as the square matrix that satisfies the condition $AX = I$. If the inverse matrix X exists, then the matrix A is called invertible. If X does not exist, A is called singular. If the inverse of a matrix A exists, it is uniquely determined and we call it A^{-1} .

Let us assume two invertible $n \times n$ matrices A and B , then the following equations hold

1. $A^{-1}A = I$
2. A^{-1} is invertible and $(A^{-1})^{-1} = A$
3. I is invertible and $I^{-1} = I$
4. AB is invertible and $(AB)^{-1} = B^{-1}A^{-1}$
5. A^T is invertible and $(A^T)^{-1} = (A^{-1})^T$

For every square matrix A , the following statements are equivalent.

1. A is invertible
2. The system of equations $Ax = b$ is solvable for every b .
3. The system of equations $Ax = 0$ has only the trivial solution $x = 0$.

A.3.2.7 Orthogonal Matrices

A square matrix A is called **orthogonal**, if the condition $A^T A = I$ holds. For two orthogonal matrices A and B , the following statements hold.

1. A is invertible and $A^{-1} = A^T$
2. A^{-1} is orthogonal
3. AB is orthogonal
4. I is orthogonal

A.4 Systems Of Equations

Systems of linear equations are introduced based on (Nipp and Stoffer, 2002) and (Searle, 1971). Solving systems of linear equations is one of the fundamental tasks of linear algebra. We start with a general example of a system of linear equations which is given as

$$\begin{aligned}x_1 + 2x_2 &= 5 \\ 2x_1 + 3x_2 &= 8\end{aligned}\tag{A.1}$$

In (A.1) we are given a system of linear equations with two equations and two unknowns x_1 and x_2 . The aim is to find numeric values for x_1 and x_2 such that both equations are satisfied. Inserting the values $x_1 = 1$ and $x_2 = 2$ into the above equations show that they are both satisfied. Hence the set $L = \{x_1 = 1, x_2 = 2\}$ consisting of the values for x_1 and x_2 that satisfy both equations is called a solution or a solution set for the above shown equations.

In general, a linear system of equations consists of m equations and n unknowns. In the example (A.1), $m = 2$ and $n = 2$.

The example in (A.2) does not have any solutions.

$$\begin{aligned}x_1 + x_2 &= 4 \\ 2x_1 + 2x_2 &= 5\end{aligned}\tag{A.2}$$

This can be seen, that if the first equation in (A.2) is multiplied by 2, we get $2x_1 + 2x_2 = 8$ which contradicts the second equation shown in (A.2).

A system with $m = 2$ equations and $n = 3$ unknowns is shown in (A.3).

$$\begin{aligned}x_1 - x_2 + x_3 &= 2 \\ 2x_1 + x_2 - x_3 &= 4\end{aligned}\tag{A.3}$$

There are infinitely many solutions consisting of $x_1 = 2$, $x_2 = \alpha$ and $x_3 = \alpha$ for any real number α .

The examples in (A.1), (A.2) and (A.3) already show all possible cases that may occur when solving linear systems of equations. The question is how to determine the set of all solutions of a system of linear equations.

A.4.1 Matrix-Vector Notation

So far, we have written systems of linear equations explicitly in the sense that every equation was written on one line. For small systems this is not a problem. But when the number of equations (m) and the number of unknowns (n) get very large, the explicit notation is no longer feasible. Hence, we need a notation that can also be used for large systems of equations. The so-called matrix-vector notation provides an efficient way to write down large systems of equations very efficiently.

We return to the example given by (A.1) and we define the matrix A to be

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix},$$

the vector x to be

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix},$$

and the vector y to be

$$y = \begin{bmatrix} 5 \\ 8 \end{bmatrix},$$

With these definitions, we can write the system of equations given in (A.1) using matrix-vector notation as

$$A \cdot x = y \tag{A.4}$$

A.5 Solving Systems of Linear Equations

If matrix A in (A.4) is not singular, i.e. the inverse Matrix A^{-1} of A does exist, the solution x to (A.4) can be written as $x = A^{-1}y$. This result is obtained by pre-multiplying both sides of (A.4) with A^{-1} and since a matrix times its inverse results in the identity matrix I , the solution is obtained as

$$\begin{aligned} A \cdot x &= y \\ A^{-1} \cdot A \cdot x &= A^{-1} \cdot y \\ I \cdot x &= A^{-1} \cdot y \\ x &= A^{-1} \cdot y \end{aligned} \tag{A.5}$$

For systems of equations with a singular matrix A , solutions can be found, if the equations are **consistent**. The linear equations $Ax = y$ are consistent, if any linear relationship existing among the rows of A also exist among the corresponding elements of y . As a simple example, the equations

$$\begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 7 \\ 21 \end{bmatrix}$$

are consistent. In the matrix on the left the second row corresponds to three times the first row and in the vector on the right, the second element is also three times the first element. In contrast the equations

$$\begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 7 \\ 24 \end{bmatrix}$$

are not consistent. From this example, we can already see that non-consistent equations do not have any solutions. But consistent equations $Ax = y$ have a solution which can be written as $x = Gy$ if and only if, $AGA = A$ which means that G is a so-called generalized inverse of A . The matrix G is often written as A^- . The proof of this statement is given on page 9 of (Searle, 1971).

Appendix B

Introduction To R and RStudio

According to the website <https://www.r-project.org/>, R (R Core Team, 2018) is a free software environment for statistical computing and graphics. Because R is free and it is available for the most commonly used operating systems such as Windows, MacOSX and Linux, it has become very popular in statistics and in data science. Furthermore, R can be extended with user-contributed code and documentation (called R-packages) in a very easy and standardised way. The number of available R-packages is growing rapidly and has reached

RStudio (<https://www.rstudio.com/>) is a private company that among a large number of different products distributes the RStudio Integrated Development Environment (IDE) for R. A great number of different resources about R and RStudio IDE is available. Some of them are listed here.

- **R:** From the R website, there is an introduction to R available at: <https://cran.r-project.org/doc/manuals/r-release/R-intro.html>.
- **RStudio IDE:** The link <https://resources.rstudio.com/> points to a series of webinars introducing R and RStudio.

Bibliography

3blue1brown (2016). What are vectors.

Arndorfer, M., Blauensteiner, P., Berger, B., Falschlunger, G., Fischerleitner, F., Frohofer, W., Gaber, R., Holzner, W., Kajtna, B., Koller, B., Kriebbaum, M., Machatschek, M., Neunteufel, M., Spitzenberger, F., Michael, S., Stekovics, E., Vogl, C., Vogl-Lukasser, B., Wallner, R. M., Wilfing, A., and Zipser, P. (2010). *Wie viele Arten braucht der Mensch?* Böhlau, Wien, Köln, Weimar.

Bayes, T. and Price, R. (1763). An Essay towards solving a Problem in the Doctrine of Chance. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton. *Philosophical Transactions of the Royal Society of London*, 53:370—418.

Bulmer, M. G. (1971). The effect of selection on genetic variability. *Am Nat*, 105(943):201–211.

Duerst, J. U. (1931). *Grundlagen der Rinderzucht*. Springer, Berlin; Heidelberg.

Falconer, D. S. and Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics*. Addison Wesley Longman Limited, Essex, 4 edition.

Gianola, D. and Fernando, R. L. (1986). Bayesian Methods in Animal Breeding Theory. *Journal of Animal Science*, 63(1):217–244.

Gianola, D. and Foulley, J. L. (1982). Non linear prediction of latent genetic liability with binary expression: An empirical Bayes approach. *Proc. 2nd World Congr. Genet. Appl. Livest. Prod., Madrid, Spain*, 7:293—303.

Gianola, D. and Rosa, G. J. M. (2015). One Hundred Years of Statistical Developments in Animal Breeding. *Annual Review of Animal Biosciences*, 3:13.1—13.38.

Gondro, C., van der Werf, J., and Hayes, B., editors (2013). *Genome-Wide Association Studies and Genomic Prediction*, volume 1019 of *Methods in Molecular Biology*. Humana Press, Totowa, NJ.

Hazel, L. N. (1943). The genetic basis for constructing selection indexes. *Genetics*, 28(6):476–490.

- Hazel, L. N. and Lush, J. L. (1942). The efficiency of three methods of selection. *Journal of Heredity*, 33(11):393–399.
- Henderson, C. R. (1953). Estimation of variance and covariance components. *Biometrics*, 9:226—252.
- Henderson, C. R. (1973a). Sire evaluation and genetics trends. In *Proceedings of Animal Breeding and Genetic Symposium in Honor of Dr Jay Lush*, volume 1, pages 10–41.
- Henderson, C. R. (1973b). Sire evaluation and genetics trends. *Proceedings of Animal Breeding and Genetic Symposium in Honor of Dr Jay Lush. American Society*, 1:10–41.
- Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31:423—449.
- Henderson, C. R. (1976). A simple method for computing the inverse of a numerator relationship matrix used in predicting of breeding values. *Biometrics*, 32:69—83.
- Meuwissen, T. H. and Sonesson, A. K. (1997). Maximizing the Response of Selection with a Predefined Rate of Inbreeding: Overlapping Generations. *Journal of Animal Science*, 76(10):2575–2583.
- Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001a). Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics*, 157:1819–1829.
- Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001b). Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics*, 157(157):1819–1829.
- Moore, G. E. (1965). Cramming more components onto integrated circuits. *Electronics*, 38.
- Mrode, R. A. (2005). *Linear Models for the Prediction of Animal Breeding Values*. CABI Publishing, Wallingford, Oxfordshire OX10 8DE, 2nd edition.
- Nipp, K. and Stoffer, D. (2002). *Lineare Algebra*. vdf Hochschulverlag, Zurich, 5 edition.
- Quaas, R. L. (1976). Computing the diagonal elements and inverse of a large numerator relationship matrix. *Biometrics*, 32:949–953.
- R Core Team (2018). R: A language and environment for statistical computing.
- Schaeffer, L. R. (1984). Sire and cow evaluation under multiple trait models. *Journal of Dairy Science*, 67:1567—1580.
- Schaeffer, L. R. (2006). Strategy for applying genome-wide selection in dairy cattle. *J Anim Breed Genet*, 123(123):218–223.

- Schaeffer, L. R. (2013). *History of Genetic Evaluation Methods in Dairy Cattle*. self.
- Searle, S. R. (1971). *Linear Models*. John Wiley & Sons, New York, Wiley classic edition.
- Thompson, R. and Meyer, K. (1986). A review of theoretical aspects in the estimation of breeding values for multi-trait selection. *Livestock Production Science*, 15:299—313.
- VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.*, 91(91):4414—4423.
- Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Miklos, G. L. G., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanagan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Francesco, V. D., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Hu, J., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferreira, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratt, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Abril, R. G., Campbell, M. J., Blick, L., Caminha, M., Carnes-stine, J., Caulk, P., Chai, Y., Chiang, Y., Coyne, M., Dahlke, C., Mays, A. D., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D.,

Majoros, W., Mcdaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001). The Sequence of the Human Genome. *Science*, 291(5507):1–38.