

Applied Statistical Methods in Animal Sciences

Peter von Rohr

2021-02-19

Contents

Preface	5
General Developments	5
Where Does This Course Fit In?	6
Course Objectives	6
Prerequisites	7
1 Introduction	9
1.1 US Presidential Campaigns	9
1.2 Health Care	9
1.3 Face Recognition	9
1.4 Feed Intake and Behavior Traits of Cows	11
1.5 Conclusions from Examples	11
1.6 Traditional Livestock Breeding	11
1.7 Genomic Selection	11
1.8 Mono-Genic Model	15
1.9 Two Step Approach	16
1.10 Single Step Approach	17
1.11 Summary	17
2 Fixed Linear Effects Models	19
2.1 Resources	19
2.2 Motivation	19
2.3 Data	19
2.4 Model	21
2.5 Definition of FLEM	23
2.6 Parameter Estimation Using Least Squares	25
2.7 Different Types of Linear Regressions	28
2.8 Predictions	29
2.9 Regression On Dummy Variables	30
3 Genomic Best Linear Unbiased Prediction (GBLUP)	39
3.1 Finding Relevant SNP Loci	39
3.2 Stepwise Approach	40
3.3 Model Selection With Genomic Data	41

3.4	Mixed Linear Effects Model	42
3.5	Genomic Relationship Matrix	45
3.6	How Does GBLUP Work	48
4	Least Absolute Shrinkage And Selection Operator (LASSO)	49
4.1	Stochastic Error Component	49
4.2	Parameter Estimation	50
4.3	Alternatives To Least Squares	51
4.4	LASSO	51
4.5	Determine λ	53
A	Introduction To Linear Algebra	55
A.1	Glimpse Ahead	55
A.2	Vectors	55
A.3	Matrices	62
A.4	Systems Of Equations	65
A.5	Solving Systems of Linear Equations	67
B	Basics in Quantitative Genetics	69
B.1	Single Locus - Quantitative Trait	69
B.2	Frequencies	71
B.3	Hardy-Weinberg Equilibrium	71
B.4	Value and Mean	72
B.5	Variances	79
B.6	Extension To More Loci	81
B.7	Genetic Models	82
B.8	Appendix: Derivations	83

Preface

This document contains the course notes for

751-7602-00L Applied Statistical Methods in Animal Sciences.

General Developments

With the advent of **Big Data** (see [Wikipedia, 2019] and [Mashey, 1998] for a reference), it became clear that the importance of statistical methods to analyze the huge amounts of collected data would increase dramatically. Many modern statistical methods are only applicable due to the vast availability of cheap computing resources. The progress of the development that happens in the hardware manufacturing industry is often referred to by the term **Moore's Law**. This law was stated as a projection as early as 1965 by one of the founders of the Intel cooperation [Moore, 1965]. In a very general term, Moore's law says that the number of circuits that could be placed on a silicon waver would double every 18 months. In a derived version the law was interpreted in a way that the performance of computers would double every 18 months. Together with the high degree of automated production of the building blocks of a computer, the prices for a single unit of computation dropped dramatically. This development made it possible that the possibility to analyze large amounts of data with modern methods can be done by almost everyone. This created very many opportunities which are actively used by many business companies. Statistical methods used to be only used by academic researchers. Nowadays almost all important decisions in business companies are done based on supporting facts that are derived from analyzing market and customer data. With that it is clear that the importance of being able to use statistical methods to analyze data is almost ubiquitous and the knowledge of these methods can be very important in many different jobs or employments.

Where Does This Course Fit In?

This course gives a short introduction to a collection of statistical methods that I believe are relevant for a wide range of topics in Animal Sciences. These methods include

- Multiple Linear Least Squares Regression (MLLSR)
- Best Linear Unbiased Prediction (BLUP) which is called GBLUP when applied in the context of genomics
- Least Absolute Shrinkage and Selection Operator (LASSO)
- Bayesian Estimation of Unknown Parameters (BEUP)

The above listed collection of statistical methods all happen to be illustrated around the same type of dataset. This dataset contains the genetic variants at many locations in the genome for a number of livestock breeding animals. Because there are many genetic locations considered in such a dataset and the locations are distributed across the complete genome, such a dataset is referred to as a **genomic** dataset. This type of dataset does appear in an area of livestock breeding which is called **Genomic Selection** (GS). GS was introduced in a seminal paper by [Meuwissen et al., 2001]. This very same paper is used as a building block to explain some of the statistical methods (MLLSR and BEUP) used in this course. Furthermore the same publication illustrates that some methods (MLLSR) are not suitable for analyzing certain aspects in a genomic dataset.

The time available for this course is just half a semester. This leaves very little time for the introduction of each topic. As a consequence of that each topic can only be presented very superficially and students are expected to work on their own during the exercise hours. Exercises consist of sets of problems related to each topic. Problems are often to be expected to be solved using the R programming language [R Core Team, 2018].

This version of the course corresponds to the fifth edition. With each additional iteration of the course, improvements are sought to be implemented. Hence any input from the students are greatly appreciated.

Course Objectives

The students are familiar with the properties of multiple linear regression and they are able to analyze simple data sets using regression methods. The students know why multiple linear regression cannot be used for problems where the number of parameters exceeds the number of observations. One such problem is the prediction of genomic breeding values used in genomic selection. The students know alternative statistical methods that can be applied in situations where the number of parameters is larger than the number of observations. Examples of such methods are BLUP-based approaches, Bayesian procedures

and LASSO. The students are able to solve simple exercise problems applying BLUP-based approaches, LASSO and BEUP. The students are expected to use the statistical language and environment R [R Core Team, 2018].

Prerequisites

Because the data that is used in this course comes from genetics, a basic level of quantitative genetics is useful for this course. All statistical models will be presented in matrix-vector notation, hence some basics of linear algebra helps in understanding the presented material. Introductory chapters to both subjects (quantitative genetics and linear algebra) are included in these course notes, but will not be discussed during the lecture. These chapters are prepared for students who feel that they need more background. But this material is left for self-studying.

Chapter 1

Introduction

According to Wikipedia [Wikipedia, 2019], the term **Big Data** has been used since the 1990s. Some credit was given to John Mashey [Mashey, 1998] for popularizing the term. Nowadays **Big Data** is used in connection with large companies, social media or governments which collect massive amounts of data. This data is then used to infer certain conclusions about behaviors of customers, or followers or voters. The following subsections show a few examples of **Big Data**-applications.

1.1 US Presidential Campaigns

The presidential election campaigns of Barack Obama were examples of how **Big Data** was used to access behaviors of voters [Issenberg, 2013].

1.2 Health Care

A different example is the use of **Big Data** in health care. An overview of the use of **Big Data** in health care is given in [Adibuzzaman et al., 2017]. The collected health data is most likely not only used by research but also by insurance companies.

1.3 Face Recognition

The Swiss TV news show 10 vor 10 showed on the 7th Feb. 2020 how a data journalist managed to build a face recognition system. The general idea how this system works is shown in Figure 1.1.

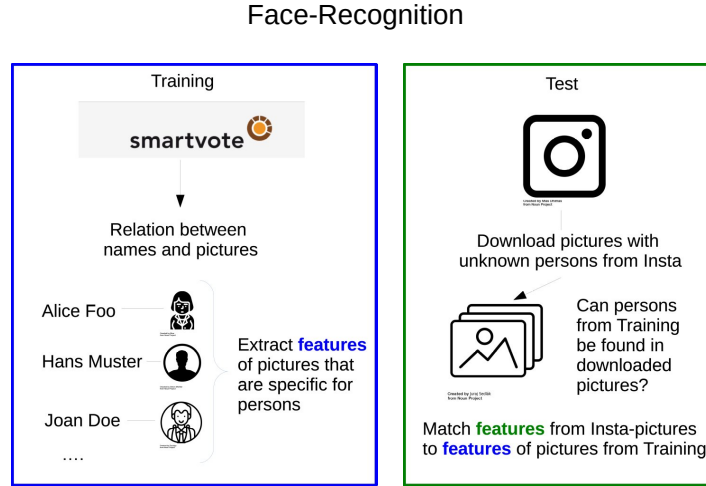


Figure 1.1: Design of Face Recognition System

The main goal of the face recognition system was to be able to identify certain persons, in this case the politicians that had a picture on the platform ‘smartvote’, on random pictures obtained from the social media platform Instagram. The data used for the face recognition system can be split into two parts.

1. **Training set.** The training set consists of pictures showing different politicians. This dataset was downloaded from the politics platform smartvote. The training set is used to establish a fixed relation between pictures of politicians and their names. Furthermore the pictures of the politicians are also used to extract characteristic features that are different among the different pictures of the different politicians. Most of these features consist of numbers which describe the faces shown on the pictures. Examples of such features might be the surface of the face, the surface of the hairs shown in the picture, the length of the mouth, etc.
2. **Test set:** The test set consists of 230000 publicly available pictures on the platform Instagram. The content of these pictures is a priori unknown. The question that the face recognition system tries to answer is whether it is possible to identify any of the politicians from the training set on any of the Instagram-pictures. This question is answered by a comparison of the features extracted from the instagram pictures to the features that were obtained from the pictures of the training set. If the feature comparison results in a match, the system suggests that we found a given person on

one of the instagram pictures.

The complete story about the face recognition system is available under <https://www.srf.ch/news/schweiz/automatische-gesichtserkennung-so-einfach-ist-es-eine-ueberwachungsmaschine-zu-bauen>.

1.4 Feed Intake and Behavior Traits of Cows

In the recent past technologies based on computer vision have been introduced into agricultural applications. Two examples of such applications are

1. Estimation of feed intake of cows based on video data as described by [Chizzotti et al., 2015].
2. An ongoing EU-Interreg project called “SESAM” aims at predicting basic behavior traits from data obtained from sensors and from video recordings.

1.5 Conclusions from Examples

The above shown examples demonstrate that data can be used for very different purposes. Using just one source of data does in most cases not give a lot of insights. But when different sources of information are combined, they can be used to make certain predictions that influence our daily lives. Hence this kind of development is becoming a general interest to all of us. In what follows, we try to show that some of these methods have been applied for a long time in the area of animal science and especially in livestock breeding.

1.6 Traditional Livestock Breeding

In livestock breeding the statistical analyses that are used together with **Big Data** technologies have long been applied to predict breeding values for livestock populations. The process of breeding value prediction uses statistical methods to assess the genetic potential of breeding animals in a population. The data used to predict the breeding values are collected mainly for quality control or management purposes. The prediction of breeding values can be viewed as a side product. In the area of cattle breeding, data collection consists of rather complex flows of information. The flow of information is shown in Figure 1.2.

1.7 Genomic Selection

The data flow shown in Figure 1.2 contains the traditional evaluation of data to result in predicted breeding values. But it is missing the newest development

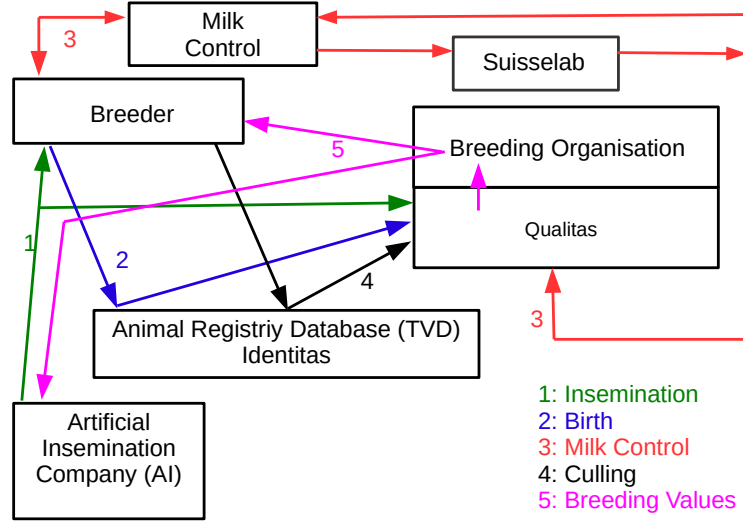


Figure 1.2: Data Flow in an Animal Breeding Program

in the breeding industry. This development is known as **Genomic Selection** (GS). GS was introduced by the work of [Meuwissen et al., 2001]. The methods presented by [Meuwissen et al., 2001] were only introduced into practical breeding programs when [Schaeffer, 2006] showed the tremendous potential of saving costs for breeding programs. The use of **genomic** information for the assessment of the genetic potential of all breeding animals represents the core of the evaluation approach presented by [Meuwissen et al., 2001]. The term **genomic** is used because genetic markers which are evenly spaced over the complete genome are used as information source. Single Nucleotide Polymorphisms (SNP) are the most widely used marker model nowadays. SNPs are single positions in the genome that occur in different variants in the whole population. A description on how to identify SNPs in a population is given in [Czech et al., 2018]. Potential use cases of SNPs are outlined by [Seidel, Jr., 2010] and [Pant et al., 2012]. The genetic configuration of an SNP in a given population is shown in Figure 1.3.

These SNPs can occur anywhere in the genome which means they can be observed in coding regions, in non-coding regions as well as in regulatory regions. In genomic selection, we are working with a large set of SNPs that are distributed over the complete genome. Hence some of the SNPs will be located close to genetic positions that are important for the expression of quantitative traits of interest. Such genetic positions which are related to quantitative traits are often called **Quantitative Trait Loci** (QTL). QTL themselves are difficult to detect and their inheritance is often manifested in complex modes. But due

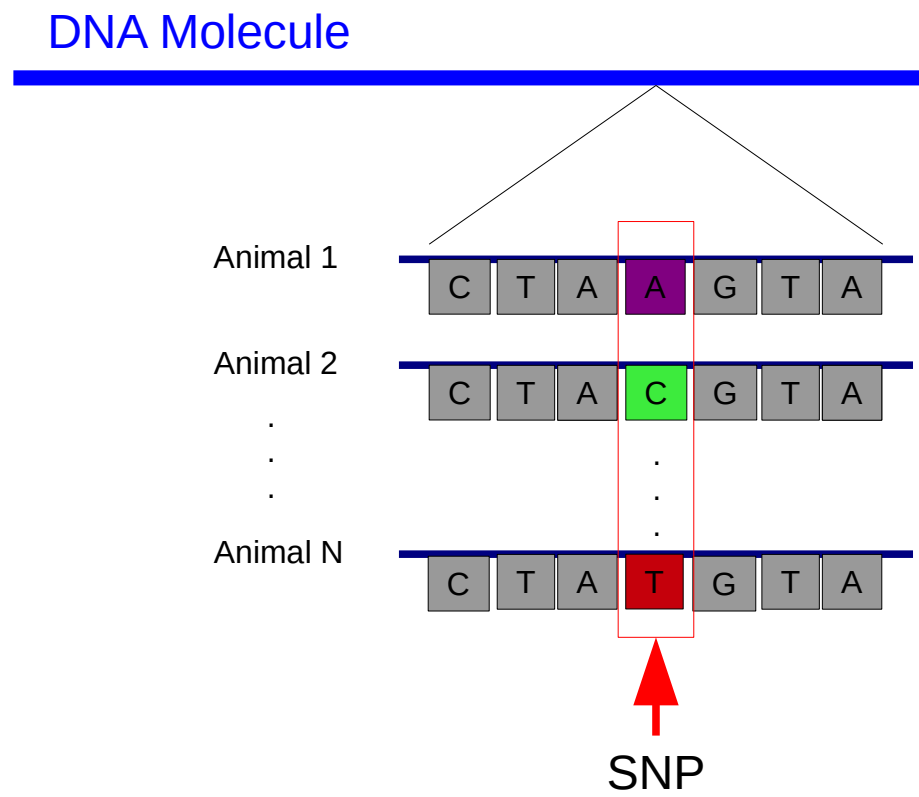


Figure 1.3: Genetic Configuration of a Single Nucleotide Polymorphism (SNP)

to the likely occurrence of several SNPs in the close proximity of a QTL, the inheritance of QTL alleles and of surrounding SNP alleles will not be independent due to linkage between SNPs and QTL. Such a linkage scenario between two SNPs flanking a QTL is shown in Figure 1.4.

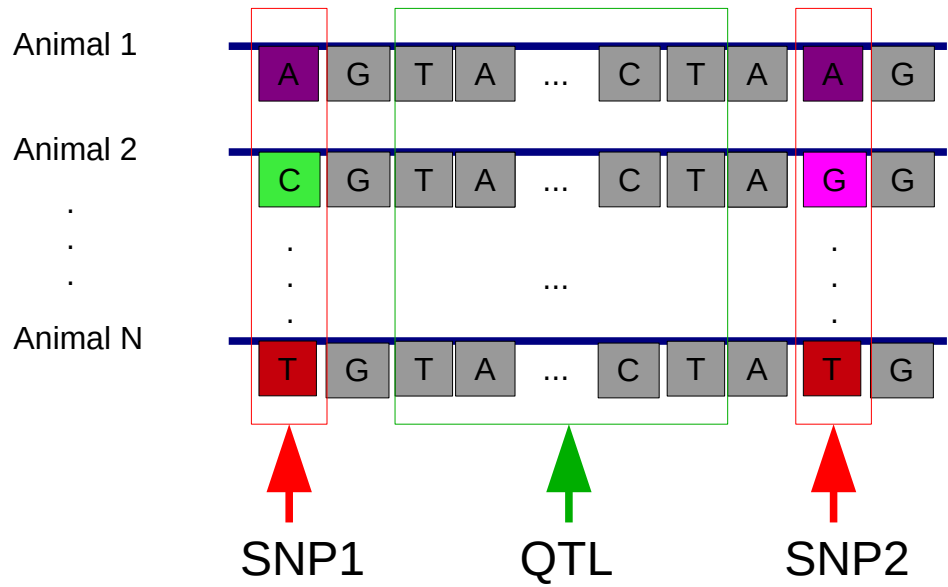


Figure 1.4: Two SNPs flanking a QTL

Although the QTL is likely to span a range of many positions on the chromosome, we can still assume the QTL to be bi-allelic with alleles Q_1 and Q_2 . In theory, any SNP position can have four different alleles according to the four different bases. But when looking at different SNPs in real-world populations, most of them only show two alleles. Hence, for the two SNPs flanking the QTL shown in Figure 1.4 they also have just two alleles $SNP1_1$, $SNP1_2$, $SNP2_1$ and $SNP2_2$. In genetics the dependency of the inheritance of neighboring loci (marker or QTL) is referred to as **linkage disequilibrium** (LD). This means that any joint allele frequency $Pr(SNP1_i, Q_j, SNP2_k)$ does not correspond to

the product of the single allele frequencies of the two SNPs (*SNP1* and *SNP2*) and the QTL. In a formula this can be written as

$$Pr(SNP1_i, Q_j, SNP2_k) \neq Pr(SNP1_i) * Pr(Q_j) * Pr(SNP2_k) \quad (1.1)$$

Assuming that the QTL allele Q_1 is favorable for the expression of a given trait of interest and using the fact of LD as expressed in (1.1), the alleles of *SNP1* and *SNP2* which occur more frequently together with Q_1 are therefore also related to favorable expression levels of the trait of interest. In real breeding populations, the position of the QTL is unknown. But because we know the allelic configuration of a large number of SNP loci from many breeding animals, we can reliably relate SNP alleles and favorable expression levels of traits of interest.

1.8 Mono-Genic Model

In quantitative genetics, the so-called mono-genic or single-locus model allows us to quantify the genetic potential of breeding animals in terms of breeding values. The standard reference in quantitative genetics in which also the mono-genic model is described is [Falconer and Mackay, 1996]. For a single locus, the breeding value depends on the allele frequencies at that locus and on the additive substitution effect which is often called α . The mono-genic model for any given SNP locus in relation to the level of expression of a given trait of interest can be visualized in the following Figure 1.5.

In a real breeding population, we assume that the effect of all loci linked to the SNPs are purely additive. Hence any values for d are all zero. As a consequence of that the breeding values at any given SNP position only depend on the allele frequencies of the SNP and the a values at every SNP. The overall breeding value of a given animal is computed as the sum of all locus-specific breeding values. This overall breeding value is called **genomic breeding value** (GBV). In order to get an estimate of such a GBV, we have to estimate all a values at any SNP position. This estimation procedure can be done in one of the following two ways.

1. Two step approach
2. Single step approach

1.9 Two Step Approach

In the two step approach the estimation of the a -values and the computation of the GBVs are done in two separate steps. For the estimation of the a values for all SNPs, a reference population is defined. In dairy cattle breeding this reference population consists of all male breeding animals. In the recent past,

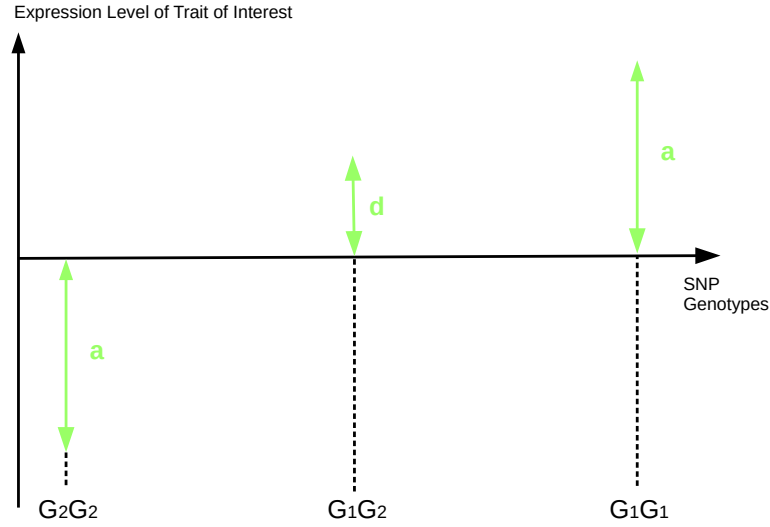


Figure 1.5: Single-Locus Model for a Quantitative Trait

the reference population has been augmented continuously with female animals. The animals in the reference population are all genotyped and they also all have phenotypic measurements¹ for the trait of interest. The estimation of the a values amounts to estimating fixed effects in a linear model. We will see in later chapters of this course what methods are available to estimate these parameters.

In the second step the estimates for all the a values are used to compute the GBVs for all animals with genomic SNP information also for those outside of the reference population. The Figure 1.6 tries to summarize the process graphically.

The big advantage of the two step method is that once we have defined a good reference population which yields reliable estimates for the a values, the computation of the GBV is a simple computation of just summing up the a contributions with the correct sign determined by the SNP genotypes of the animals for which the GBVs should be determined. All animals with SNP genotypes can get GBV values. The difficult part in the two step approach is to define a reliable reference population and to determine good phenotypic measurements (y).

¹Whenever phenotypic measurements are not available, traditionally predicted breeding values are transformed back into pseudo phenotypes which are then used to estimate a values.

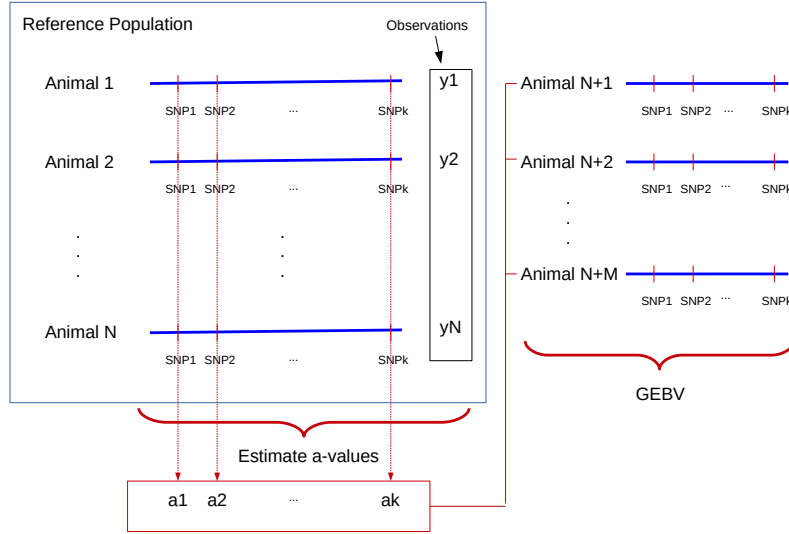


Figure 1.6: Two Step Approach To Estimate Genomic Breeding Values

1.10 Single Step Approach

The estimation of the a values and the prediction of the genomic breeding values is done in one step using linear mixed effects models. In this single step evaluation animals with and without genomic information can get predicted genomic breeding values in a single analysis. One possibility to get to this predicted breeding values is via the use of **Genomic BLUP** (GBLUP). This will be the topic of a complete chapter in this course. The problem with the single step approach is to get an estimate of the covariance between animals with and without genomic information. This is a problem of ongoing research.

1.11 Summary

The main difference between traditional predictions of breeding values using a BLUP animal model and the prediction of GBV is that the former uses the so called **infinitesimal** model to assess the genetic potential and the latter uses sufficiently dense genomic information and uses a **polygenic** model. This difference is illustrated in Figure 1.7.

In the remaining chapters, different approaches for the prediction of GBVs are described. Chapter 2 gives a description of the fixed linear effects model and how it was tried to be used for GBV prediction by [Meuwissen et al., 2001].

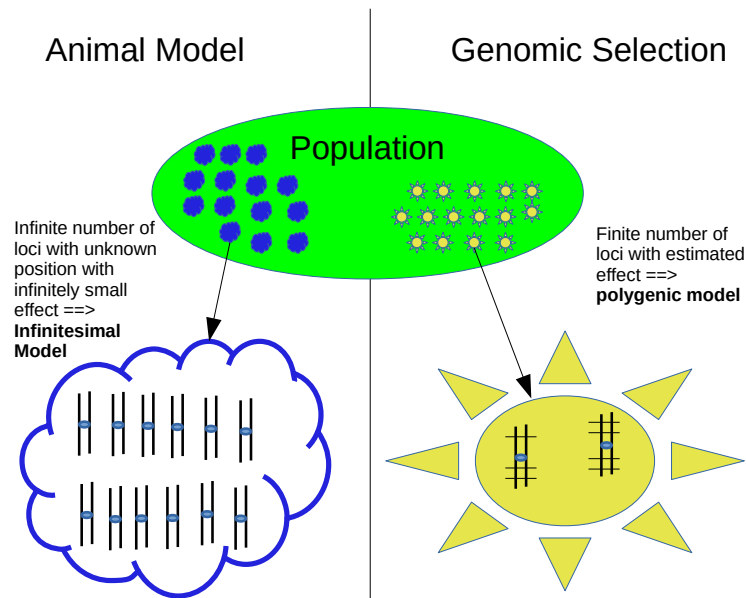


Figure 1.7: Infinitesimal Versus Polygenic Model

Chapter 3 introduces BLUP methodology in the context of predicting GBVs. In Chapter 4 the method called LASSO is introduced. Interestingly enough, this method is used very seldom in the area of animal breeding. Last but not least, Chapter 5 makes an excursion into Bayesian estimation approaches. The Bayesian methods are important because they are used in practical breeding programs of Swiss Dairy cattle.

Chapter 2

Fixed Linear Effects Models

2.1 Resources

This chapter on **fixed linear effects models** (FLEM) is based on the work of [Buehlmann and Maechler, 2014]. Besides that, there are many other resources for the topic of **Multiple Linear Regressions**. An interesting online book is [Lilja, 2016].

2.2 Motivation

Why is the topic of fixed linear effects models important for the analysis of genomic data? This question is best answered when looking at the data. In chapter 1, we saw that genomic breeding values can either be estimated using a two-step procedure (see section 1.9) or by a single step approach (see section 1.10). At the moment, we assume that we are in the first step of the two step approach where we estimate the marker effects (a -values) in a reference population or alternatively we have a perfect data set with all animals genotyped and with a phenotypic observation in a single step setting. Both situations are equivalent when it comes to the structure of the underlying dataset. Furthermore the same class of models can be used to analyse the data.

2.3 Data

As already mentioned in section 2.2, we are assuming that we have a perfect dataset for a given population of animals. That means each animal i has a phenotypic observation y_i for a given trait of interest. Furthermore, we assume to have a map of only three SNP markers. The marker loci are called G , H and

I. All markers have two alleles each. Figure 2.1 tries to illustrate the structure of a dataset used to estimate genomic breeding values (GBV).

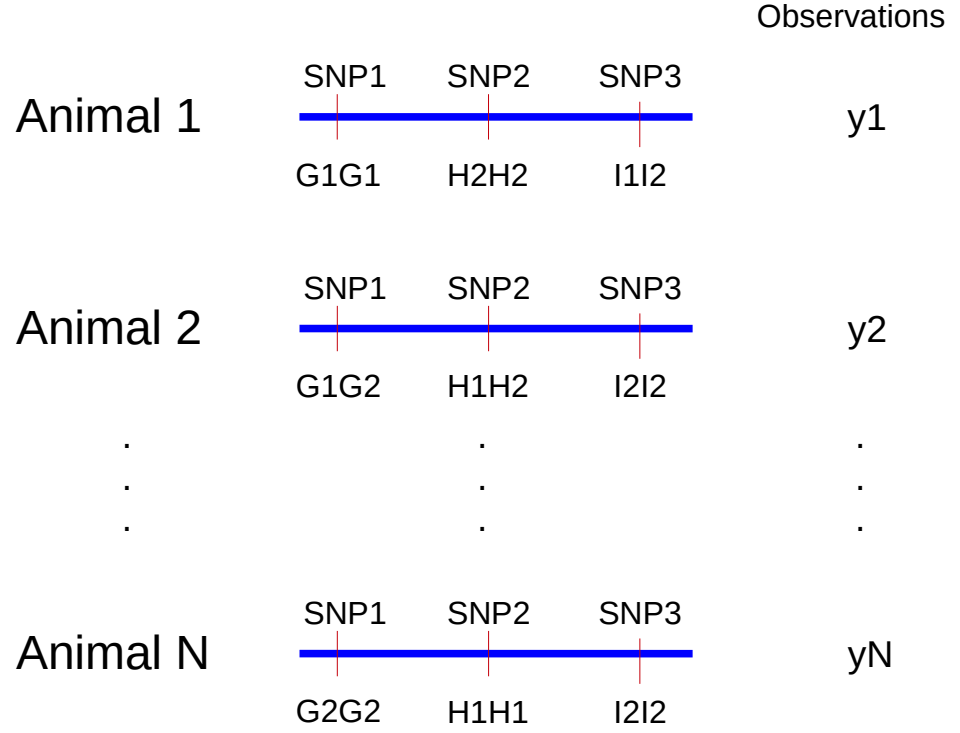


Figure 2.1: Structure of Dataset To Estimate GBV

As can be seen from Figure 2.1 each of the N animals have known genotypes for all three SNP markers and they all have a phenotypic observation y_i ($i = 1, \dots, N$). Because we are assuming each SNP marker to be bi-allelic, there are only three possible marker genotypes at every marker position. Hence marker genotypes are discrete entities with a fixed number of levels. Due to the nature of the SNP marker genotype data, we can already say that they could be modeled as fixed effects in a fixed linear effects model. More details about the model will follow in section 2.4.

2.4 Model

The goal of our data analysis using the dataset described in section 2.3 is to come up with estimates for genomic breeding values for all animals in our dataset. The genomic breeding values will later be used to rank the animals. The ranking of the animals according to the GBV is used to select the parents of the future generation of livestock animals. It probably makes sense to distinguish between two different types of models that we have to set up. On the one side we need a model that describes the underlying genetic architecture which is present in our dataset. We will be using a so-called **genetic** model to describe this. On the other side, we have to be able to get estimates for the GBVs which requires a **statistical** model which is able to estimate unknown parameters as a function of observed data. In the end, we will realize that the two models are actually the same model but they are just different ways of looking at the same structure of the underlying phenomena. These phenomena characterize the relationship between genetic architecture of an animal and the expression of a certain phenotypic trait in that same animal.

2.4.1 Genetic Model

The availability of genomic information for all animals in the dataset makes it possible to use a polygenic model. In contrast to an infinitesimal model, a polygenic model uses a finite number of discrete loci to model the genetic part of an expressed phenotypic observation. From quantitative genetics (see e.g. [Falconer and Mackay, 1996] for a reference) we know that every phenotypic observation y can be separated into a genetic part g and an environmental part e . This leads to the very simple genetic model

$$y = g + e \quad (2.1)$$

The environmental part can be split into some fixed known systematic factors such as **herd**, **season effects**, **age** and more and into a random unknown part. The systematic factors are typically grouped into a vector of fixed effects called β . The unknown environmental random part is usually called ϵ . This allows to re-write the simple genetic model in (2.1) as

$$y = \beta + g + \epsilon \quad (2.2)$$

The genetic component g can be decomposed into contributions from the finite number of loci that are influencing the observation y . In our example dataset (see Figure 2.1) there are three loci¹ that are assumed to have an effect on y .

¹Implicitly, we are treating the SNP-markers to be identical with the underlying QTL. But based on the fact that we have very many SNPs spread over the complete genome, there will always be SNP sufficiently close to every QTL that influences a certain trait. But in reality

Ignoring any interaction effects between the three loci and thereby assuming a completely additive model, the overall genetic effect g can be decomposed into the sum of the genotypic values of each locus. Hence

$$g = \sum_{j=1}^k g_j \quad (2.3)$$

where for our example k is equal to three².

Considering all SNP loci to be purely additive which means that we are ignoring any dominance effects, the genotypic values g_j at any locus j can just take one of the three values $-a_j$, 0 or $+a_j$ where a_j corresponds to the a value from the mono-genic model (see Figure 1.5). For our example dataset the genotypic value for each SNP genotype is given in the following table.

Table 2.1: Genotypic Values For All Three SNP-Loci

SNP Locus	Genotype	Genotypic Value
SNP_1	G_1G_1	a_1
SNP_1	G_1G_2	0
SNP_1	G_2G_2	$-a_1$
SNP_2	H_1H_1	a_2
SNP_2	H_1H_2	0
SNP_2	H_2H_2	$-a_2$
SNP_3	I_1I_1	a_3
SNP_3	I_1I_2	0
SNP_3	I_2I_2	$-a_3$

From the Table 2.1 we can see that always the allele with subscript 1 is taken to be that with the positive effect. Combining the information from Table 2.1 together with the decomposition of the genotypic value g in (2.3), we get

$$g = m^T \cdot a \quad (2.4)$$

where m is an indicator vector taking values of -1 , 0 and 1 depending on the SNP marker genotype and a is the vector of a values for all SNP marker loci. Combining the decomposition in (2.4) together with the basic genetic model in (2.2), we get

$$y = \beta + m^T \cdot a + \epsilon \quad (2.5)$$

the unknown QTL affect the traits and not the SNPs.

²In reality k can be $1.5 * 10^5$ for some commercial SNP chip platforms. When working with complete genomic sequences, k can also be in the order of $3 * 10^7$.

The result obtained in (2.2) is the fundamental decomposition of the phenotypic observation y into a genetic part represented by the SNP marker information (m) and an environmental part (β and ϵ). The a values are unknown and must be estimated. The estimates of the a values will then be used to predict the GBVs. How this estimation procedure works is described in the next section 2.4.2.

2.4.2 Statistical Model

When looking at the fundamental decomposition given in the genetic model presented in (2.5) from a statistics point of view, the model in (2.5) can be interpreted as **fixed linear effects model** (FLEM). FLEM represent a class of linear models where each model term except for the random residual term is a fixed effect. Furthermore, besides a random error term, the response is explained by a linear function of the predictor variables.

Using the decomposition given in our genetic model (see equation (2.5)) for our example dataset illustrated in Figure 2.1, every observation y_i of animal i can be written as

$$y_i = W_i \cdot \beta + M_i \cdot a + \epsilon_i \quad (2.6)$$

where

- y_i is the observation of animal i
- β is a vector of unknown systematic environmental effects
- W_i is an indicator row vector linking β to y_i
- a is a vector of unknown additive allele substitution effects (a values)
- M_i is an indicator row vector encoding the SNP genotypes of animal i and
- ϵ_i is the random unknown environmental term belonging to animal i

In the following section, we write down the definition of a FLEM and compare it to the statistical model given in (2.6).

2.5 Definition of FLEM

The multiple fixed linear effects model is defined as follows.

Definition 2.1 (Fixed Linear Effects Model). In a fixed linear effects model, every observation i in a dataset is characterized by a **response variable** and a set of **predictors**. Up to some random errors the response variable can be expressed as a linear function of the predictors. The proposed linear function contains unknown parameters. The goal is to estimate both the unknown parameters and the error variance.

2.5.1 Terminology

For datasets where both the predictors and the response variables are on a continuous scale, which means that they correspond to measured quantities such as body weight, breast circumference or milk yield, the model is referred to as **multiple linear regression model**. Because the statistical model in (2.6) contains the SNP genotypes as discrete fixed effects, we are not dealing with a regression model but with a more general fixed linear effects model.

2.5.2 Model Specification

An analysis of the model given in (2.6) shows that it exactly corresponds to the definition 2.1. In this equivalence, the observation y_i corresponds to the response variable. Furthermore, the unknown environmental term ϵ corresponds to the random residual part in the FLEM. Except for the random residuals the response variable y_i is a linear function of the fixed effects which corresponds to all systematic environmental effects and to all SNP genotype effects.

For the description of how to estimate the unknown parameter β and a in the model (2.6), it is useful to combine β and a into a single vector of unknown parameters and we call it b .

$$b = \begin{bmatrix} \beta \\ a \end{bmatrix} \quad (2.7)$$

Taking the equations as shown in (2.6) for all observations ($i = 1, \dots, N$) and expressing them in matrix-vector notation, we get

$$y = Xb + \epsilon \quad (2.8)$$

where

- y is the vector of N observations
- b is the vector of all unknown fixed effects
- X is the incidence matrix linking the parameters of b to y
- ϵ is the vector of random residuals

The incidence matrix X in (2.8) can be composed from the matrices W and M by concatenating the latter two matrices, i.e.,

$$X = [\ W \quad M \] \quad (2.9)$$

2.6 Parameter Estimation Using Least Squares

The method of parameter estimation is explained using the simpler case of a regression model. That means both the predictors and the response variables are on a continuous scale. As a further simplification, we assume that there is only one predictor variable and one response variable. The predictor variable is called x and the response variable is called y . The model is still the same as shown in (2.8). The matrix X has just one column with the measured values of the predictor variable and b is just a scalar unknown parameter. The vector y contains the observed values for the response values.

The goal of the analysis of the simple dataset is to find an estimate of the scalar b such that the linear combination of X and b best explains the values in y . How we can find such an estimation procedure that allows us to calculate an estimate of b is explained using a small example dataset in the following subsection.

2.6.1 An Example Dataset

A widely use example dataset for such a simple regression analysis in animal breeding consists of measurements of **body weight** (BW) and **breast circumference** (BC) for a given group of animals.

Table 2.2: Dataset for Regression of Body Weight on Breast Circumference for ten Animals

Animal	Breast Circumference	Body Weight
1	176	471
2	177	463
3	178	481
4	179	470
5	179	496
6	180	491
7	181	518
8	182	511
9	183	510
10	184	541

The dataset shown above is taken from Table 9.1 in [Essl, 1987]. One of the possible reasons for fitting a regression from BW on BC is that the latter is easier to measure. The measured values of BC can be used to predict BW once we have determined the regression coefficient. For this prediction, we use BW as response variable y and BC as predictor variable x . This leads to the regression model

$$y = x * b + \epsilon \quad (2.10)$$

where y is the vector of body weights and x is the vector of breast circumferences. b is a scalar value which is unknown and ϵ is the vector of random unknown error terms. The goal is to determine b such that the predictor variable best explains the response variable. How b is determined is explained with the following plot.

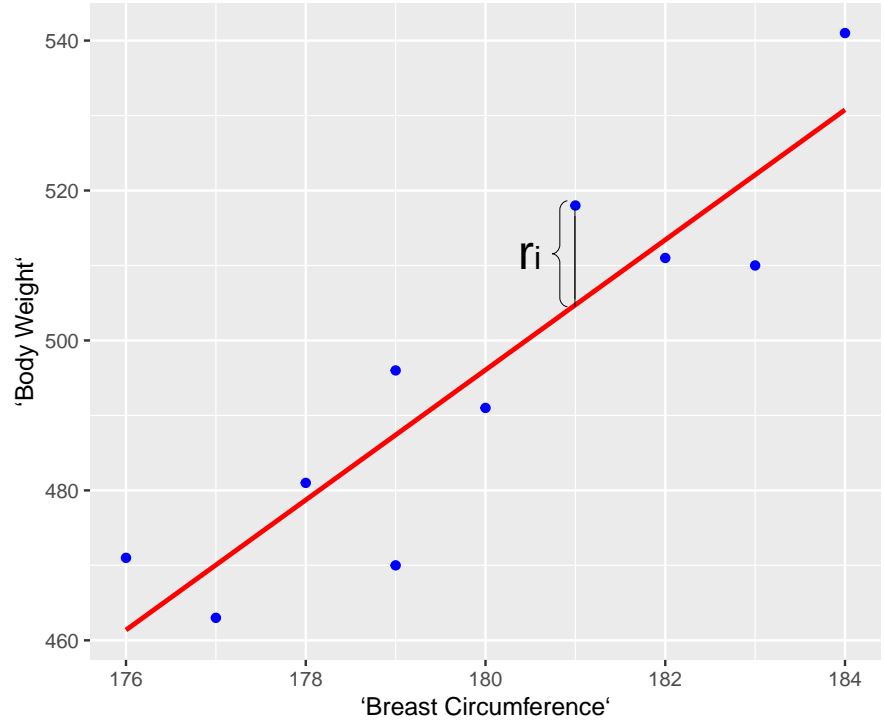


Figure 2.2: Regression of Body Weight On Breast Circumference

In Figure 2.2 the blue points correspond to the data points given by the dataset shown in Table 2.2. The red line corresponds to the regression line defined by the unknown regression parameter b . The distance between the data points to the projection in the direction of the y -axis corresponds to the residual r . For a given data point i , the residual r_i is computed as

$$r_i = y_i - x_i * \hat{b} \quad (2.11)$$

where \hat{b} denotes a concrete estimated value of b . For a different choice of a value of \hat{b} , different values for the residuals r_i can be computed. Our goal is to find the value of \hat{b} that results in the smallest residuals r_i . In order to avoid cancellation of positive and negative values of the residuals, the r_i values are squared and added. This sum of the squared residuals is used as a measure of how good a given regression line determined by \hat{b} fits a given set of data points. Because we want to have a good fit this means that the sum of the squared residuals should be as small as possible.

The method that determines \hat{b} such that the sum of the squared residuals is minimal is called **Least Squares**. In a general formula with more than one predictor variables we can write the least squares estimate \hat{b}_{LS} as

$$\hat{b}_{LS} = \operatorname{argmin}_b \|y - Xb\|^2 \quad (2.12)$$

where $\|\cdot\|$ denotes the Euclidean norm. The estimate \hat{b}_{LS} can be found by finding the minimum of $\|y - Xb\|^2$. The minimum of $\|y - Xb\|^2$ is found by first taking the derivative with respect to b and then setting that derivative to 0. The derivative of $\|y - Xb\|^2$ with respect to b can be computed by starting with the derivation of the least squares criterium LS as follows

$$LS = \|y - Xb\|^2 = (y - Xb)^T (y - Xb) = y^T y - y^T Xb - b^T X^T y + b^T X^T Xb \quad (2.13)$$

The derivative of LS with respect to b corresponds to the gradient vector of LS with respect to b which contains the derivative of LS with respect to every component b_i of the vector b . This gradient vector is

$$\frac{\partial LS}{\partial b} = -X^T y - X^T y + 2 * X^T Xb \quad (2.14)$$

The minimum is found by setting $\frac{\partial LS}{\partial b}$ to 0 which is satisfied by a specific vector called \hat{b} .

$$\frac{\partial LS}{\partial b} = -X^T y - X^T y + 2 * X^T X\hat{b} = 0 \quad (2.15)$$

From equation (2.15), we get the so-called least squares **Normal Equations** for \hat{b} .

$$X^T X\hat{b} = X^T y \quad (2.16)$$

For a regression model, we know that X has full column rank³. That means we can solve the normal equations (2.16) explicitly for \hat{b} .

$$\hat{b} = (X^T X)^{-1} X^T y \quad (2.17)$$

Equation (2.17) presents a solution to the estimation problem of the unknown parameter b in the regression problem. There is one additional unknown parameter that we have not mentioned so far. The regression model contains the random error terms ϵ . Because ϵ is random, we have to specify the expected value and the variance. The error terms are deviations of the predicted values from the observed data points. Hence the expected values $E[\epsilon]$ must be 0. The variance σ^2 of the error terms is an additional unknown parameter that has to be estimated from the data. One way of estimating the error variance from the data is shown in subsection 2.6.2.

2.6.2 Variance of Errors

The least squares procedure itself does not yield an estimate of the error variance σ^2 . But the estimate of σ^2 based on the residuals is often declared to be the **least squares estimate** of σ^2 . The residuals r_i as defined in (2.11) are estimates of the error terms ϵ_i . As a matter of fact the residuals can be used to estimate σ^2 . This estimate is given by

$$\widehat{\sigma^2} = \frac{1}{n-p} \sum_{i=1}^n r_i^2 \quad (2.18)$$

The factor $(n-p)^{-1}$ in (2.18) is used, because it leads the estimate $\widehat{\sigma^2}$ to be unbiased, which means $E[\widehat{\sigma^2}] = \sigma^2$.

2.7 Different Types of Linear Regressions

2.7.1 Regression Through The Origin

The regression model as it was proposed in (2.10) for the dataset of body weight and breast circumference defines a line in the $x-y$ -plane. This line shown in Figure 2.2. What is not shown in the plot, but what becomes clear from the model is that the regression line goes through the origin of the coordinate system. Mathematically the origin is given by $x = 0$ and $y = 0$. In this regression model, the origin is the fixed point which is on the regression line. The fixed point together with the estimated regression coefficient b uniquely define

³In a regression model, all values in the matrix X are real values. Hence no column of X will be a linear combination of any other columns and therefore X has full column rank.

the regression line. From a geometrical point of view the estimated regression coefficient defines the slope of the regression line.

2.7.2 Regression With Intercept

Depending on the data analysed with a regression model, it does not make sense to force the regression line to run through the origin. This can be avoided by including an additional fixed term in the regression model. This term is called the **intercept**. A regression model with an intercept can be written as

$$y_i = b_0 + x_i * b_1 + \epsilon_i \quad (2.19)$$

The term b_0 corresponds to the value of the response variable y when the value of the predictor x is 0. Then the fixed point of the regression line is no longer the origin, but the point $x = 0$ and $y = \widehat{b_0}$. The slope of the regression line is determined by $\widehat{b_1}$. In matrix-vector notation the intercept b_0 is added to the vector of unknown parameters b and the design-matrix X has to be augmented by a column of all ones on the left.

2.7.3 Regression With Transformed Predictor Variables

Regression models can also contain different transformations of the predictor variables. As an example, we can include any higher order polynomial functions of predictor variables such as

$$y_i = b_0 + b_1 * x_i + b_2 * x_i^2 + \dots + b_k * x_i^k + \epsilon_i \quad (2.20)$$

Although the model (2.20) contains non-linear functions of the predictors x_i , the function is still linear in the unknown parameters b_j ($j = 1, \dots, k$) and hence the model (2.20) is still a linear regression model.

Transformations of the predictor variables are not restricted to polynomial functions. Many different kinds of transformations are possible. An example is shown in the following equation

$$y_i = b_0 + b_1 * \log(x_i) + b_2 * \sin(\pi x_i) + \epsilon_i \quad (2.21)$$

2.8 Predictions

One goal of estimating the regression coefficient was that we want to be able to predict the response based on concrete values of the predictor variables. For our example with the body weight and the breast circumference, this means that

we want to measure the breast circumference of an animal for which we do not know the body weight. Then based on the estimated regression coefficient, we want to be able to predict the body weight of that animal.

The computation of the regression coefficient for the dataset shown in Table 2.2 will be the topic of an exercise. But let us assume that we have computed the value of \hat{b} , then the predicted value of the body weight \hat{y}_s for an animal s is computed based on the measured breast circumference x_s of animal s as follows

$$\hat{y}_s = \hat{b} * x_s \quad (2.22)$$

It has to be noted that the prediction \hat{y}_s is only valid, if the measured value x_s is close to the measured predictors that were used to estimate \hat{b} . For our example with body weight and breast circumference, we could not use the same regression line to predict the body weight for calves, if \hat{b} was estimated with data of adult bulls.

2.9 Regression On Dummy Variables

In a regression model (such as shown in (2.10)) both the response variable and the predictor variables are continuous variables. Examples of such variables are **body weight** and **breast circumference** which are both measured and the measurements are expressed as real numbers. In contrast to such a regression model, the statistical model shown in (2.6) has a continuous response, but the predictor variables are discrete variables. The predictor variables are assumed to be genotypes of a certain set of SNP genotypes and hence these genotypes can only have a fixed number of states. Under the assumption of bi-allelic Loci, a SNP locus can have just three genotypes and hence the predictor variable that is used to represent any given SNP-locus can only take three discrete states.

Figure 2.3 shows the difference between a regression model as the one of **body weight** on **breast circumference** and a fixed linear effects model where one locus has an effect on a quantitative trait. In the left diagram of Figure 2.3 the red line denotes the regression line. This line is meaningful because on the x-axis and on the y-axis every single point of the red line would be valid observations. On the x-axis of the diagram on the righthand side, only three values are possible. In the diagram they are shown as Genotypes G_1G_1 , G_1G_2 and G_2G_2 . We will see very soon that in our statistical model, they will be encoded by 1, 0 and -1 . The response variable in the diagram on the right of Figure 2.3 is a continuous random variable, similarly to the regression model shown in the left diagram. This combination of continuous response variable on a discrete type of variable lead to the term **regression on dummy variables** because the predictor variables are not continuous but just discrete levels of a certain factor. In this lecture, we are using **fixed linear effects model** rather than regression on dummy variables for the same type of model. The

term of fixed linear effects model was used, because in the next chapter in Genomic BLUP we are going to introduce mixed linear effects model which are an extension of the fixed linear effects model used in this chapter.

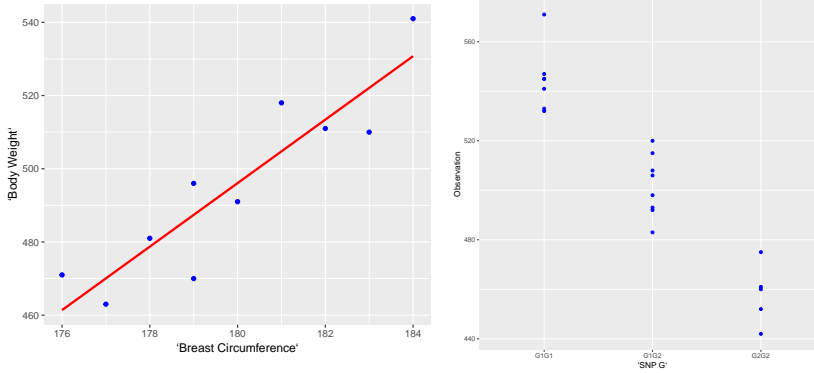


Figure 2.3: Comparison Between Regression Model And Fixed Linear Effects Model With An SNP-Locus As A Discrete Predictor Variables

2.9.1 Fixed Linear Effects Model For SNP Data

We are using genetic data and assume that the SNP genotypes have an effect on a quantitative trait. Our goal is to predict genomic breeding values based on the information from the SNP genotypes for the quantitative traits. We have seen that under some simplifying assumptions of additivity of the genetic effects, the genomic breeding values depend on the absolute value of the genotypic values (a values) of the homozygous SNP genotypes. Hence all we need to know from our analysis of the data under a fixed linear effects model are the a values for each SNP locus. The decomposition of the phenotypic observation shown in 2.4.1 under the assumed genetic model tells us that the phenotypic observation can be explained as a linear function of the genotypic values of the SNP genotypes plus a random error term. The fact that our genetic model is a fixed linear effects model that uses phenotypic observations as response and SNP loci as predictors allows us to set up the following model for an example data set shown in the following subsection.

2.9.2 Example Data Set With SNP Loci And A Phenotypic Observation

We are using the dataset shown in Table 2.3 as an example on how to use a fixed linear effects model to estimate the genotypic value of the SNP genotypes.

Table 2.3: Animals With Two SNP Loci Affecting A Quantitative Trait

Animal	SNP G	Genotypic Value G	SNP H	Genotypic Value H	Observation
1	G_1G_1	a_G	H_1H_2	0	510
2	G_1G_2	0	H_1H_1	a_H	528
3	G_1G_2	0	H_1H_1	a_H	505
4	G_1G_1	a_G	H_2H_2	$-a_H$	539
5	G_1G_1	a_G	H_1H_1	a_H	530
6	G_1G_2	0	H_1H_2	0	489
7	G_1G_2	0	H_2H_2	$-a_H$	486
8	G_2G_2	$-a_G$	H_1H_1	a_H	485
9	G_1G_2	0	H_2H_2	$-a_H$	478
10	G_2G_2	$-a_G$	H_1H_2	0	479
11	G_1G_1	a_G	H_1H_2	0	520
12	G_1G_1	a_G	H_1H_1	a_H	521
13	G_2G_2	$-a_G$	H_1H_2	0	473
14	G_2G_2	$-a_G$	H_1H_2	0	457
15	G_1G_2	0	H_1H_1	a_H	497
16	G_1G_2	0	H_1H_2	0	516
17	G_1G_1	a_G	H_1H_2	0	524
18	G_1G_1	a_G	H_1H_2	0	502
19	G_1G_1	a_G	H_2H_2	$-a_H$	508
20	G_1G_2	0	H_1H_2	0	506

Instead of fitting individual effects for the different SNP genotypes to explain the response variable, we are directly including the genotypic values a_G and a_H into the fixed effects linear model. How the genotypic values are related to the SNP genotypes is also shown in Table 2.3. For all animals in Table 2.3, we can write the model equations in matrix-vector notation as

$$y = Xb + \epsilon \quad (2.23)$$

where y is the vector of observations, b is a vector of genotypic values plus an intercept, X is a design matrix linking the elements in b to y and ϵ is a vector of random errors. Writing out the matrices and vectors leads to

$$\begin{bmatrix} 510 \\ 528 \\ 505 \\ 539 \\ 530 \\ 489 \\ 486 \\ 485 \\ 478 \\ 479 \\ 520 \\ 521 \\ 473 \\ 457 \\ 497 \\ 516 \\ 524 \\ 502 \\ 508 \\ 506 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & -1 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & -1 \\ 1 & -1 & 1 \\ 1 & 0 & -1 \\ 1 & -1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & -1 & 0 \\ 1 & -1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & -1 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} b_0 \\ a_G \\ a_H \end{bmatrix} + \epsilon \quad (2.24)$$

2.9.3 Parameter Estimation In A Fixed Linear Effects Model

The goal for model (2.23) is to get an estimate for the unknown parameters b_0 , a_G and a_H . In section 2.6 we saw how unknown parameters can be estimated for a regression model using least squares. When applying the least squares method, we did not make any assumptions about the predictor variables. The minimization of the sum of the squared residuals can also be applied for the fixed linear effects model. This minimization leads to the same normal equations

$$X^T X b^{(0)} = X^T y \quad (2.25)$$

So far everything was identical to the case of the regression model. But when trying to find a solution for (2.25) we have to account for the different nature of the design matrix X . In the regression model this matrix X contains real numbers. In our example of a fixed linear effects model, the matrix X just contains just the three number -1 , 0 and 1^4 . The fact that the matrix X contains only a few discrete values makes it very likely that X does not have full column rank. That means it is very likely that some columns of X can be expressed as linear combinations of other columns. This linear dependence of the columns of X causes the matrix $X^T X$ to be singular and hence the inverse of $X^T X$ cannot be computed. Whenever the matrix $X^T X$ is singular, the solution given in (2.17) cannot be computed.

The normal equations in (2.25) are written with the symbol $b^{(0)}$ to denote that the equations do not have a single solution $b^{(0)}$ in the sense that we were able to compute them in the case of the regression model. In the case where $X^T X$ is singular, there are infinitely many solutions $b^{(0)}$. These solutions can be expressed as

$$b^{(0)} = (X^T X)^- X^T y \quad (2.26)$$

where $(X^T X)^-$ stands for a **generalized inverse** of the matrix $X^T X$.

2.9.4 Generalized Inverse Matrices

A generalized inverse matrix G of a given matrix A is defined as the matrix that satisfies the equation $AGA = A$. The matrix G is not unique. Applying the concept of a generalized inverse to a system of equations $Ax = y$, it can be shown that $x = Gy$ is a solution, if G is a generalized inverse of A . Because G is not unique, there are infinitely many solutions corresponding to $\tilde{x} = Gy + (GA - I)z$ where z can be an arbitrary vector of consistent length. Applying these statements concerning generalized inverses and solutions to systems of equations to (2.26), it means that $b^{(0)}$ is not a unique solution to (2.25) because the generalized inverse $(X^T X)^-$ is not unique. As a consequence of that non-uniqueness, the solution $b^{(0)}$ cannot be used as an estimate of the unknown parameter vector b .

2.9.5 Estimable Functions

The numeric solution of the analysis of the example dataset given in Table 2.3 is the topic of an exercise. When developing that solution, we will see that some linear functions of $b^{(0)}$ can be found which do not depend on the choice of the generalized inverse $(X^T X)^-$. Such functions are called **estimable functions**

⁴In most other fixed linear effects models, the design matrix contains just 0 and 1.

Table 2.4: Example Showing Estimable Functions

Animal	SNP G	SNP H	SNP I	Observation
6	G_1G_1	H_1H_2	I_1I_2	16
6	G_1G_1	H_1H_2	I_1I_2	10
6	G_1G_1	H_1H_2	I_1I_2	19
6	G_1G_2	H_1H_1	I_1I_2	11
6	G_1G_2	H_1H_1	I_1I_2	13
6	G_1G_2	H_1H_2	I_1I_1	27

and can be used as estimates for the respective functions of the unknown parameter vector b . The idea of estimable functions can be demonstrated with the following example.

Let us assume that we have a small data set of 6 animals with observations in a particular traits and given genotypes at three loci. The dataset for that example is given in Table 2.4.

As shown before, we want to estimate the marker effects at the three loci G , H and I . This can be done with the following fixed effects model.

$$y = Xb + e$$

with

$$y = \begin{bmatrix} 16 \\ 10 \\ 19 \\ 11 \\ 13 \\ 27 \end{bmatrix}, X = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \text{ and } b = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix}$$

The vector b of unknown parameters consist of the general mean μ and the three marker effects α_1 , α_2 and α_3 . Based on the above information, the normal equations can be written as

$$\begin{bmatrix} 6 & 3 & 2 & 1 \\ 3 & 3 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu^0 \\ \alpha_1^0 \\ \alpha_2^0 \\ \alpha_3^0 \end{bmatrix} = \begin{bmatrix} 96 \\ 45 \\ 24 \\ 27 \end{bmatrix}$$

The above equations have infinitely many solutions. Four of them are shown below in Table 2.5.

Table 2.5: Solution of Normal Equations

Elements of Solution	b_1^0	b_2^0	b_3^0	b_4^0
μ^0	16	14	27	-2982
α_1^0	-1	1	-12	2997
α_2^0	-4	-2	-15	2994
α_3^0	11	13	0	3009

Table 2.6: Estimates of Estimable Functions

Linear Function	b_1^0	b_2^0	b_3^0	b_4^0
$\alpha_1^0 - \alpha_2^0$	3.0	3.0	3.0	3.0
$\mu^0 + \alpha_1^0$	15.0	15.0	15.0	15.0
$\mu^0 + 1/2(\alpha_2^0 + \alpha_3^0)$	19.5	19.5	19.5	19.5

The differences between the same elements in the four numerical solutions make it clear why no solution b^0 can be used as estimates for the unknown parameters in b .

This problem can be addressed, if we are not considering the single elements of a solution vector b^0 , but linear functions of these elements. Examples of such linear functions are shown in Table 2.6.

The values of the expressions shown in Table 2.6 are invariant to whatever solution b^0 is selected. Because this invariance statement is true for all solutions b^0 , these functions are of special interest which corresponds to

- $\alpha_1^0 - \alpha_2^0$: estimate of the difference between marker effect of locus G and locus H
- $\mu^0 + \alpha_1^0$: estimate of the general mean plus the marker effect of locus G
- $\mu^0 + 1/2(\alpha_2^0 + \alpha_3^0)$: estimate of the general mean plus mean effect of loci H and I

2.9.5.1 Definition of Estimable Functions

In summary the underlying idea of estimable functions are that they are linear functions of the parameters b that do not depend on the numerical solutions b^0 of the normal equations. Because estimable functions are functions of the parameters b , they can be expressed as $q^T b$ where q^T is a row vector. In a more formal way estimable functions can be described by the following definition.

Definition 2.2 (Estimable Function). A (linear) function of the parameters b is defined as **estimable**, if it is identically equal to some linear function of the expected value of the vector of observations y .

This means the linear function $q^T b$ is estimable, if

$$q^T b = t^T E(y)$$

for some vector t . That means, if there exists a vector t , such that $t^T E(y) = q^T b$, then $q^T b$ is said to be estimable. For our example shown in Table 2.4, the expected value of the observations of all animals with a genotype $G_1 G_1$ is obtained by

$$E(y_{G_1 G_1}) = \mu + \alpha_1$$

with $t^T = [1 \ 1 \ 1 \ 0 \ 0 \ 0]$ and $q^T = [1 \ 1 \ 0 \ 0 \ 0]$

2.9.6 Properties of Estimable Functions

Among the many properties we are here just listing the ones that are considered important. The complete list of properties can be found in [Searle, 1971].

- *Form of estimable function.* If $q^T b$ is estimable, then $q^T b = t^T E(y)$ for some t . By definition $E(y) = Xb$ and therefore, $q^T b = t^T Xb$. Because estimability is not a concept that depends on b , this result is true for all values of b . Therefore

$$q^t = t^T X$$

for some vector t .

- *Invariance to solutions b^0 .* If $q^T b$ is estimable, the linear function $q^T b^0$ is invariance to whatever solution of the normal equation

$$X^T X b^0 = X^T y$$

is used for b^0 . This is because

$$q^T b^0 = t^T X b^0 = t^T X G X^T y$$

where G is a generalized inverse of $X^T X$ and $X G X^T$ is invariant to G which means that it is the same for any choice of G .

2.9.6.1 Testing for Estimability

A given function $q^T b$ is estimable, if some vector t can be found, such that $t^T X = q^T$. For a known value of q , it might not be easy to find a vector t satisfying $t^T X = q^T$. Alternatively to finding a vector t , estimability of $q^T b$ can also be investigated by seeing whether q has the property that

$$q^T H = q^T$$

with $H = GX^T X$. A proof of that can be found in [Searle, 1971].

Chapter 3

Genomic Best Linear Unbiased Prediction (GBLUP)

In chapter 2 we introduced the fixed linear effects model to estimate additive genotypic values for SNP-Loci. In most real-world genomic datasets the number of SNP loci is larger than the number of observations. But from the point of view of quantitative genetics, we still assume that only a subset of the observed SNP-Loci is linked to a QTL and could therefore have an estimable effect on our trait of interest. Hence the original problem of estimating SNP-effect parameters is extended by a new problem of determining which SNPs are important for the expression of a given trait of interest.

3.1 Finding Relevant SNP Loci

Unfortunately it is not as easy as it may have seemed when we were looking at the monogenic model in Figure 1.5. When there are many SNPs that are observed and that are potentially influencing a trait, the different loci are interacting with each other and the distribution of the different trait values across the different genotypes is much more blurry. Furthermore when we use real-world observations of livestock animals these are phenotypic values which are influenced by many different environmental factors for which the phenotypic measurements all have to be corrected for.

This new problem of determining which SNP locus is linked to a QTL may sound like a not so difficult problem. But the number of possible SNP combinations is quite large. For a given number of k SNP loci the possible number of SNP combinations that might affect a trait is determined by the cardinality of the

powerset of k elements which is in the order of 2^k . Typical values of k might be $1.5 * 10^5$ and hence the number of possible combinations of any number SNP loci is a very large number. As a consequence of that a brute force approach where all possible combinations of SNP loci are tried cannot be used. Figure 3.1 tries to illustrate the problem of selecting important SNPs for a given trait.

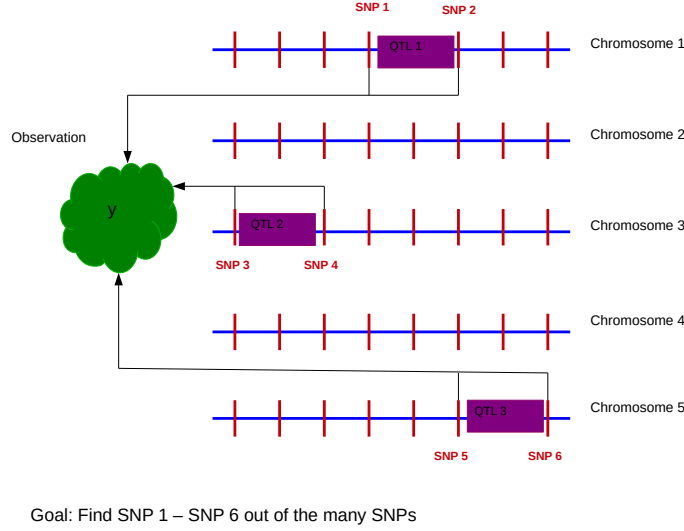


Figure 3.1: Finding SNP Loci Important For The Expression Of A Quantitative Trait

3.2 Stepwise Approach

In fixed linear effects model when the number of predictors is excessively large it is often desirable to find the subset of predictors that have a relevant effect on the response variable. Having too many predictors in a model decreases the power to predict future values of responses. To find a subset of relevant predictor variables out of a large set of predictors can be done with two stepwise approaches.

1. Forward selection
2. Backward elimination

In Statistics textbooks, the procedure of finding relevant predictor variables in a fixed linear model is called **model selection**. In what follows these two approaches are only described very shortly. Due to some practical problems with these techniques they did not find their way into the practical analyses of genomic data.

3.2.1 Forward Selection

The forward selection approach is described by the following step-wise procedure.

1. Start with an empty model \mathcal{M}_0 that contains no predictor variables but only an intercept.
2. Include the predictor variable that reduces the residual sum of squares the most.
3. Continue with step 2 until all predictor variables are included. This results in a series of models $\mathcal{M}_0 \subseteq \mathcal{M}_1 \subseteq \mathcal{M}_2 \subseteq \dots$
4. From the series of models select the one that optimizes a previously defined selection criterion.

As model quality criterion, different quantities can be used. Possible quantities are the Mallows C_p criterion, the Akaike Information Criterion (AIC) or Bayes Information Criterion (BIC).

3.2.2 Backward Elimination

Backward elimination can be viewed as the reverse process of forward selection. The following steps constitute the backward elimination algorithm

1. Start with the full model \mathcal{M}_0 containing all available predictor variables.
2. Drop the predictor variable that increases the residual sum of squares the least.
3. Continue with step 2 until all predictor variables are dropped. This results in a series of models $\mathcal{M}_0 \supseteq \mathcal{M}_1 \supseteq \mathcal{M}_2 \supseteq \dots$
4. From the series of models select the one that optimizes a defined selection criterion.

Backward elimination yields often better results compared to forward selection but it is more computationally expensive. In the case where the number of predictor variables (p) is larger than the number of observations (n), the full model cannot be fit to the data. Hence in that case forward selection would be a possible way to select an optimal model. Although if the number of predictors is very large and therefore many different predictor variables cause a similar reduction of the residual sum of squares, no unique series of models can be generated in the process of forward selection. The result of the forward selection depends on the order of the inclusion of the predictor variables.

3.3 Model Selection With Genomic Data

In real-world genomic data analyses, the number of predictors (p) can be as high as $1.5 * 10^5$. Therefore the backward elimination approach as described in 3.2.2 cannot be used because of the problem of $n \ll p$. Also the forward selection

approach does not yield a stable procedure for finding the subset of relevant SNPs. This has several reasons which are shortly described in the following subsections.

3.3.1 Fitting The Full Model

Fitting the full model with such a high number of predictors leads to the problem that the design matrix X will not have full column rank. The solution of the least squares normal equation then depends on a generalized inverse $(X^T X)^-$ of $X^T X$. Generalized inverses are not unique and furthermore for a given generalized inverse, there are infinitely many solutions that satisfy the normal equation coming out of least squares. Instead of the non-unique solutions, we have to focus on estimable functions (see section @ref(#asm-flem-estimable-functions)) of the solutions which are independent of the choice of a concrete solution. Although, even if it is possible to fit the full model of a genomic dataset, applying the backward elimination procedure is very time consuming and due to its greediness is not expected to result in a stable subset of relevant SNPs that has an influence on a given trait of interest.

3.4 Mixed Linear Effects Model

Based on the above described problems with the use of the fixed linear effects model for analyzing genomic data, animal breeders were looking for an alternative. In traditional genetic evaluation in animal breeding the BLUP animal model was used world-wide. The term **traditional genetic evaluation** refers to the prediction of breeding values based on phenotypic observation and pedigree relationships between animals in a given population. When looking at Figure 1.7 the traditional genetic evaluation is shown on the left side. The BLUP animal model is a mixed linear effects model where the breeding values of all animals in the populations are taken as random effects. In most of these traditional genetic evaluations the number of predicted breeding values exceeds the number of observations. This is possible due to the BLUP methodology which uses the variance-covariance matrix between the random effects to distribute the information of the observations also to predicted breeding values of animals which do not have any observations. In a BLUP animal model the variance-covariance matrix is proportional to the numerator relationship matrix A . We will see later that when using the genomic version of BLUP the matrix A will be replaced by its equivalent which is called the **genomic relationship matrix**.

Mixed linear effects models can be applied to genomic data using two different parametrisations. At this point, we are using the terminology proposed by [Fernando et al., 2016]. In the first approach SNP loci also referred to as markers are modeled as random effects. These models are called **marker effect models**

(MEM). In a second parametrisation, breeding values of animals corresponding to a linear combination of marker effects are modeled as random effects. This second type of models are referred to as **breeding value models** (BVM). Figure 3.2 illustrates the difference between the two types of models.

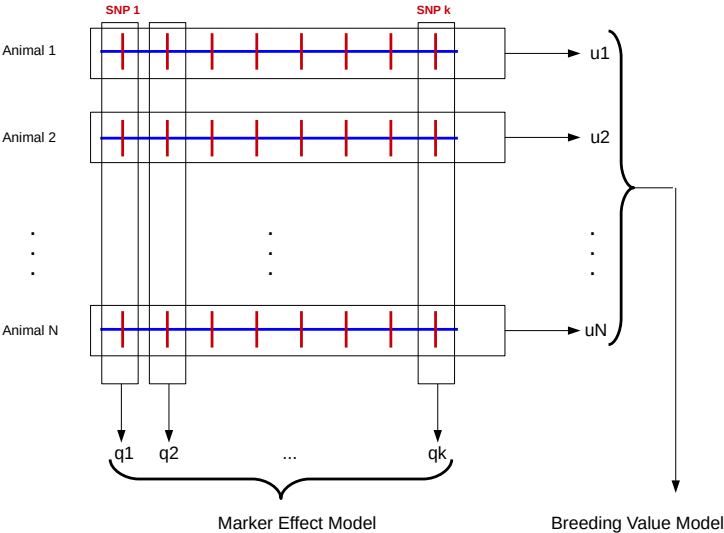


Figure 3.2: Two Types Of Mixed Linear Effects Models For Genomic Data

3.4.1 Marker Effect Models

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

$$y = 1_n \mu + Wq + e \quad (3.1)$$

where

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

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

3.4.2 Breeding Value Models

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

$$y = Xb + Zg + e \quad (3.2)$$

where

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

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

traditional BLUP animal model. The matrix G is called **genomic relationship matrix** (GRM). The variance $var(e)$ of the random error terms is given by $var(e) = I * \sigma_e^2$.

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

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

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

3.5 Genomic Relationship Matrix

The variance-covariance matrix between the genetic effects g in model (3.2) is proportional to the genomic relationship matrix G . Analogously to the traditional BLUP animal model where the variance-covariance matrix of the random breeding values is proportional to the numerator relationship matrix A .

3.5.1 Derivation of G

Because the traditional pedigree-based BLUP animal model is very well respected in animal breeding and the defined model (3.2) produces an analogy of the genomic evaluation model to the already known animal model the following properties of g and the genomic relationship matrix G are essential.

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

off-diagonal elements of animals that are related should have higher values than elements between unrelated animals.

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

3.5.2 Linear Combination of SNP Effects

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

$$g = U \cdot q \quad (3.3)$$

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

3.5.3 Deviation

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

Let us have a look at the random variable w which takes the SNP-genotype codes in the matrix W in the MEM model given in (3.1). Let us further assume that the SNP loci are in Hardy-Weinberg equilibrium. Then w can take the following values

$$w = \begin{cases} -1 & \text{with probability } (1-p)^2 \\ 0 & \text{with probability } 2p(1-p) \\ 1 & \text{with probability } p^2 \end{cases} \quad (3.4)$$

The expected value of w corresponds to

$$E[w] = (-1) * (1-p)^2 + 0 * 2p(1-p) + 1 * p^2 = -1 + 2p - p^2 + p^2 = 2p - 1 \quad (3.5)$$

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

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

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

$$\begin{aligned} E[g]_j &= [(1 - p_j)^2 * (-2p_j) + 2p_j(1 - p_j)(1 - 2p_j) + p_j^2(2 - 2p_j)] q_j \\ &= 0 \end{aligned} \quad (3.6)$$

3.5.4 Variance of Genetic Effects

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

$$\text{var}(g) = G * \sigma_g^2 \quad (3.7)$$

Computing the same variance-covariance matrix based on equation (3.3)

$$\text{var}(g) = U \cdot \text{var}(q) \cdot U^T \quad (3.8)$$

The variance-covariance matrix of the SNP effects is $\text{var}(q) = I * \sigma_q^2$. Inserting this into (3.8) we get $\text{var}(g) = UU^T \sigma_q^2$.

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

$$\sigma_g^2 = 2 \sum_{j=1}^m p_j(1 - p_j) \sigma_q^2 \quad (3.9)$$

Now we combine all relationships for $\text{var}(g)$ leading to

$$\text{var}(g) = G * \sigma_g^2 = UU^T \sigma_q^2 \quad (3.10)$$

In (3.10), σ_g^2 is replaced by the result of (3.9).

$$G * 2 \sum_{j=1}^m p_j(1 - p_j) \sigma_q^2 = UU^T \sigma_q^2 \quad (3.11)$$

Dividing both sides of (3.11) by σ_q^2 and solving for G gives us a formula for the genomic relationship matrix G

$$G = \frac{UU^T}{2 \sum_{j=1}^m p_j(1-p_j)} \quad (3.12)$$

3.6 How Does GBLUP Work

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

The BVM model given in (3.2) is a mixed linear effects model. The solution for the unknown parameters can be obtained by solving the mixed model equations shown in (3.13). 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.

$$\begin{bmatrix} X^T X & X^T Z & 0 \\ Z^T X & Z^T Z + G^{(11)} & G^{(12)} \\ 0 & G^{(21)} & G^{(22)} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{g}_1 \\ \hat{g}_2 \end{bmatrix} = \begin{bmatrix} X^T y \\ Z^T y \\ 0 \end{bmatrix} \quad (3.13)$$

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 (3.13) 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.

$$\hat{g}_2 = -(G^{22})^{-1} G^{21} \hat{g}_1 \quad (3.14)$$

Equation (3.14) 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 4

Least Absolute Shrinkage And Selection Operator (LASSO)

The linear model

$$y_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} + \epsilon_i \quad (4.1)$$

for an observation i (with $i = 1, \dots, n$) is used to model the relationship between the predictor variables x_{i1}, \dots, x_{ip} and a response variable y_i . In a fixed linear effect model the unknown parameters β_j (with $j = 1, \dots, p$) are estimated with least squares.

The values for β_j and for the error terms ϵ_i are unknown. We expect the predictor variables x_{i1}, \dots, x_{ip} to be known without any error. For a given dataset with n observations the model can be written in matrix-vector notation as

$$y = X\beta + \epsilon \quad (4.2)$$

4.1 Stochastic Error Component

The error terms ϵ are random effects in the model (4.2). The expected value $E(\epsilon) = 0$ and the variance $\text{var}(\epsilon) = I * \sigma^2$. This means the error terms of the different observations are assumed to be uncorrelated. The single error terms

ϵ_i are not so interesting, but the variance component σ^2 is besides the vector β of coefficients an additional unknown parameter.

4.2 Parameter Estimation

For the fixed linear effects model we have seen that the parameters β can be estimated using **Least Squares**. The condition implied by least squares corresponds to

$$\hat{\beta}_{LS} = \operatorname{argmin}_{\beta} \|y - X\beta\|^2 \quad (4.3)$$

where $\|\cdot\|$ corresponds to the euclidean norm. The solution of (4.3) for $\hat{\beta}_{LS}$ leads to the least squares normal equations given by

$$(X^T X) \hat{\beta}_{LS} = X^T y \quad (4.4)$$

If the matrix X has full column rank, the inverse of $(X^T X)$ exists and we can write the least squares estimator $\hat{\beta}_{LS}$ as

$$\hat{\beta}_{LS} = (X^T X)^{-1} X^T y \quad (4.5)$$

In all cases where X does not have full column rank, $(X^T X)$ is singular and one solution to (4.4) can be written in terms of a generalized inverse $(X^T X)^-$ of $(X^T X)$. This solution is called b_0 and can be written as

$$b_0 = (X^T X)^- X^T y \quad (4.6)$$

The solution in (4.6) are called b_0 and not $\hat{\beta}$ because b_0 is not an estimate of β . Furthermore, b_0 is not unique, because $(X^T X)^-$ is not unique. The solution b_0 can be used to generate estimates of estimable functions of β . An estimable function is a linear function of the parameters for which an estimator can be found from b_0 that is invariant to whatever solution is used for the normal equations. Any linear function of the parameters β is defined as estimable, if it is identically equal to a linear function of the expected value of the observations $E(y)$. This means that the linear function $q^T \cdot \beta$ is estimable if $q^T \cdot \beta = t^T \cdot E(y)$ for some vector t . For more details on estimable functions of parameters, we refer to section 4 of chapter 5 in [Searle, 1971].

Although, least squares provides us a tool to get estimates of either the unknown parameters or of estimable functions of the parameters, we are still not able to determine the important predictor variables out of a large set of available variables that characterize our responses in a given dataset. In statistics terms this problem is referred to as model selection or variable selection. In what follows, a method is described that solves the problem of variable selection.

4.3 Alternatives To Least Squares

The fixed linear effect model ((4.1)) is a very useful tool. Least Squares provides a very well established and an efficient method to estimate the unknown parameters. In the recent past, with the advent of a phenomenon called **Big Data** which stands for recent tendencies of systematically collecting large amounts of data, we have datasets available where each response has a very large number of potentially meaningful predictor variables. Finding the relevant predictors for a given response has become an important problem. Possible solutions can be divided into the following three classes of methods.

1. **Subset Selection:** Out of a set of p predictor variables, a subset of “relevant” variables are selected. All other variables are ignored. The relevant variables are often selected based on the significance of the hypothesis test against the Null-Hypothesis (H_0) of a given model coefficient β_j being 0 which means $H_0 : \beta_j = 0$.
2. **Regularisation (Shrinkage):** All p parameters are used in the model. The estimated coefficients are “forced” towards the origin. This process is called **shrinkage**. This causes a reduction of the variability of the estimates which is called *regularisation*.
3. **Dimension Reduction:** The p predictors are reduced to m linear combinations of the predictors. This reduction is achieved with techniques such as principal components analysis (PCA) or factor analysis (FA).

4.4 LASSO

Some procedures to estimate parameters can be found in more than one of the above three classes. Such methods are very popular, because they combine multiple of the above described properties which are desirable. An example of such a procedure is LASSO. LASSO is an abbreviation for Least Absolute Shrinkage and Selection Operator. It combines subset selection and regularization. The regularization is achieved by adding a penalty term to the least squares condition given in (4.3).

4.4.1 Regularisation With LASSO

With LASSO a penalty term is added to the least squares condition. The penalty term corresponds to $\lambda \sum_{j=1}^p |\beta_j|$. This term is a penalty, because when the sum of the absolute values of all β_j parameters is larger, the contribution due to that term is also larger and since the least squares criterion must be minimized, larger values are acting like a penalty. As a consequence of that parameter values with smaller absolute values are preferred and this leads to the desired effect of regularization. The resulting LASSO criterion can be expressed by the following formula.

$$\begin{aligned}
\hat{\beta}_{LASSO} &= \operatorname{argmin}_{\beta} \left\{ \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\} \\
&= \operatorname{argmin}_{\beta} \left\{ RSS + \lambda \sum_{j=1}^p |\beta_j| \right\}
\end{aligned} \tag{4.7}$$

4.4.2 Subset Selection With LASSO

The penalty term $\lambda \sum_{j=1}^p |\beta_j|$ in (4.7) is also responsible for the effect of subset selection. Due to the absolute value operator in the penalty term, some of the coefficients β_j in the linear model are explicitly set to zero. Why this effect occurs is shown in Figure 4.1.

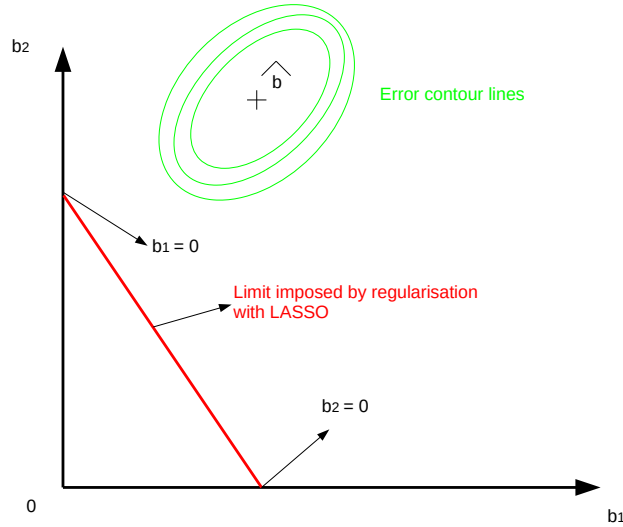


Figure 4.1: Subset Selection With LASSO

In Figure 4.1 a simple case with $p = 2$ parameters. The coefficients are called b and not β . With an infinite amount of data the coefficients b_j would be estimated with minimal error at the point which is labelled \hat{b} in Figure 4.1. The green ellipses denote the contour lines that have a constant value for the error. The red line symbolizes the limit which is created by the regularization effect imposed by LASSO. The regularization forces the parameter estimates to be inside of the triangle defined by the red line and the coordinate axes. Furthermore we want parameter estimates with minimal error. Hence the best estimate

is at the intersection of a green ellipsis with the red line. This intersection is very likely to happen at one of the corners of the regularization triangle. At these corners, one of the coefficients is set to zero which is the source of the desirable property of subset selection.

4.5 Determine λ

The penalty term in (4.7) contains the parameter λ . This parameter is used to determine the strength of the regularization and it has to be estimated from the data. One possibility to determine λ is via a procedure called **cross validation** (CV). In a cross validation the dataset is divided randomly into a training set and a test set. The complete data set is separated such that the test set is smaller than the training set. For a given training set with an assumed value of λ the other parameters are estimated. With the estimated parameters, the data in the test set are tried to be predicted. This is repeated for different values of λ and different separations into training and test sets. The value of λ with minimal prediction error is selected as optimal estimate for λ .

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

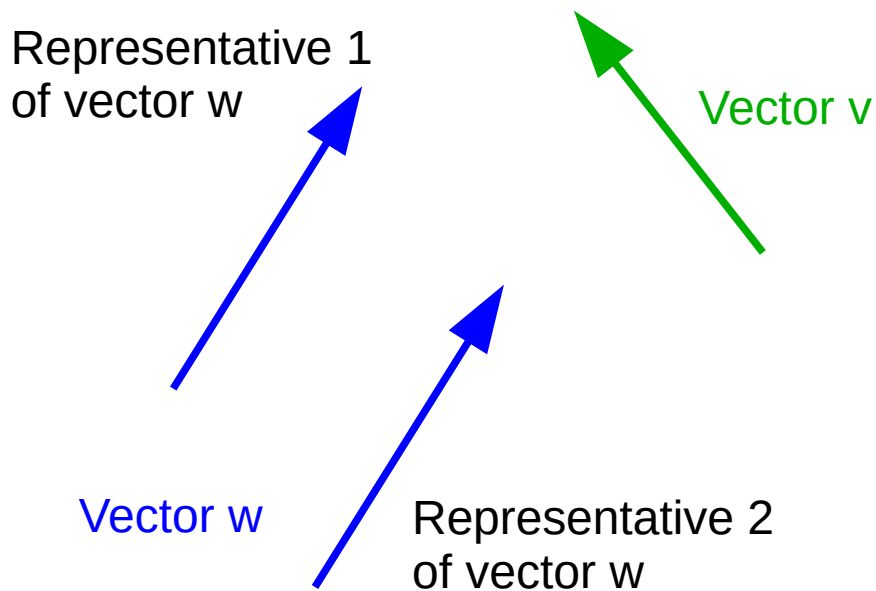
The material of this section is largely based on the video tutorial (https://youtu.be/fNk_zzaMoSs) from [3blue1brown, 2016]. We try to give a summarized transcript of the video. The vector is the fundamental building block of linear algebra. There are three different but related concepts about what vectors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

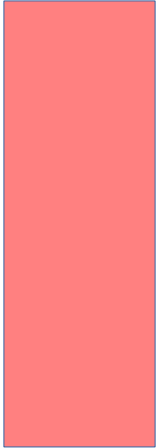
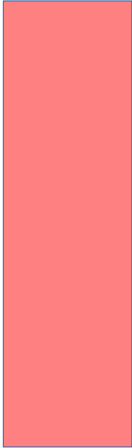
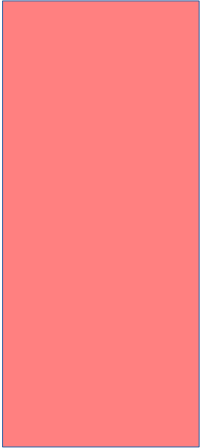
A.2.1 Physics Perspective

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

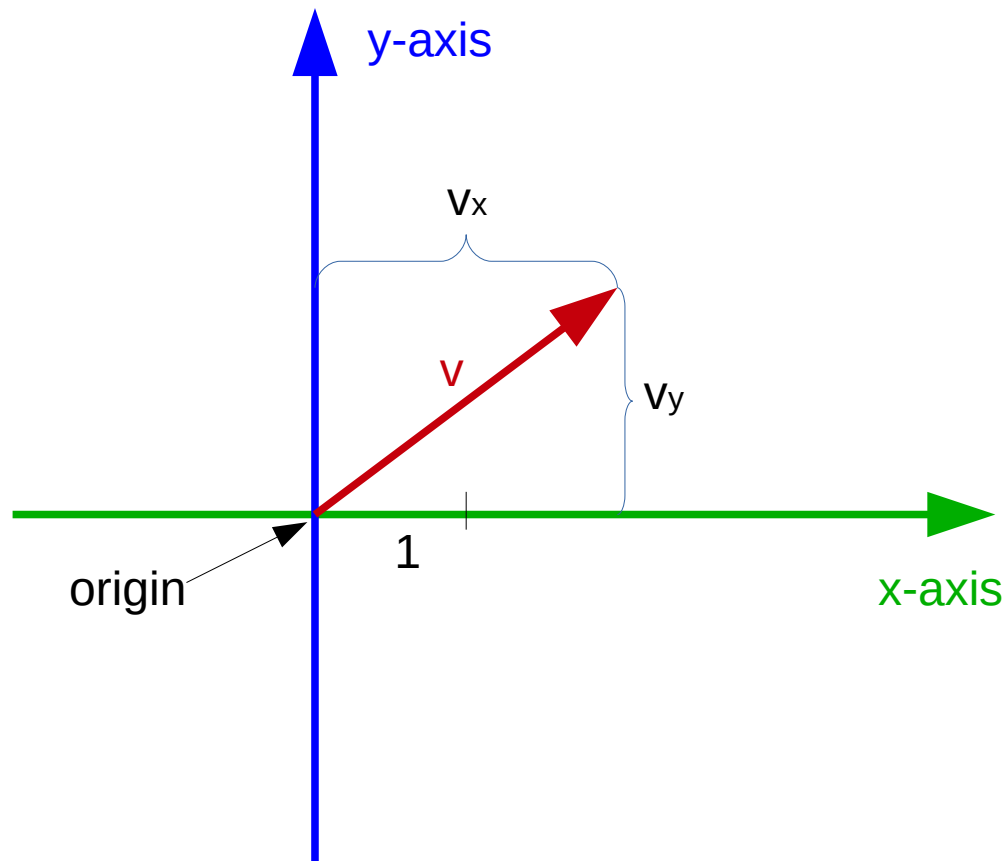
A.2.3 Geometric Context

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).

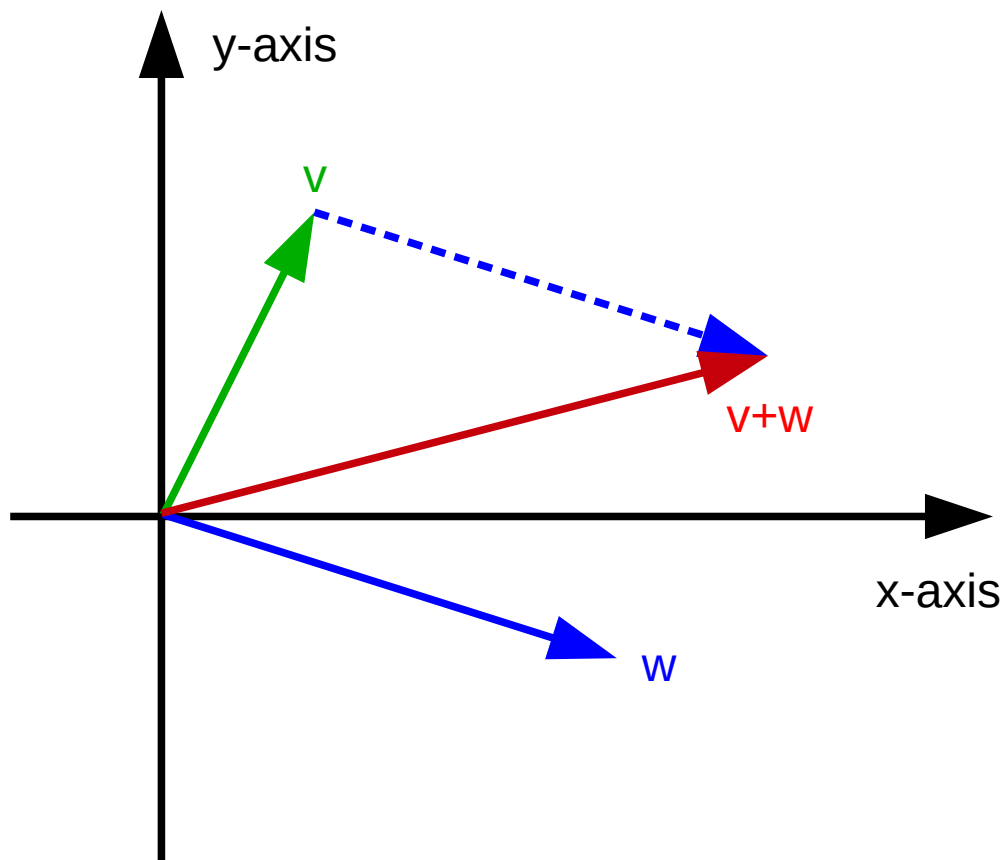


Figure A.1: Addition of two vectors

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

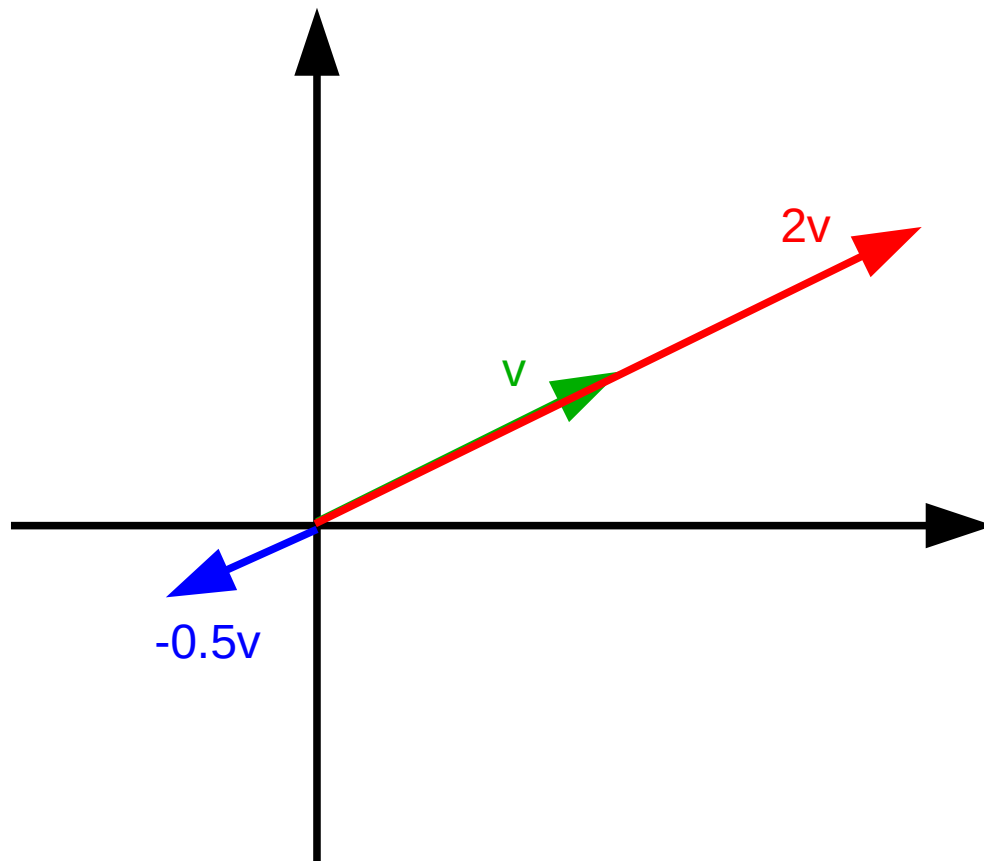
$$v = \begin{bmatrix} v_x \\ v_y \end{bmatrix}, w = \begin{bmatrix} w_x \\ w_y \end{bmatrix}$$

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.

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

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 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 invertable. 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 invertable $n \times n$ matrices A and B , then the following equations hold

1. $A^{-1}A = I$

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

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

1. A is invertable
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 invertable 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

$$A \cdot x = y \quad (\text{A.4})$$

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

$$\begin{aligned} A \cdot x &= y \\ A^{-1} \cdot A \cdot x &= A^{-1} \cdot y \\ I \cdot x &= A^{-1} \cdot y \\ x &= A^{-1} \cdot y \end{aligned} \quad (\text{A.5})$$

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

Basics in Quantitative Genetics

As already mentioned in section ??, the central dogma of molecular biology tells us that the genotype is the basis 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 (??). 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.

B.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 B.1).

B.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

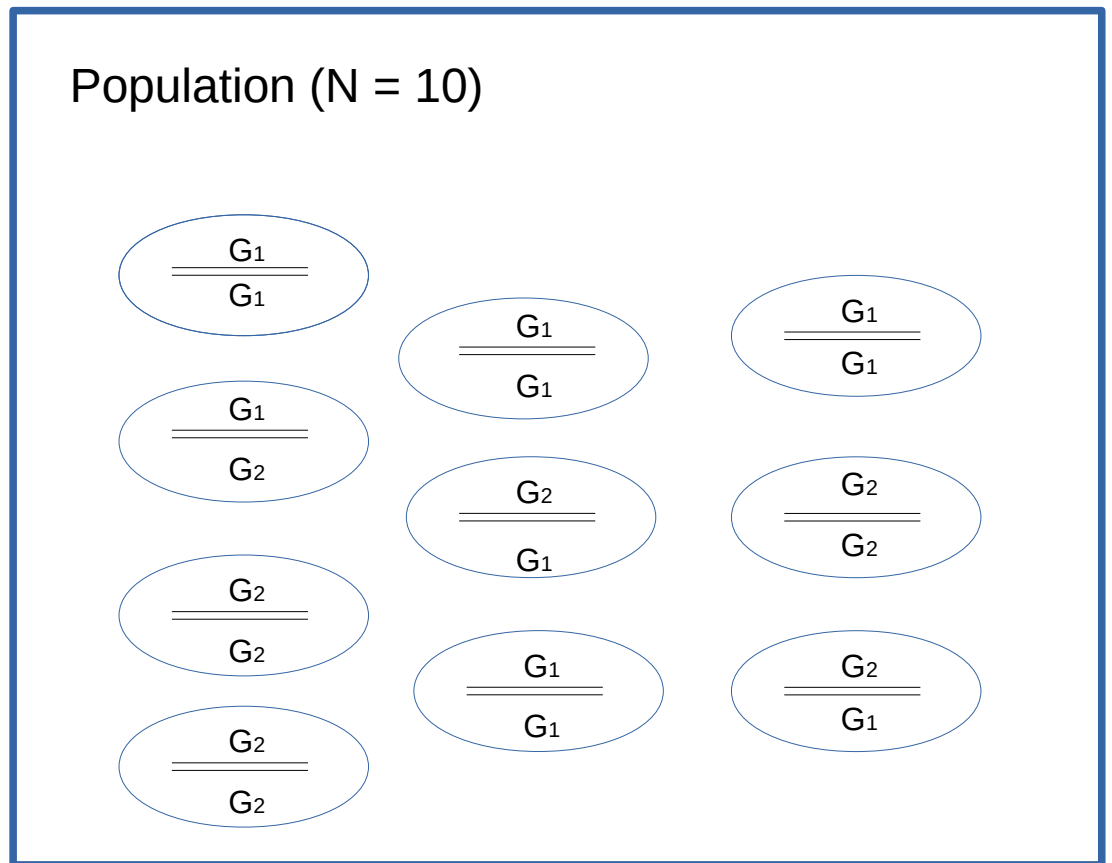


Figure B.1: Idealized Population With A Single Locus

are called the individuals **genotype**. All possible combinations of the two alleles 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**.

B.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 B.1, the **genotype frequencies** are

$$\begin{aligned} f(G_1G_1) &= \frac{4}{10} = 0.4 \\ f(G_1G_2) &= \frac{3}{10} = 0.3 \\ f(G_2G_2) &= \frac{3}{10} = 0.3 \end{aligned} \tag{B.1}$$

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

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

B.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

$$\begin{aligned} f(G_1) &= p \\ f(G_2) &= q = 1 - p \end{aligned} \tag{B.3}$$

During mating, we assume that in an idealized population alleles are combined independently. This leads to the genotype frequencies shown in Table B.1.

Summing up the heterozygous frequencies leads to

Table B.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	$f(G_1G_2) = p * q$	$f(G_2G_2) = q^2$

Table B.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

$$\begin{aligned}
 f(G_1G_1) &= p^2 \\
 f(G_1G_2) &= 2pq \\
 f(G_2G_2) &= q^2
 \end{aligned} \tag{B.4}$$

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 B.1, we can clearly say that the small example population is not in Hardy-Weinberg equilibrium.

B.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.

B.4.1 Genotypic Values

The values V_{ij} to each genotype G_iG_j are assigned as shown in Figure B.2.

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 B.2 summarizes the values for all genotypes.

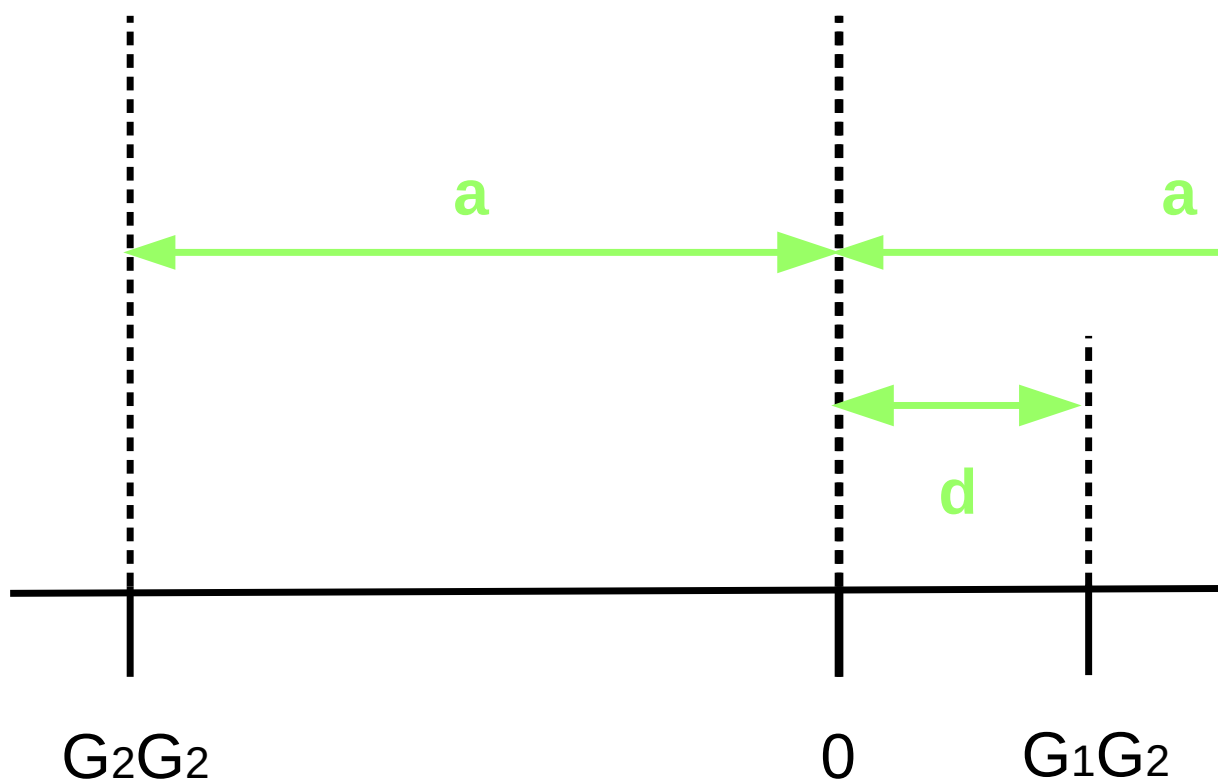


Figure B.2: Genotypic Values

B.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

$$\begin{aligned}\mu &= V_{11} * f(G_1G_1) + V_{12} * f(G_1G_2) + V_{22} * f(G_2G_2) \\ &= a * p^2 + d * 2pq + (-a) * q^2 \\ &= (p - q)a + 2pqd\end{aligned}\tag{B.5}$$

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 $\mu = 0$.

B.4.3 Breeding Values

The term **breeding value** is defined as shown in Definition B.1.

Definition B.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.

B.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

$$\mu_{11} = p * a + q * d \quad (\text{B.6})$$

Applying definition B.1, we can compute the breeding value (BV_{11}) for animal S as shown in equation (B.7) while using the results given by equations (B.6) and (B.5).

$$\begin{aligned} BV_{11} &= 2 * (\mu_{11} - \mu) \\ &= 2 (pa + qd - [(p - q)a + 2pqd]) \\ &= 2 (pa + qd - (p - q)a - 2pqd) \\ &= 2 (qd + qa - 2pqd) \\ &= 2 (qa + qd(1 - 2p)) \\ &= 2q (a + d(1 - 2p)) \\ &= 2q (a + (q - p)d) \end{aligned} \quad (\text{B.7})$$

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

B.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

$$\mu_{22} = pd - qa \quad (\text{B.8})$$

The breeding value BV_{22} corresponds to

$$\begin{aligned}
BV_{22} &= 2 * (\mu_{22} - \mu) \\
&= 2 (pd - qa - [(p - q)a + 2pqd]) \\
&= 2 (pd - qa - (p - q)a - 2pqd) \\
&= 2 (pd - pa - 2pqd) \\
&= 2 (-pa + p(1 - 2q)d) \\
&= -2p(a + (q - p)d)
\end{aligned} \tag{B.9}$$

B.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

$$\mu_{12} = 0.5pa + 0.5d - 0.5qa = 0.5[(p - q)a + d] \tag{B.10}$$

The breeding value BV_{12} corresponds to

$$\begin{aligned}
ZW_{12} &= 2 * (\mu_{12} - \mu) \\
&= 2 (0.5(p - q)a + 0.5d - [(p - q)a + 2pqd]) \\
&= 2 (0.5pa - 0.5qa + 0.5d - pa + qa - 2pqd) \\
&= 2 (0.5(q - p)a + (0.5 - 2pq)d) \\
&= (q - p)a + (1 - 4pq)d \\
&= (q - p)a + (p^2 + 2pq + q^2 - 4pq)d \\
&= (q - p)a + (p^2 - 2pq + q^2)d \\
&= (q - p)a + (q - p)^2d \\
&= (q - p)[a + (q - p)d]
\end{aligned} \tag{B.11}$$

B.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	$-2p\alpha$

B.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

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

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

$$\begin{aligned}
 BV_{11} - BV_{12} &= 2q\alpha - (q - p)\alpha \\
 &= (2q - (q - p))\alpha \\
 &= \alpha
 \end{aligned} \tag{B.13}$$

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**.

B.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.

$$\begin{aligned}
 V_{11} - BV_{11} &= a - 2q\alpha \\
 &= a - 2q[a + (q - p)d] \\
 &= a - 2qa - 2q(q - p)d \\
 &= a(1 - 2q) - 2q^2d + 2pqd \\
 &= [(p - q)a + 2pqd] - 2q^2d \\
 &= \mu + D_{11}
 \end{aligned} \tag{B.14}$$

$$\begin{aligned}
 V_{12} - BV_{12} &= d - (q - p)\alpha \\
 &= d - (q - p)[a + (q - p)d] \\
 &= [(p - q)a + 2pqd] + 2pqd \\
 &= \mu + D_{12}
 \end{aligned} \tag{B.15}$$

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

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

B.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
$G_i G_j$	V_{ij}	BV_{ij}	D_{ij}
$G_1 G_1$	a	$2q\alpha$	$-2q^2d$
$G_1 G_2$	d	$(q - p)\alpha$	$2pqd$
$G_2 G_2$	$-a$	$-2p\alpha$	$-2p^2d$

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 (B.16).

$$V_{ij} = \mu + BV_{ij} + D_{ij} \quad (\text{B.16})$$

Taking expected values on both sides of equation (B.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

$$\begin{aligned} E[V] &= E[\mu + BV + D] \\ &= E[\mu] + E[BV] + E[D] \\ &= \mu \end{aligned} \quad (\text{B.17})$$

From the last line in equation (B.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.

B.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 (B.18).

$$Var[X] = \sum_{x_i \in \mathcal{X}} (x_i - \mu_X)^2 * f(x_i) \quad (\text{B.18})$$

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 (B.18) to the genotypic values V_{ij} , we obtain the following expression.

$$\begin{aligned}\sigma_G^2 = Var[V] &= (V_{11} - \mu)^2 * f(G_1 G_1) \\ &+ (V_{12} - \mu)^2 * f(G_1 G_2) \\ &+ (V_{22} - \mu)^2 * f(G_2 G_2)\end{aligned}\quad (B.19)$$

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

Based on the decomposition of the genotypic value V_{ij} given in (B.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

$$\begin{aligned}\sigma_G^2 = Var[V] &= (BV_{11} + D_{11})^2 * f(G_1 G_1) \\ &+ (BV_{12} + D_{12})^2 * f(G_1 G_2) \\ &+ (BV_{22} + D_{22})^2 * f(G_2 G_2)\end{aligned}\quad (B.20)$$

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 (B.21). A more detailed derivation of σ_G^2 is given in the appendix (B.8) of this chapter.

$$\begin{aligned}\sigma_G^2 &= 2pq\alpha^2 + (2pqd)^2 \\ &= \sigma_A^2 + \sigma_D^2\end{aligned}\quad (B.21)$$

The formula in equation (B.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 (B.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 (B.27)) it is called dominance variance.

B.6 Extension To More Loci

When only a single locus is considered, the genotypic values (V_{ij}) can be decomposed according to equation (B.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.

B.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

$$V = V_A + V_B + I_{AB} \quad (\text{B.22})$$

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.

$$V = \mu + U + D + I \quad (\text{B.23})$$

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 B.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.

B.6.2 Interaction Variance

If genotypes at different loci show epistatic interaction effects as described in section B.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.

B.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 (B.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.

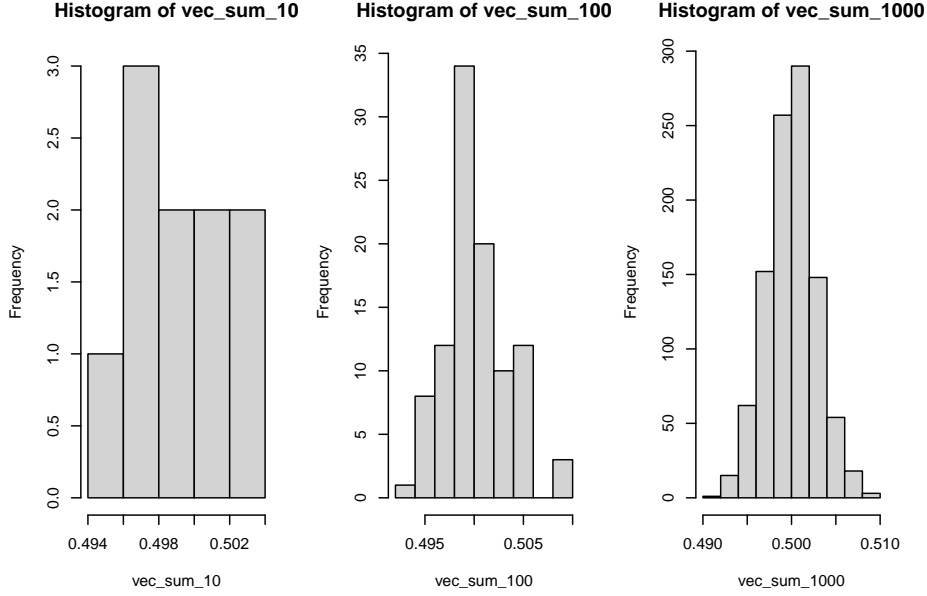


Figure B.3: Distribution of Sums of Different Numbers of Components

B.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.

B.8 Appendix: Derivations

This section shows how the genetic variance in equation (B.21) is computed.

$$\begin{aligned}
\sigma_G^2 &= (BV_{11} + D_{11})^2 * p^2 \\
&+ (BV_{12} + D_{12})^2 * 2pq \\
&+ (BV_{22} + D_{22})^2 * q^2 \\
&= (2q\alpha - 2q^2d)^2 * p^2 \\
&+ ((q - p)\alpha + 2pqd)^2 * 2pq \\
&+ (-2p\alpha - 2p^2d)^2 * q^2 \\
&= (4q^2\alpha^2 - 8q^3d\alpha + 4q^4d^2) * p^2 \\
&+ (q^2\alpha^2 - 2pq\alpha^2 + p^2\alpha^2 - 4(q - p)pqd\alpha + 4p^2q^2d^2) * 2pq \\
&+ (4p^2\alpha^2 + 8p^3d\alpha + 4p^4\alpha^2) * q^2 \\
&= 4p^2q^2\alpha^2 - 8p^2q^3d\alpha + 4p^2q^4d^2 \\
&+ 2pq^3\alpha^2 - 4p^2q^2\alpha^2 + 2p^3q\alpha^2 \\
&- 8p^3q^2d\alpha + 8p^2q^3d\alpha + 8p^3q^3d^2 \\
&+ 4p^2q^2\alpha^2 + 8p^3q^2d\alpha + 4p^4q^2d^2 \\
&= 4p^2q^2\alpha^2 + 4p^2q^4d^2 \\
&+ 2pq^3\alpha^2 + 2p^3q\alpha^2 \\
&+ 8p^3q^3d^2 \\
&+ 4p^4q^2d^2 \\
&= 2pq\alpha^2 (p^2 + 2pq + q^2) \\
&+ (2pqd)^2 (p^2 + 2pq + q^2) \\
&= 2pq\alpha^2 + (2pqd)^2 \\
&= \sigma_A^2 + \sigma_D^2
\end{aligned} \tag{B.24}$$

From the last two lines of (B.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.

$$\begin{aligned}
\sigma_A^2 &= Var[BV] = (BV_{11} - E[BV])^2 * f(G_1G_1) + (BV_{12} - E[BV])^2 * f(G_1G_2) + (BV_{22} - E[BV])^2 * \\
&= BV_{11}^2 * f(G_1G_1) + BV_{12}^2 * f(G_1G_2) + BV_{22}^2 * f(G_2G_2) \\
&= (2q\alpha)^2 * p^2 + ((q - p)\alpha)^2 * 2pq + (-2p\alpha)^2 * q^2 \\
&= 4p^2q^2\alpha^2 + (q^2\alpha^2 - 2pq\alpha^2 + p^2\alpha^2) * 2pq + 4p^2q^2\alpha^2 \\
&= 8p^2q^2\alpha^2 + 2pq^3\alpha^2 - 4p^2q^2\alpha^2 + 2p^3q\alpha^2 \\
&= 4p^2q^2\alpha^2 + 2pq^3\alpha^2 + 2p^3q\alpha^2 \\
&= 2pq\alpha^2 (2pq + q^2 + p^2) \\
&= 2pq\alpha^2
\end{aligned} \tag{B.25}$$

In the above derivation in (B.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

$$\begin{aligned}
 E[BV] &= BV_{11} * f(G_1 G_1) + BV_{12} * f(G_1 G_2) + BV_{22} * f(G_2 G_2) \\
 &= 2q\alpha * p^2 + (q - p)\alpha * 2pq + (-2p\alpha) * q^2 \\
 &= 2p^2 q\alpha + 2pq^2\alpha - 2p^2 q\alpha - 2pq^2\alpha \\
 &= 0
 \end{aligned} \tag{B.26}$$

Similarly to (B.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.

$$\begin{aligned}
 \sigma_D^2 &= D_{11}^2 * f(G_1 G_1) + D_{12}^2 * f(G_1 G_2) + D_{22}^2 * f(G_2 G_2) \\
 &= (-2q^2 d)^2 * p^2 + (2pqd)^2 * 2pq + (-2p^2 d)^2 * q^2 \\
 &= 4p^2 q^4 d^2 + 8p^3 q^3 d^2 + 4p^4 q^2 d^2 \\
 &= 4p^2 q^2 d^2 (q^2 + 2pq + p^2) \\
 &= 4p^2 q^2 d^2
 \end{aligned} \tag{B.27}$$

Bibliography

- 3blue1brown. What are vectors, 2016. URL https://youtu.be/fNk{__}zzaMoSs.
- M Adibuzzaman, P DeLaurentis, J Hill, and B D Benneyworth. Big data in healthcare - the promises, challenges and opportunities from a research perspective: A case study with a model database. *AMIA Annu Symp Proc*, 2017:384–392, 2017. ISSN 1942-597X. URL <https://www.ncbi.nlm.nih.gov/pubmed/29854102>.
- Peter Buehlmann and Martin Maechler. Computational Statistics Course, 2014.
- M. L. Chizzotti, F. S. Machado, E. E.L. Valente, L. G.R. Pereira, M. M. Campos, T. R. Tomich, S. G. Coelho, and M. N. Ribas. Technical note: Validation of a system for monitoring individual feeding behavior and individual feed intake in dairy cattle. *Journal of Dairy Science*, 98(5):3438–3442, 2015. ISSN 15253198. doi: 10.3168/jds.2014-8925. URL <http://dx.doi.org/10.3168/jds.2014-8925>.
- Bartosz Czech, Magdalena Fraszczak, Magda Mielczarek, and Joanna Szyda. Identification and annotation of breed-specific single nucleotide polymorphisms in *Bos taurus* genomes. *PLoS ONE*, 13(6):1–9, 2018. ISSN 15498328. doi: 10.1109/TCSI.2014.2341116. URL <https://doi.org/10.1371/journal.pone.0198419>.
- Alois Essl. *Statistische Methoden in der Tierproduktion*. Österreichischer Agrarverlag, Wien, 1987. ISBN 3-7040-0859-1.
- D. S. Falconer and Trudy F. C. Mackay. *Introduction to Quantitative Genetics*. Addison Wesley Longman Limited, Essex, 4 edition, 1996. ISBN 0582-24302-5.
- Rohan L Fernando, Hao Cheng, and Dorian J Garrick. An efficient exact method to obtain GBLUP and single-step GBLUP when the genomic relationship matrix is singular. *Genetics Selection Evolution*, 48(1):80, dec 2016. ISSN 1297-9686. doi: 10.1186/s12711-016-0260-7. URL <http://gsejournal.biomedcentral.com/articles/10.1186/s12711-016-0260-7>.
- Daniel Gianola, Gustavo De Los Campos, William G. Hill, Eduardo Manfredi, and Rohan Fernando. Additive genetic variability and the Bayesian alphabet.

- Genetics*, 183(1):347–363, 2009. ISSN 00166731. doi: 10.1534/genetics.109.103952.
- Sasha Issenberg. How Obama’s Team Used Big Data to Rally Voters. *MIT Technology Review*, 116(1):38–49, 2013.
- David J. Lilja. *Linear regression using R : an introduction to data modeling*. University of Minnesota Libraries Publishing, Minneapolis, 2016. ISBN 9781946135001. URL <https://open.umn.edu/opentextbooks/textbooks/linear-regression-using-r-an-introduction-to-data-modeling>.
- John R Mashey. Big Data ... and the Next Wave of InfraStress, 1998. URL <http://static.usenix.org/event/usenix99/invited{ }talks/mashey.pdf>.
- T H E Meuwissen, B J Hayes, and M E Goddard. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics*, 157:1819–1829, 2001.
- G E Moore. Cramming more components onto integrated circuits. *Electronics*, 38, 1965.
- Kaspar Nipp and Daniel Stoffer. *Lineare Algebra*. vdf Hochschulverlag, Zurich, 5 edition, 2002. ISBN 3 7281 2818 X.
- Sameer D. Pant, Flavio S. Schenkel, Chris P. Verschoor, and Niel A. Karrow. Use of breed-specific single nucleotide polymorphisms to discriminate between Holstein and Jersey dairy cattle breeds. *Animal Biotechnology*, 23(1):1–10, 2012. ISSN 10495398. doi: 10.1080/10495398.2012.636224. URL <https://doi.org/10.1080/10495398.2012.636224>.
- R Core Team. R: A language and environment for statistical computing., 2018. ISSN 16000706. URL <https://www.r-project.org/>.
- L. R. Schaeffer. Strategy for applying genome-wide selection in dairy cattle. *J Anim Breed Genet*, 123(123):218–223, 2006. ISSN 09312668. doi: 10.1111/j.1439-0388.2006.00595.x.
- S R Searle. *Linear Models*. John Wiley & Sons, New York, wiley clas edition, 1971. ISBN 0-471-18499-3.
- G. E. Seidel, Jr. Brief introduction to whole-genome selection in cattle using single nucleotide polymorphisms. *Reproduction Fertility and Development*, 22(1):138–144, 2010. ISSN 1031-3613. doi: 10.1071/RD09220. URL <https://doi.org/10.1071/RD09220>.
- P.M. M VanRaden. Efficient methods to compute genomic predictions. *J. Dairy Sci.*, 91(91):4414—4423, nov 2008. ISSN 00220302. doi: 10.3168/jds.2007-0980. URL <https://www.sciencedirect.com/science/article/pii/S0022030208709901?via{ }%3Dihubhttp://dx.doi.org/10.3168/jds.2007-0980>.
- Contributors Wikipedia. Big data - Wikipedia, 2019. URL <https://en.wikipedia.org/w/index.php?title=Big{ }data{ }&oldid=881938730>.