Applied Statistical Methods in Animal Sciences

Peter von Rohr

2020-02-07

Contents

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 analyse 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 development of the hardware manufacturing industry is called **Moore's Law** and 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 analyse large amounts of data with moder methods can be done by almost anyone. 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 analyse 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 a topic 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 is the fourth edition overall and the first time that the course is taught in English. 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 analyse 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 presidential election campaigns of Barack Obama were examples of how `Big Data` was used to access behaviors of voters [Issenberg, 2013]. 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. The Swiss TV news show `10 vor 10` showed on the 7^{th} Feb. 2020 how a data journalist managed to build a face recognition system. He used open-source software together with portrait pictures from politicians and a database of pictures downloaded from the social network instagram. The face recognition program was able to find several politicians on the pictures obtained from instagram. The complete story is available under https://www.srf.ch/news/schweiz/automatische-gesichtserkennung-so-einfach-ist-es-eine-ueberwachungsmaschine-zu-bauen. These examples show that data can be used for 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 influences 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.

1.1 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 1.1.

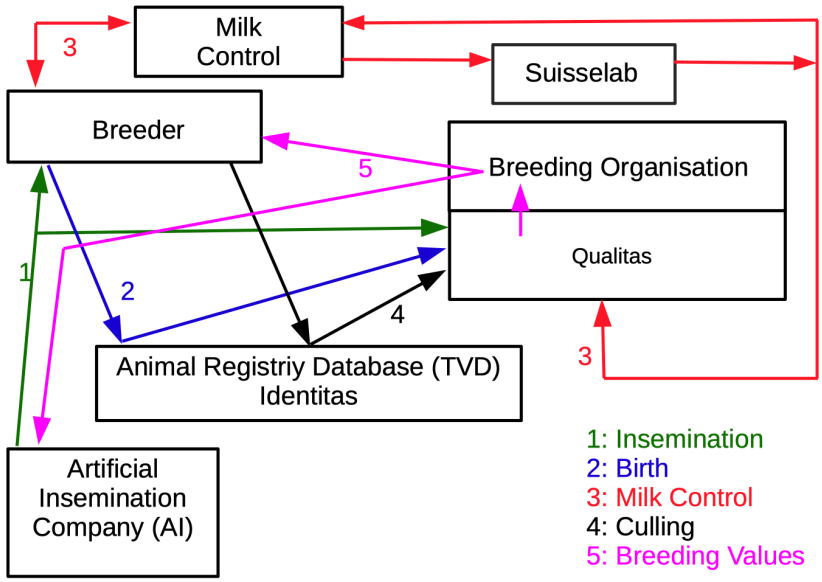

Figure 1.1: Data Flow in an Animal Breeding Program

1.2 Genomic Selection

The data flow shown in Figure 1.1 contains the traditional evaluation of data to result in predicted breeding values. But it is missing the newest development 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 1.2.

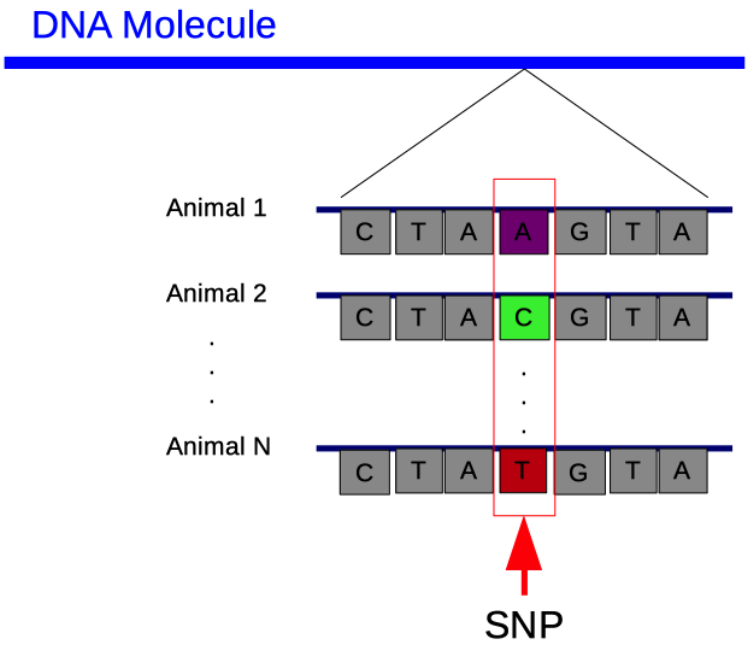

Figure 1.2: Genetic Configuration of a Single Nucleotide Polymorphism (SNP)

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 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 1.3.

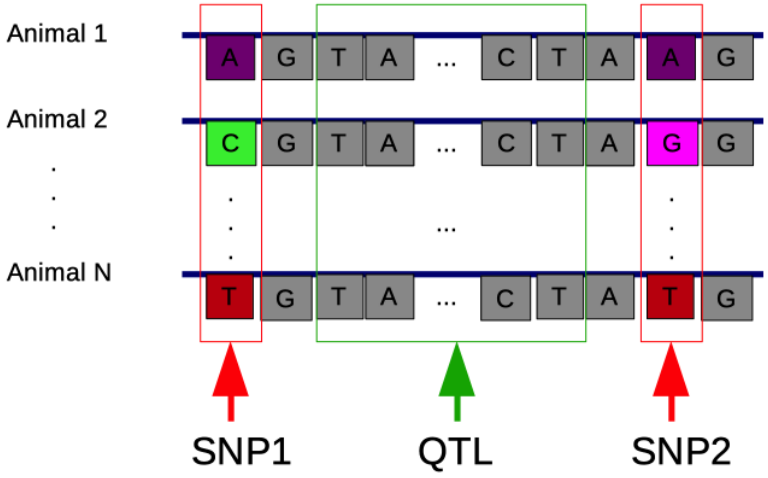

Figure 1.3: Two SNPs flanking a QTL

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 1.3 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

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

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.

1.3 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 1.4.

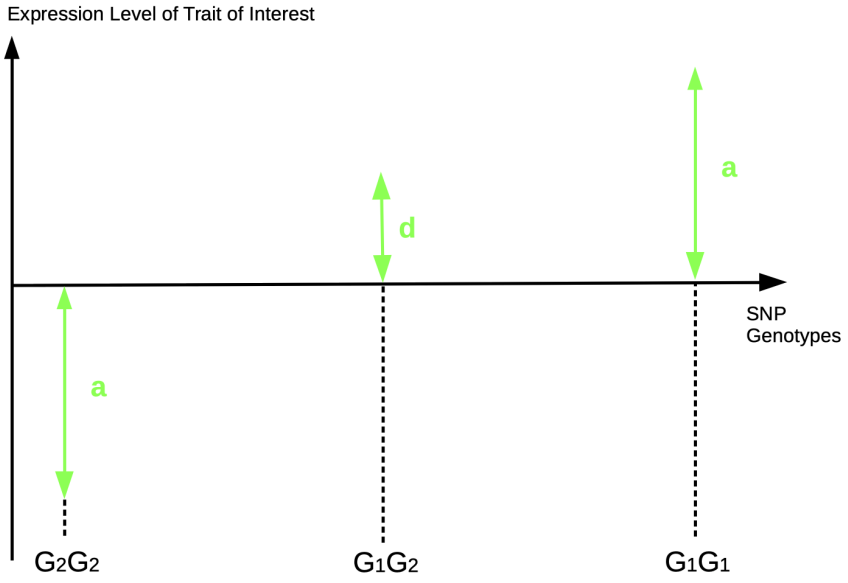

Figure 1.4: Single-Locus Model for a Quantitative Trait

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

1.4 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, 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 measurements[1] 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 1.5 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).

1.5 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

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

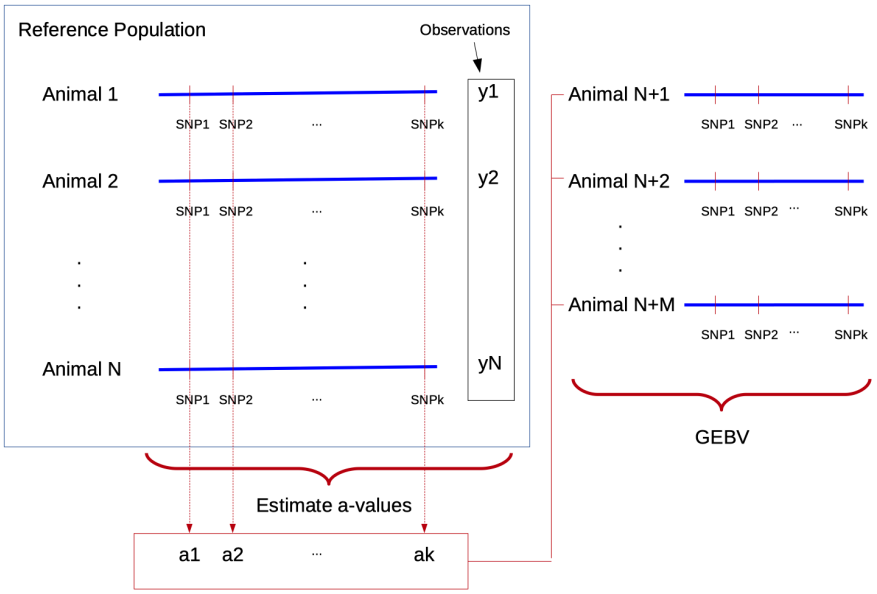

Figure 1.5: Two Step Approach To Estimate Genomic Breeding Values

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.

1.6 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 1.6.

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]. 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

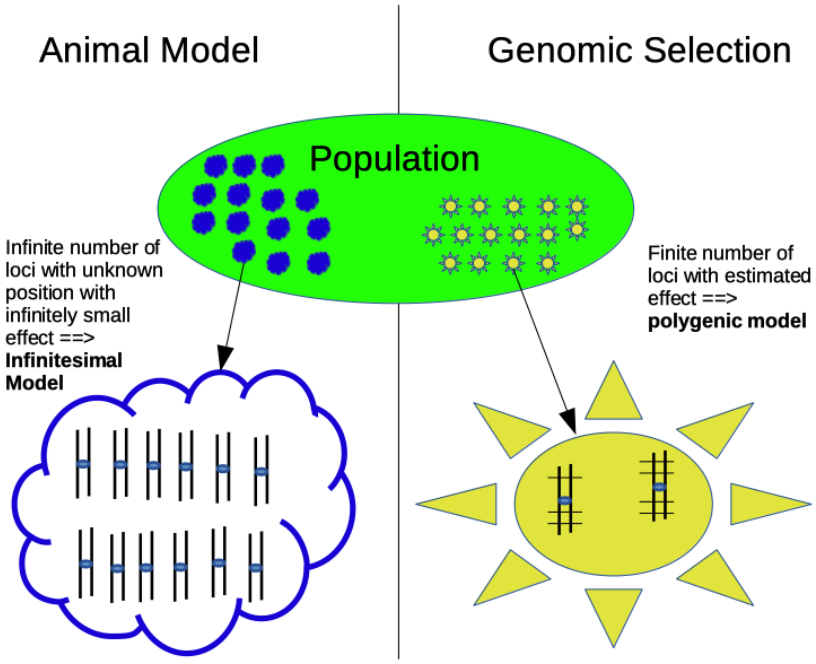

Figure 1.6: Infinitesimal Versus Polygenic Model

Bayesian methods are important because they are used in practical breeding programs of Swiss Dairy cattle.