Applied Statistical Methods in Animal Sciences

Spring Semester 2024

Peter von Rohr

2024-01-28

Table of 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 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, Hayes, and Goddard 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 rather short. 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 2022).

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 2022).

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.