Applied Statistical Methods In Animal Science

Peter von Rohr

22.02.2021

Administration

- Course: 2 hours of lecture (2 V)
- Plan: 2 V → 1 U + 1 V (i.e., 1 hour of lecture intersperced with time to do exercises)
- Exercises: Work on problems in R
- Material: course notes, slides, solution to exercises
- Exam: written, date: 31.05.2021, 08:15-09:00

Exercise platform, w02

Objectives

The students

- are familiar with the properties of **fixed linear effects models**
- are able to analyse simple data sets
- know why least squares cannot be used for genomic selection.
- know the statistical methods used in genomic selection, such as
  - BLUP-based approaches,
  - Bayesian procedures and
  - LASSO.
- are able to solve simple exercise problems using the statistical framework R.

Program

Week	Date	Topic
1	22.02	Introduction to Applied Statistical Methods
2	01.03	Linear Fixed Effect Models
3	08.03	GBLUP - Marker-Effects Models
4	15.03	GBLUP - Breeding Value Models
5	22.03	Lasso
6	29.03	Bayesian Approaches
7	05.04	**Easter Monday**
8	12.04	Introduction to Genetic Evaluation of Livestock
9	19.04	Model Selection
10	26.04	Variance Components
11	03.05	Genetic Groups and Longitudinal Data
12	10.05	Genomic Selection
13	17.05	Questions, Test Exam
14	24.05	**Pfingstmontag**
15	31.05	Exams

Applied Statistics

for both courses

Information

- Website: https://charlotte-ngs.github.io/gelasmss2021/
- Topics for master thesis: https://charlotte-ngs.github.io/gelasmss2021/misc/MasterThesisTopics_SS2021.html
- Exam: 31.05.2021 08:15 – 09:00

This Course

Bachelor Statistics: Multiple Linear Regression (MLR)
Applied Statistics: Aim: Further develop the concepts started in MLR

- Use dataset that is used to predict genomic breeding values and introduce four methods

1. Fixed Linear Effects Models - Least Squares Parameter estimation
2. GBLUP - genomic version of BLUP
3. LASSO - still fixed linear effects model, but modified parameter estimation
4. Bayesian approach to estimate unknown parameter

Methods 2, 3 and 4 are solving problems found with method 1

Assumption: Population of livestock animals. From animals of this population, we have a dataset of observations, and genomic information

Significance

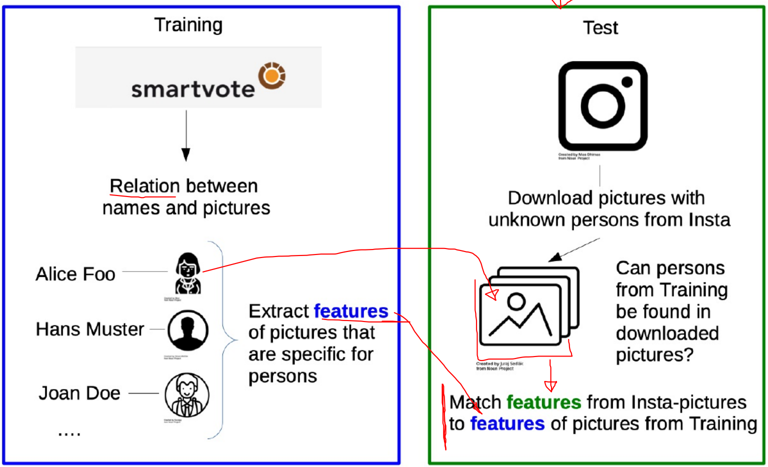
Ex: Corona Pandemic:
Goverments: Develop measures and rules of behavior based on the number of infections, R-value which the reproduction number

- Why is this important?
- Is this only relevant for animal breeding?
- What about the rest of animal science?
- General trend of collecting data has led to development of `Big Data`
- Examples
  - Presidential campains in the US
  - Health care
  - Face recognition
  - Agriculture: Smart Farming
  - Animal Science

Face Recognition

Swiss TV (SRF): 10 vor 10 in 2019

Face-Recognition



Training

smartvote

Relation between names and pictures

Alice Foo

Hans Muster

Joan Doe

....

Extract **features** of pictures that are specific for persons

Test

Download pictures with unknown persons from Insta

Can persons from Training be found in downloaded pictures?

Match **features** from Insta-pictures to **features** of pictures from Training

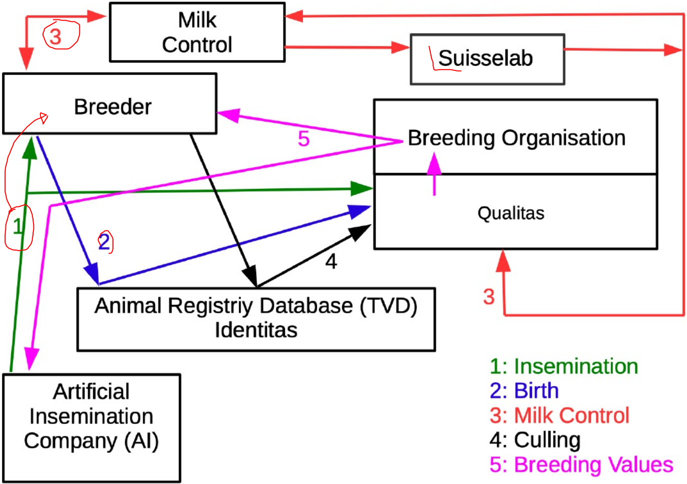
Traditional Animal Breeding

traditional or "pre-genomic" era (before 2006)
==> breeding values are predicted only based
on phenotypic information and pedigree data

- Before 2006
- Data collected for other purposes were used to predict breeding values
- Predicted breeding values as side-product

Data Logistics

After birth of a calf, lactation starts



Milk Control
3
Suisselab
Breeder
5
Breeding Organisation
Qualitas
1
2
4
3
Animal Registriy Database (TVD) Identitas
Artificial Insemination Company (AI)

1: Insemination
2: Birth
3: Milk Control
4: Culling
5: Breeding Values

3 times a year:
April, August, Dec

Genomic Selection

- Same goal as in traditional breeding: Find animals with best genetic potential as parents of next generation
- New: use additional source of information
- **Genomic** information
  - spread accross whole genome
  - single nucleotide polymorphisms (SNP)
- Introduction:

“> Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829”

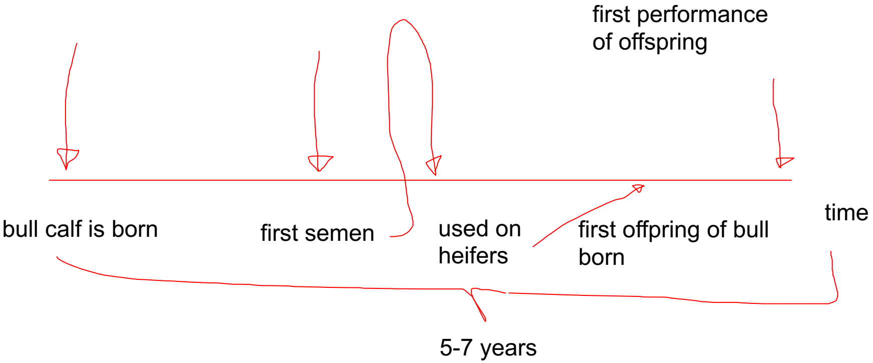
- Popularisation:

Use genomic selection to save about 90% of the total costs of cattle breeding program

“> L. R. Schaeffer. Strategy for applying genome-wide selection in dairy cat- tle. Journal of Animal Breeding and Genetics, 123(4):218–223, 2006. ISSN 09312668. doi: 10.1111/j.1439-0388.2006.00595.x.”

Tradtional breeding programs in dairy cattle:

- selection of bulls is based on evaluation of daughter performance
- most important traits can only be observed in cows



first performance of offspring

bull calf is born

first semen

used on heifers

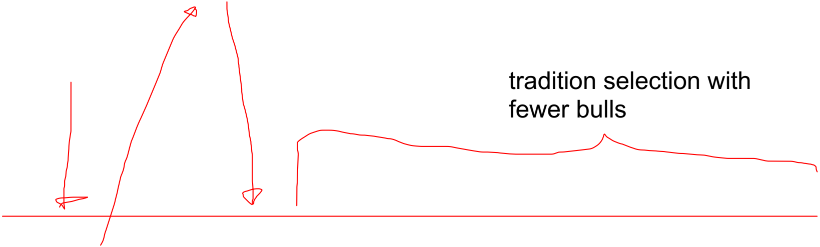
first offpring of bull born

time

5-7 years

Progeny tests of bulls: start with 300-400 bulls in test, kept 15-20

Breeding program with Genomic Selection



Hair sample and is analysed for SNP in the lab

tradition selection with fewer bulls

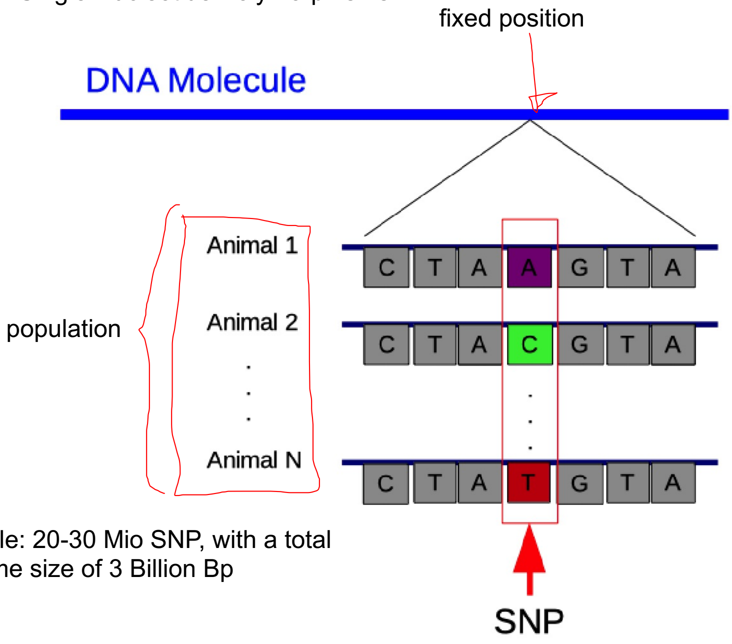
birth of bull

Genomic breeding value 3-6 months after birth
==> first selection of bull

time

Cost saving: Reduction of time until the first selection decision from 7 years to 6 months.

SNP Single Nucleotide Polymorphisms



fixed position

DNA Molecule

population

Animal 1 C T A A G T A

Animal 2 C T A C G T A

.
.
.

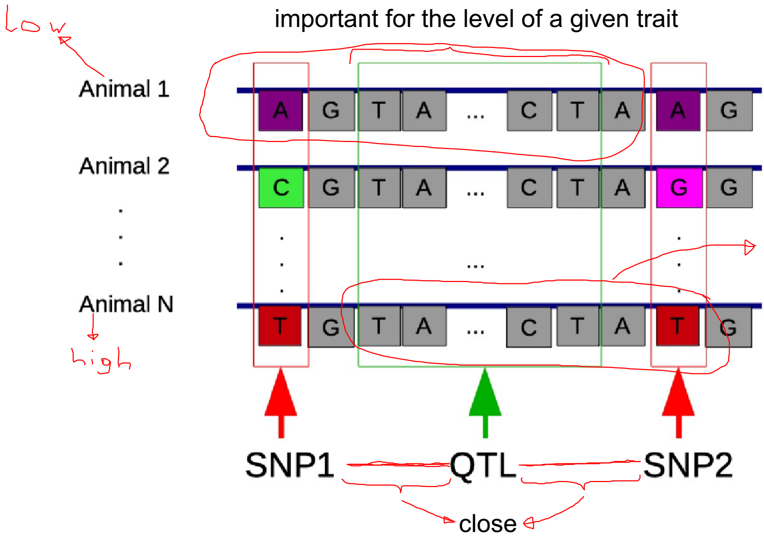
Animal N C T A T G T A

SNP

in cattle: 20-30 Mio SNP, with a total genome size of 3 Billion Bp

QTL

Quantitative Trait Locus, with unknown positions



important for the level of a given trait

Low

Animal 1 A G T A ... C T A A G

Animal 2 C G T A ... C T A G G

Animal N T G T A ... C T A T G

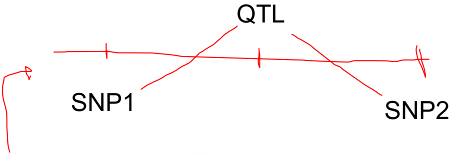
high

association between QTL allel and SNP-allel

SNP1 QTL SNP2

close

Linkage

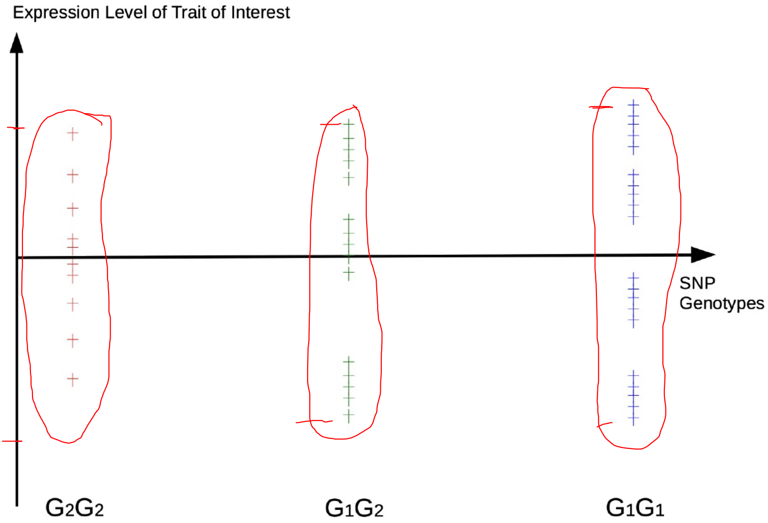


- Flanking SNPs and QTL not independent passed on from parents to progeny
- Favorable QTL-allele linked with a given SNP-allele
- QTL is unknown, but use SNPs close to QTL as information for selection

Monogenic Model

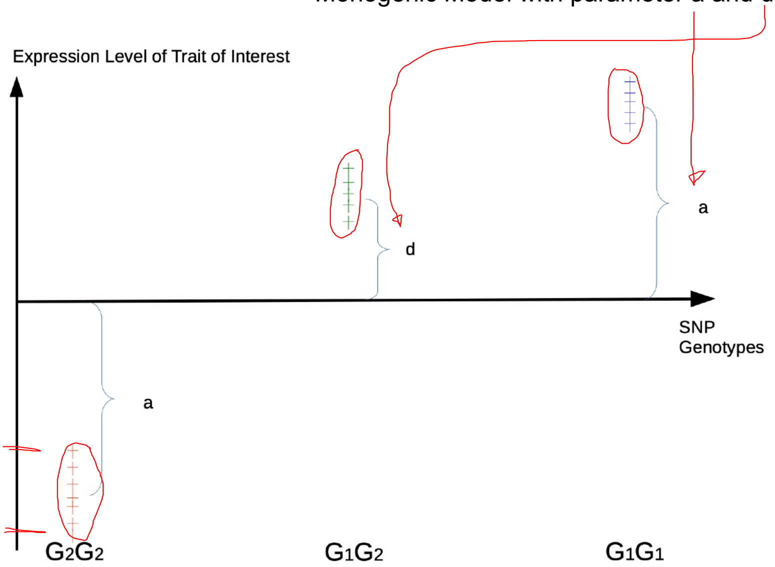
- Assume quantitative trait is influenced by one locus only
- Locus is bi-allelic → two alleles (G_1 and G_2) and three genotypes
- Look at Distribution of trait values for three different genotypes

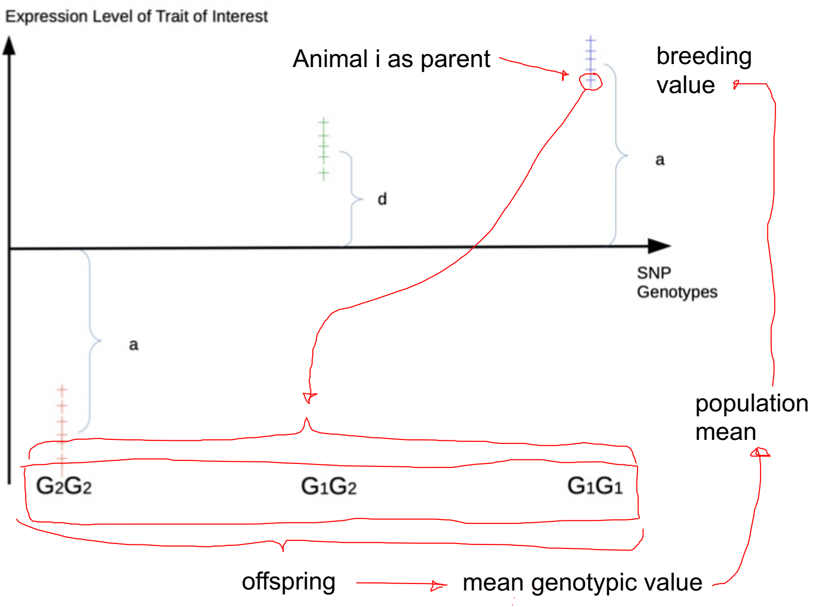
Distribution No Effect



Distribution With Effect

Monogenic Model with parameter a and d





Animal i as parent

breeding value

population mean

offspring

mean genotypic value

Mono-genic Breeding Value and Direct Genomic Breeding Value

* Breeding Value: Mono-genic, that means single locus

G1G1= $2q\alpha$

G1G2= $(q-p)\alpha$

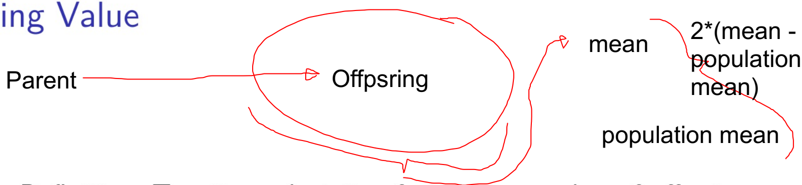
G2G2= $-2p\alpha$

p, q are allele frequencies
f(G1) = p, f(G2) = q
Assume, d = 0 ==> \alpha = a

* Direct Genomic Breeding Value: Sum of marker effects

Marker effects correspond to the a-values
Assume, that p is small and d=0, then ranking of animals according to Direct genomic breeding value and the mono-genic breeding value will be the same.

Breeding Value



Parent

Offpsring

mean

2*(mean - population mean)

population mean

- Definition: Two times deviation from large number of offspring from population mean
- Assume: Hardy-Weinberg equilibrium
- Compute population mean as expected value of genotypic values
- Compute expected genotypic value of offspring for each of the three parental genotypes
- Assume purely additive loci, hence $d = 0$

Genomic Breeding Value

- Take into account many loci
- Approximate unknown QTL with linked SNP
- Estimate *a*-effects from monogenic model
- Compute genomic breeding values for all loci based on *a* effects

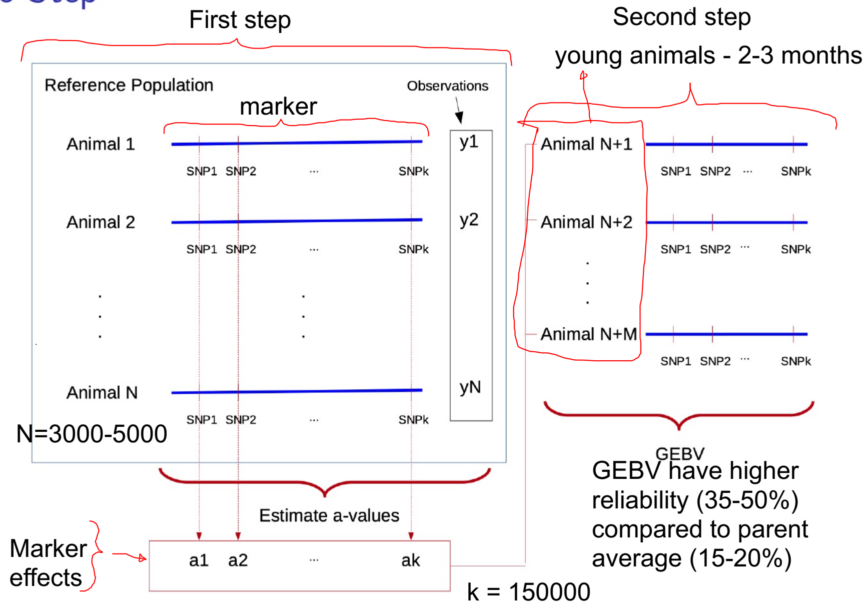
Two Approaches

also used for Swiss Beef Cattle

1. Two Step Procedure (used currently in Swiss Dairy Cattle)
2. Single Step

starting to develop analyses with single step

Two Step



N=3000-5000

k = 150000

GEBV have higher reliability (35-50%) compared to parent average (15-20%)

Advantages of 2-step approach

* Easy computation of GEBV for young animals
* Done every 2 weeks

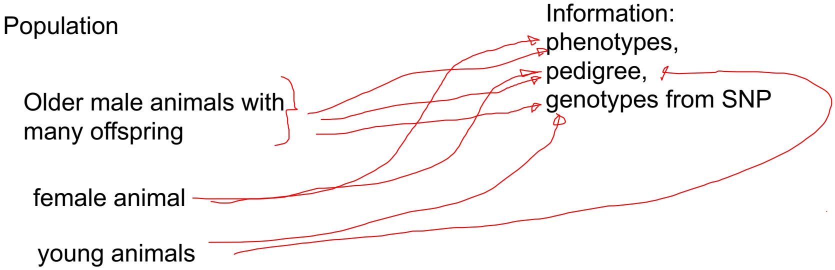
Problems with 2-step approach

* heavily depends on the availability of a good reference population
* reliabile estimates of marker effects
* For new traits (health traits, mastitis, ketosis, feed intake) with only few data, it is difficult to come up with a reference population that is large enough
* Wait for 2-3 years

Single Step

Philosophy: Combine all information

- Combine all information into one single BLUP-based analysis
- Problem: Determine covariance between animals with and without genomic information



Population

Older male animals with many offspring

female animal

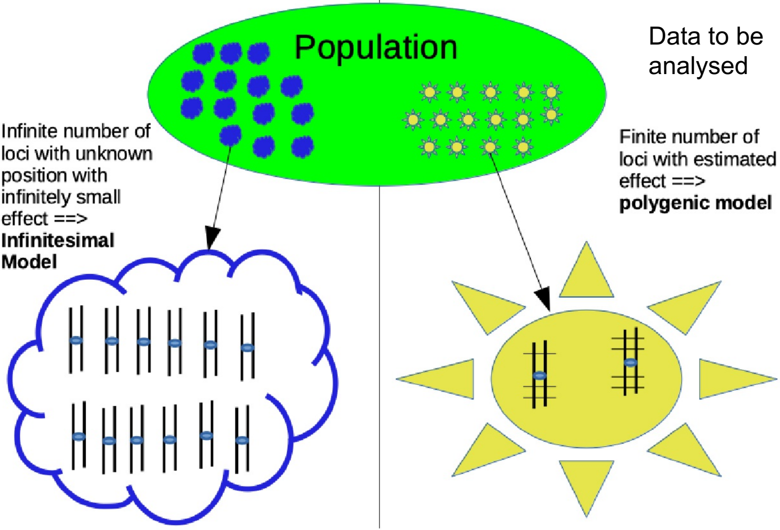
young animals

Information:
phenotypes,
pedigree,
genotypes from SNP

Summary: Traditional versus Genomic Selection

Animal Model

Genomic Selection



Population

Data to be analysed

Infinite number of loci with unknown position with infinitely small effect ==> **Infinitesimal Model**

Finite number of loci with estimated effect ==> **polygenic model**