Aim: Predict genomic breeding values
> Problems with least squares estimatimation in fixed linear effect models
> GBLUP with mixed linear effect models
> LASSO
> Further approach: Use Estimation techniques from Bayesian Statistics

Bayesian Approaches

Peter von Rohr

2021-03-29

Statistics

Frequentists

Bayes → measure of uncertainty

The world of statistics is divided into

- **Frequentists** and
- **Bayesians**

$$Pr = \frac{\#\ success}{\#\ total}$$

Divergence in

- understanding of probability
- differentiation between components of a model and the data
- techniques to estimate parameters

for us: marker effects or genomic breeding values

F vs B

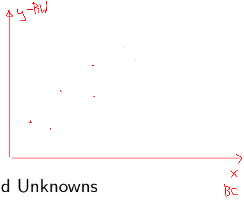
Topic	Frequentists	Bayesians
Probability	Ratio between cardinalities of sets	Measure of uncertainty
Model and Data	Parameter are unknown, data are known	Differentiation between knowns and unknowns missing data
Parameter Estimation	ML or REML are used for parameter estimation	MCMC techniques to approximate posterior distributions

ML: maximum likelihood
REML: restricted ML

MCMC: Markov Chain Monte Carlo

Linear Model

Example: Regression, with BW and BC



Frequentist:
Modell: y = Xb + e, with b and e unknown
Data:

BC	BW

$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$$

Table 1: Separation Into Knowns And Unknowns

Term	Known	Unknown
y_i	X	
x_1	X	
β_0		X
β_1		X
σ^2	X	

Bayesian Analysis

Assumption for first analysis

Example Dataset

Table 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

Estimation Of Unknowns

Aim of Bayesian Analysis: Estimates of unknows given the observed realisations of the knowns (data set)

- Estimates of unknowns $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$ — intercept, slope
- Using Bayes Theorem:

joint density of beta and y

marginal density of y

prior

posterior probability of the unknowns given the knowns

likelihood

$$\underbrace{f(\beta|y)}_{\text{posterior probability of the unknowns given the knowns}} = \frac{f(\beta, y)}{f(y)} = \frac{f(y|\beta)f(\beta)}{f(y)} \propto f(y|\beta)f(\beta)$$

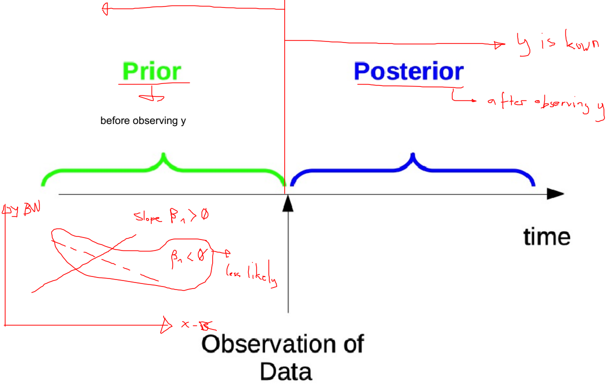
proportional to

where $f(\beta)$: prior distribution and $f(y|\beta)$: likelihood

posterior density of the unknowns given the knows is proportional to the likelihood times the prior

Prior and Posterior

measure of uncertainty related to the unknowns is
quantified by the the prior density of the unknowns.
Our regression f(\beta)



Posterior Distribution

make a quantitative statement of the uncertainty of the unknowns after observing y

- How to get to posterior distribution $f(\beta|y)$
- Use regression as example
- β is a vector with two components, $\beta^T = \begin{bmatrix} \beta_0 & \beta_1 \end{bmatrix}$
- **Solution**: accumulation of samples from full conditional posterior distributions leads to samples from posterior distribution

For the two unknowns (slope and intercept), we get two full conditional distributions:
1. f(\beta_0 | \beta_1, y)
2. f(\beta_1 | \beta_0, y)

random numbers from 1.

random numbers from 2.

pool all random numbers which result in a random sample of the posterior distribution

Prior and Likelihood

the posterior density depends on two components
1. prior
2. likelihood

specify

- What are the distributional assumptions (for regression example and in general)
- Prior: $f(\beta)$ usually assumed to be uniform

no prior information

- Likelihood: $f(y|\beta)$ assumed to be multivariate normal

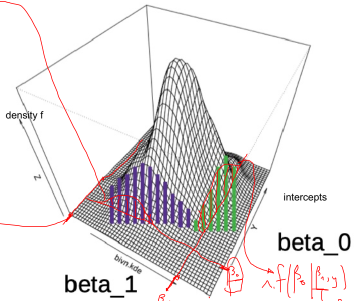
Regression

- Full conditional distributions
  - intercept: $f(\beta_0|\beta_1, y)$ is a normal distribution
  - slope: $f(\beta_1|\beta_0, y)$ is normal distribution
- Draw random numbers from full conditional distributions in turn
- Result will be samples from posterior distribution

Full Conditional Distributions

2. $f(\beta_1 | \beta_0, y)$

β_0: fixed



intercepts

$\beta_1 = \beta_0 = 0$

starting point

β_0

1. $f(\beta_0 | \beta_1, y) \Rightarrow \underbrace{N(\ldots|\ldots)}_{\text{random}}$

β_1: fixed

β_1

slope

Estimates from Samples

- Given Samples from posterior distribution $f(\beta|y)$
- Estimates are computed as empirical means and standard deviation based on the samples

$$\beta_{Bayes} = \frac{1}{N}\sum_{t=1}^{N}\beta^{(t)}$$

with N samples drawn from full conditional distributions

Gibbs Sampler

How to get to random numbers and how are \beta_0 and \beta_1 defined
for the full conditional distribution

- Implementation using full conditional distributions
- Use Gibbs Sampler for regression example
- Step 1: Start with initial values $\beta_0 = \beta_1 = 0$
- Step 2: Compute mean and standard deviation for full conditional distribution of β_0 → $f(\beta_0 | \beta_1, y)$
- Step 3: Draw 1 random sample for β_0 → $\beta_0 = \dots$
- Step 4 and 5: same for β_1 → $f(\beta_1 | \beta_0, y)$ → sample → β_1
- Step 6: Repeat 2-5 N times → 10^5
- Step 7: Compute mean from samples

Starting point: