Applied Statistical Methods - Exercise 3

Peter von Rohr

2023-03-12

Problem 1: Linear Regression on Genomic Information

Use the following dataset which is also given in:

https://charlotte-ngs.github.io/asmss2023/data/asm_flem_genomic_data.csv

to estimate marker effects for the single loci using a linear regression model.

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

Problem 2: Regression On Dummy Variables

Use the dataset with the breeds assigned to every animal and find out the influence of the breed on the response variable `body weight`. The data is available from

```
## [1] "https://charlotte-ngs.github.io/asmss2023/data/asm_bw_flem.csv"
```

Start by fitting a linear model with `Breed` as the only factor in the model, hence ignore the independent variables such as `Breast Circumference`, `BCS` and `HEI`.

Problem 3: Estimable Function

Use the matrix vector-notation to setup the model for a regression on dummy variable with the data on breeds and body weight as used in Problem 2. The aim of this problem is to find the estimable functions used in the output of `lm()`.

The model is given by

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

Setup the least squares normal equations. Find a solution for $\mathbf{b}^0$ and construct the estimable function that is used in the output `lm()`.