Applied Statistical Methods - Exercise 3

Peter von Rohr

2021-03-15

Problem 1: Fixed Linear Effects Model

We want to analyse a dataset with genetic information using a fixed linear effects model. The dataset is taken from the course notes and is shown in Table 1.

We assume that the SNP loci have a purely additive effect on the trait. That means for a SNP locus L the absolute value of the genotypic value of the homozygous genotypes (L_1L_1 and L_2L_2) is taken to be a_L and the genotypic value of the heterozygous genotype (L_1L_2) is taken to be 0. The fixed linear effects model contains the observation in Table 1 as the response variable, an intercept and the genotypic values of the the genotypes at the two SNP Loci G and H as predictor variables.

For the observation y_i of animal i, we can specify the model as

$$y_i = \beta_0 + W_i \cdot a + \epsilon_i$$

where β_0 is the intercept, a is the vector of additive SNP-effects, W_i is a row vector denoting the SNP-Genotypes and ϵ_i is the random error term.

The data can be read from https://charlotte-ngs.github.io/gelasmss2021/data/ex03p01_data.csv. The address from where the data can be downloaded is assigned to a variable.

```
### # specify path to data file depending on online status
s_data_file <- "https://charlotte-ngs.github.io/gelasmss2021/data/ex03p01_data.csv"
```

It can be read using the following statement

```
### # read the data into a tibble
tbl_geno_data <- readr::read_csv(file = s_data_file)
```

Table 1: Genotypic Data Used for Fitting a Fixed Linear Effect Model

Animal	SNP G	SNP H	Observation
1	1	0	510
2	0	1	528
3	0	1	505
4	1	-1	539
5	1	1	530
6	0	0	489
7	0	-1	486
8	-1	1	485
9	0	-1	478
10	-1	0	479
11	1	0	520

12	1	1	521
13	-1	0	473
14	-1	0	457
15	0	1	497
16	0	0	516
17	1	0	524
18	1	0	502
19	1	-1	508
20	0	0	506

Your Tasks

- Specify the fixed linear effects model in matrix-vector notation by putting the information from the dataset into the model. Use the same parametrization as shown in the course notes where the intercept β_0 and the vector a are combined into a single parameter vector b. The design matrix that links elements in b to observations y is then called X.
- Use the function `Matrix::rankMatrix()` from the `Matrix` package on the matrix X to find out the rank of the design matrix.
- Depending on the rank of X compute an estimate for b, if the rank of the matrix is equal to the number of columns of matrix X, then the same forumla as was used in the regression model can be used
- Verify your results using the `lm()` function

Hints

- Read the data using the function `readr::read_csv()`

Problem 2: Genomic Relationship Matrix

From the given dataset that can be obtained from

https://charlotte-ngs.github.io/gelasmss2021/data/ex03p02_data.csv,

compute the genomic relationship matrix G. The dataset is organised such that animals are in rows and SNPs are in columns.

Hints

- Read the data using the function `readr::read_csv()`
- Convert the input data with the function `as.matrix()` to a matrix
- Use the function `apply(mat, 2, mean)` to compute the columnwise mean of matrix `mat`
- Use the `matrix` function to construct the matrix P from the vector of SNP allele frequencies

Additional Problem

Write a function in R that accepts a matrix of genotypes and that computes the genomic relationship matrix. Verify your results from Problem 2.

Hints

- A function in R can be declared using the keyword `function`
- A function in R consists of three parts
  1. the name of the function
  2. the function arguments and
  3. the body of the function.

The following figure shows the general structure of an R-function