Livestock Breeding and Genomics - Solution 10

Peter von Rohr

2022-12-09

Problem 1: Marker Effect Model

We are given the dataset that is shown in the table below. This dataset contains gentyping results of 10 for 2 SNP loci.

Animal	SNP A	SNP B	Observation
1	0	0	156
2	1	0	168
3	0	1	161
4	1	0	164
5	-1	0	128
6	-1	1	124
7	0	-1	143
8	1	1	178
9	1	0	163
10	0	0	151

The above data can be read from:

```
## https://charlotte-ngs.github.io/lbgfs2022/data/geno_data.csv
```

Your Task

- The goal of this problem is to estimate SNP marker effects using a `marker effect model`. Because we have just 2 SNP loci, you can use a fixed effects linear model with the 2 loci as fixed effects. Furthermore you can also include a fixed intercept into the model.
- Specify all the model components including the vector of observations, the design matrix X, the vector of unknowns and the vector of residuals.
- You can use the R-function `lm()` to get the solutions for estimates of the unknown SNP effects.

Solution

The fixed effects model to estimate the marker effects can be written as

$$y = X\beta + e$$

where y is the vector of observations, β is the vector of fixed effects and e is the vector of residuals. Inserting the data from the dataset into the model components leads to

$$y = \begin{bmatrix} 156 \\ 168 \\ 161 \\ 164 \\ 128 \\ 124 \\ 143 \\ 178 \\ 163 \\ 151 \end{bmatrix} \beta = \begin{bmatrix} \beta_0 \\ \beta_A \\ \beta_B \end{bmatrix} e = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \\ e_9 \\ e_{10} \end{bmatrix}$$

where β_0 is the intercept and β_A and β_B correspond to the marker effects (a-values) for both SNPs A and B.

The design matrix X is taken from the dataset as

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & -1 & 0 \\ 1 & -1 & 1 \\ 1 & 0 & -1 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

The solution for the intercept and the marker effects are obtained with

```
lm_snp_eff <- lm(tbl_all_data$Observation ~ tbl_all_data$`SNP A` + tbl_all_data$`SNP B`,
                 data = tbl_all_data)
summary(lm_snp_eff)
```

```
##
## Call:
## lm(formula = tbl_all_data$Observation ~ tbl_all_data$`SNP A` +
##     tbl_all_data$`SNP B`, data = tbl_all_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -9.40  -4.02   0.52   3.02   7.72
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            148.280      2.172  68.270  3.8e-11 ***
## tbl_all_data$`SNP A`   20.740      2.660   7.797 0.000107 ***
## tbl_all_data$`SNP B`    5.860      3.318   1.766 0.120691
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.27 on 7 degrees of freedom
## Multiple R-squared:  0.8985, Adjusted R-squared:  0.8695
## F-statistic: 30.97 on 2 and 7 DF,  p-value: 0.0003335
```

Problem 2: Breeding Value Model

Use the same data as in Problem 1 to estimate genomic breeding values using a `breeding value model`.

Hints

- The only fixed effect in this model is the mean μ which is the same for all observations.
- You can use the following function to compute the genomic relationship matrix

```r
#' Compute genomic relationship matrix based on data matrix
computeMatGrm <- function(pmatData) {
  matData <- pmatData
  # check the coding, if matData is -1, 0, 1 coded, then add 1 to get to 0, 1, 2 coding
  if (min(matData) < 0) matData <- matData + 1
  # Allele frequencies, column vector of P and sum of frequency products
  freq <- apply(matData, 2, mean) / 2
  P <- 2 * (freq - 0.5)
  sumpq <- sum(freq*(1-freq))
  # Changing the coding from (0,1,2) to (-1,0,1) and subtract matrix P
  Z <- matData - 1 - matrix(P, nrow = nrow(matData),
                            ncol = ncol(matData),
                            byrow = TRUE)
  # Z%*%Zt is replaced by tcrossprod(Z)
  return(tcrossprod(Z)/(2*sumpq))
}
matG <-computeMatGrm(pmatData = t(mat_geno_snp))
matG_star <- matG + 0.01 * diag(nrow = nrow(matG))
n_min_eig_matG_start <- min(eigen(matG_star, only.values = TRUE)$values)
if (n_min_eig_matG_start < sqrt(.Machine$double.eps))
  stop(" *** Genomic relationship matrix singular with smallest eigenvalue: ",
       n_min_eig_matG_start)
```

- The resulting genomic relationship matrix is given by

$$G = \begin{bmatrix}
0.093 & -0.125 & -0.125 & -0.125 & 0.292 & 0.083 & 0.292 & -0.333 & -0.125 & 0.083 \\
-0.125 & 0.718 & -0.333 & 0.708 & -0.958 & -1.167 & 0.083 & 0.5 & 0.708 & -0.125 \\
-0.125 & -0.333 & 0.718 & -0.333 & 0.083 & 0.917 & -0.958 & 0.5 & -0.333 & -0.125 \\
-0.125 & 0.708 & -0.333 & 0.718 & -0.958 & -1.167 & 0.083 & 0.5 & 0.708 & -0.125 \\
0.292 & -0.958 & 0.083 & -0.958 & 1.552 & 1.333 & 0.5 & -1.167 & -0.958 & 0.292 \\
0.083 & -1.167 & 0.917 & -1.167 & 1.333 & 2.177 & -0.75 & -0.333 & -1.167 & 0.083 \\
0.292 & 0.083 & -0.958 & 0.083 & 0.5 & -0.75 & 1.552 & -1.167 & 0.083 & 0.292 \\
-0.333 & 0.5 & 0.5 & 0.5 & -1.167 & -0.333 & -1.167 & 1.343 & 0.5 & -0.333 \\
-0.125 & 0.708 & -0.333 & 0.708 & -0.958 & -1.167 & 0.083 & 0.5 & 0.718 & -0.125 \\
0.083 & -0.125 & -0.125 & -0.125 & 0.292 & 0.083 & 0.292 & -0.333 & -0.125 & 0.093
\end{bmatrix}$$

Your Tasks

- Specify all model components of the linear mixed model, including the expected values and the variance-covariance matrix of the random effects.

Solution

The breeding value model is a linear mixed effects model which can be written as

$$y = X\beta + Wu + e$$

where

- y is the vector of observations
- β is the vector of fixed effects
- u is the vector of random genomic breeding values
- e is the vector of random residuals
- X and W are design matrices linking fixed effects and genomic breeding values to observations.

Inserting the information from the dataset into the model leads to

$$y = \begin{bmatrix} 156 \\ 168 \\ 161 \\ 164 \\ 128 \\ 124 \\ 143 \\ 178 \\ 163 \\ 151 \end{bmatrix} \beta = \begin{bmatrix} \mu \end{bmatrix} u = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \\ u_7 \\ u_8 \\ u_9 \\ u_{10} \end{bmatrix} e = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \\ e_9 \\ e_{10} \end{bmatrix}$$

The design matrices X and W correspond to

$$X = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} W = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

The expected values of the random effects are

$$E(u) = 0$$
$$E(e) = 0$$
$$E(y) = X\beta$$

The variance-covariance matrices of the random effects are

$$var(u) = G * \sigma_u^2$$

where G is the genomic relationship matrix and σ_u^2 the genetic additive variance explained by the SNPs

$$var(e) = I * \sigma_e^2 = R$$

where I is the identity matrix and σ_e^2 the residual variance.

$$var(y) = WGW^T * \sigma_u^2 + R$$

The solutions for the fixed effects are obtained from mixed model equations.

$$\begin{bmatrix} X^TX & X^TW \\ W^TX & W^TW + G^{-1} * \lambda \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X^Ty \\ W^Ty \end{bmatrix}$$

```
lambda <- 3
```

The parameter $\lambda = \sigma_e^2/\sigma_u^2$ is the ratio between residual variance and genetic variance. We assume that this value corresponds to $\lambda = 3$.

The single components of the mixed model equations are

```
mat_xtx <- crossprod(mat_x_bv)
mat_xtw <- crossprod(mat_x_bv, mat_w_bv)
mat_wtx <- t(mat_xtw)
mat_wtw_ginv_lam <- crossprod(mat_w_bv) + solve(matG_star) * lambda
mat_coeff <- rbind(cbind(mat_xtx, mat_xtw), cbind(mat_wtx, mat_wtw_ginv_lam))
mat_rhs <- rbind(crossprod(mat_x_bv, mat_obs_y), crossprod(mat_w_bv, mat_obs_y))
mat_sol <- solve(mat_coeff, mat_rhs)
```

$$X^TX = \begin{bmatrix} 10 \end{bmatrix}, X^TW = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}, W^TX = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

$$W^TW + G^{-1} = \begin{bmatrix} 295.015 & 5.991 & 11.964 & 5.991 & -17.961 & -0.011 & -23.935 & 23.94 & 5.991 & -5.985 \\ 5.991 & 265.061 & 17.967 & -35.939 & 47.921 & 59.897 & -5.985 & -23.963 & -35.939 & 5.991 \\ 11.964 & 17.967 & 247.14 & 17.967 & 5.962 & -59.862 & 77.789 & -47.858 & 17.967 & 11.964 \\ 5.991 & -35.939 & 17.967 & 265.061 & 47.921 & 59.897 & -5.985 & -23.963 & -35.939 & 5.991 \\ -17.961 & 47.921 & 5.962 & 47.921 & 217.158 & -59.919 & -41.884 & 71.844 & 47.921 & -17.961 \\ -0.011 & 59.897 & -59.862 & 59.897 & -59.919 & 181.23 & 59.839 & 0.046 & 59.897 & -0.011 \\ -23.935 & -5.985 & 77.789 & -5.985 & -41.884 & 59.839 & 175.342 & 95.738 & -5.985 & -23.935 \\ 23.94 & -23.963 & -47.858 & -23.963 & 71.844 & 0.046 & 95.738 & 205.239 & -23.963 & 23.94 \\ 5.991 & -35.939 & 17.967 & -35.939 & 47.921 & 59.897 & -5.985 & -23.963 & 265.061 & 5.991 \\ -5.985 & 5.991 & 11.964 & 5.991 & -17.961 & -0.011 & -23.935 & 23.94 & 5.991 & 295.015 \end{bmatrix}$$

with

$$rhs = \begin{bmatrix} X^Ty \\ W^Ty \end{bmatrix}$$

$$rhs = \begin{bmatrix} 1536.388 \\ 156.41 \\ 168.379 \\ 161.35 \\ 163.533 \\ 127.857 \\ 124.478 \\ 142.925 \\ 177.661 \\ 162.853 \\ 150.941 \end{bmatrix}$$

The solution vector for the estimate of the fixed effect μ and the genomic breeding values for all animals are given by

$$sol = \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix}$$

$$sol = \begin{bmatrix} 153.6388 \\ -3.2397 \\ 10.1794 \\ -0.3581 \\ 10.1633 \\ -16.7139 \\ -13.8599 \\ -6.1497 \\ 13.0754 \\ 10.161 \\ -3.2579 \end{bmatrix}$$