Practical 4: Estimation of Breeding Values

Time schedule of practical session 4:

11:15	Question to lectures, multiple-choice question and follow up on previous multiple choice questions
11:30	Today´s exercise and assignment to groups
12:00	15 minutes break
12:30	Go through excercises using final word
12:50	Repeat multiple-choice questions
13:00	End of practical session 4

Introduction:

In this practical we will estimate breeding values for quantitative traits in the mouse population. We will be using the BLUP method. This method allow for estimation of breeding values using phenotypic information for individuals from a general pedigree. BLUP is based on linear mixed model methodology and estimates of breeding values can be obtained by solving the mixed model equations. The BLUP method also require a genetic relationship matrix and estimates of variance components (e.g., σ^2_a and σ^2_e). Furthermore, we will compute reliabilities to determine how well we have estimated the breeding value in relation to the true breeding value. These methods and algorithms are implemented in the R package `qgg` introduced previously.

Load R packages that will be used in this practical

```
library(qgg) # R package used for REML analysis
```

Explore mouse pedigree data

The mouse data and pedigre set can be loaded using the following commands:

```
mouse <- readRDS(url("https://github.com/psoerensen/bgcourse/raw/main/data/mouseqtl.rds"))
pedigree <- readRDS(url("https://github.com/psoerensen/bgcourse/raw/main/data/pedigree.rds"))
```

First let us have a quick look at the mouse data again. Use the `str` function to get a fast overview of the pedigree you are working.

```
str(pedigree)
```

```
## 'data.frame':    1267 obs. of  6 variables:
##  $ id        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ sire      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ dam       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ family    : Factor w/ 68 levels "0/0","1/2","11/12",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ sex       : chr  "Male" "Female" "Male" "Female" ...
##  $ generation: chr  "M6" "IC" "IC" "M6" ...
```

The number of individuals and generations in the pedigree can be found using the following commands:

```
nrow(pedigree)
```

```
## [1] 1267
```

```
dim(pedigree)
```

```
## [1] 1267    6
```

```
table(pedigree$generation)
```

```
##
##   F1   F2   IC   M6
##   66 1177   12   12
```

Computing genetic relationship matrix for the mouse pedigree:

The genetic relationship matrix A is used for estimating breeding values. The matrix A can be computed using the recursive algorithm implemented in the function `grm` from the qgg package. Use the command below to compute the genetic relationship matrix for the mouse pedigree:

```
A <- grm(pedigree=pedigree)
```

The dimension of the genetic relationship matrix can be determined using the following command:

```
dim(A)
```

```
## [1] 1267 1267
```

The number of rows and columns should be equal to the number of individuals in the pedigree.

Specifying the linear mixed model for the mouse data:

The next step is to prepare the linear mixed model for the mouse data. Recall that the linear mixed model contains the observation vector for the trait(s) of interest (y), the `fixed effects` that explain systematic differences in y, and the `random genetic effects` a and random residual effects e.

A matrix formulation of a general model equation is:

$$y = Xb + a + e$$

where

y : is the vector of observed values of the trait,
b : is a vector of fixed effects,
a : is a vector of random genetic effects,
e : is a vector of random residual effects,
X : is a known design matrix that relates the elements of b to their corresponding element in y.

In the statistical model (specified above) the random effects (a and e) and the phenotypes (y) are considered to be random variables which follow a multivariate normal distribution. In general terms the expectations of these random variables are:

$$\begin{aligned} a &\sim MVN(0, A\sigma_a^2) \\ e &\sim MVN(0, I\sigma_e^2) \\ y &\sim MVN(Xb, V) \end{aligned} \tag{1}$$

where $A\sigma_a^2$, and $I\sigma_e^2$ are square matrices of genetic and residual (co)variances among the individuals, respectively, and $V = A\sigma_a^2 + I\sigma_e^2$ is the overall phenotypic covariance matrix. In the previous section we have allready constructed the genetic relationship matrix A.

In order to specify the linear mixed model we need to construct y and X from the mouse data. Let us just have a quick look at the mouse data again:

Here we will estimate breeding values for body weight. The vector of observed trait values for body weight can be extracted from the mouse data as follows:

```
y <- mouse[,"BW"]
```

Let us explore the trait values using the `head`, `tail` and `summary` functions:

```
head(y)
```

```
## [1] 36.65 33.29 42.07 37.15 38.39 39.82
```

```
tail(y)
```

```
## [1] 39.67 39.35 44.80 52.23 47.63 54.10
```

```
summary(y)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   23.04   34.06   38.32   38.72   43.40   60.28
```

To make the X matrix we need to decide which variables we should include as fixed effects in the model. Here we use the variables sex and reps. The `model.matrix` function can be used to construct the X matrix in the linear mixed model specified above:

```
X <- model.matrix(BW ~ sex + reps, data=mouse)
```

We can use the `head` and `tail` functions to look at the X matrix:

```
head(X)
```

```
##    (Intercept) sexMale reps2 reps3
## 91           1       0     0     0
## 92           1       0     0     0
## 93           1       0     0     0
## 94           1       0     0     0
## 95           1       1     0     0
## 96           1       1     0     0
```

```
tail(X)
```

```
##      (Intercept) sexMale reps2 reps3
## 1262           1       0     0     1
## 1263           1       0     0     1
## 1264           1       1     0     1
## 1265           1       1     0     1
## 1266           1       1     0     1
## 1267           1       1     0     1
```

Question 1: Why do we not include the effect of sire and dam in the model?

Answer:

Estimating genetic parameters on the mouse data using REML:

The BLUP analysis is based on estimates of the variance components (i.e. σ_a^2 and σ_e^2). The variance components are estimated using REML method. The input required the vector of observed values of the trait (y), the design matrix for the fixed effects (X), and the genetic relationship matrix (A).

The genetic relationship matrix A include relationships for all individuals in the pedigree. However only a subset of the individuals have phenotypes recorded for body weight and glucose levels in blood. Therefore we need to subset the A matrix:

```
ids <- rownames(X)
A <- A[ids,ids]
```

The REML analysis is done using the following command:

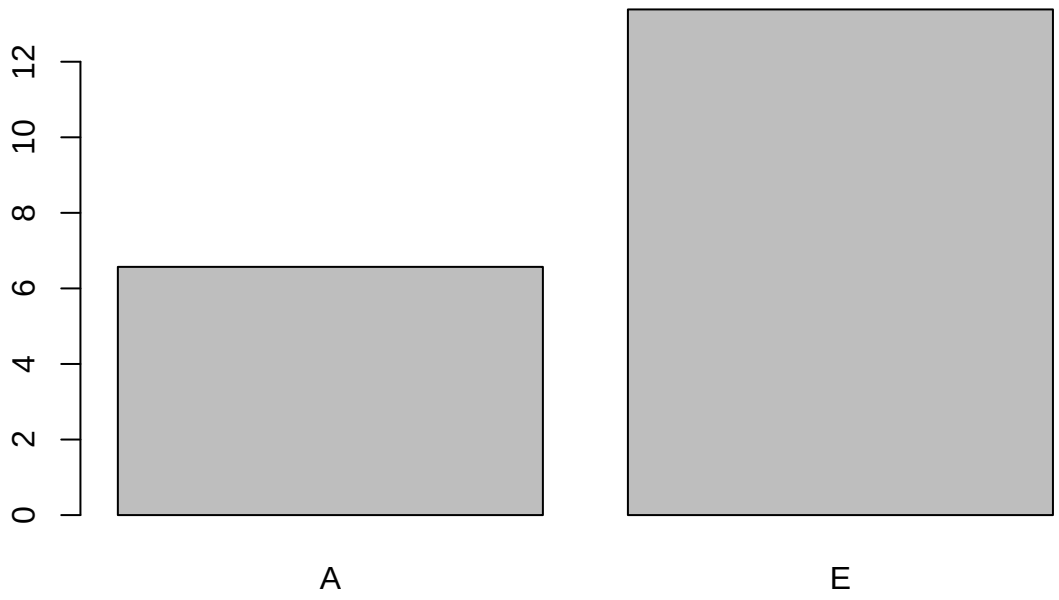
```
fit <- greml(y=y,X=X, GRM=list(A=A))
```

The fit object contains estimates of variance components, fixed and random effects, first and second derivatives of log-likelihood, and the asymptotic standard deviation of parameter estimates. Our main interest is the variance components σ_a^2 and σ_e^2 which are in the `fit$theta` slot of the fit. The following commands extract and makes a barplot of the estimates of the variance components:

```
fit$theta
```

```
##         A         E
##  6.569611 13.384147
```

```
Va <- fit$theta[1] # First element in theta is the additive genetic variance
Ve <- fit$theta[2] # Second element in theta is the residual variance
barplot(fit$theta)
```



Estimating breeding values for traits in the mouse data using BLUP:

The goal of the BLUP analysis is the estimate the fixed ,b, and random genetic effects, a, in the linear mixed model specified above. This can be done using the `BLUE` and 'BLUP´ equations shown below:

The best linear unbiased prediction (BLUP) of $\hat{a}$ is:

$$\hat{a} = A\sigma_a^2 V^{-1}(y - X\hat{b}) \tag{2}$$

The best linear unbiased estimator (BLUE) of $\hat{b}$ is:

$$\hat{b} = (X'V^{-1}X)^{-1}X'V^{-1}y \tag{3}$$

The matrix $(X'V^{-1}X)^{-1}$ denotes the inverse of the matrix $(X'V^{-1}X)$.

We have allready determined y and X and therefore just need to construct the phenotypic covariance matrix V (and it's inverse). This can be done using the following lines of R code:

```
n <- nrow(X)        # Number of individuals in the data set
I <- diag(1,n)      # Identity matrix for residual effects
V <-  A*Va + I*Ve   # Phenotypic variance covariance matrix
Vi <- solve(V)      # Inverse of phenotypic covariance matrix
```

The solution to the fixed effects, b, can be found using the following R command:

```
bhat <- solve(t(X) %*% Vi %*% X)%*%t(X) %*% Vi %*% y
bhat
```

```
##                   [,1]
## (Intercept) 33.8873546
## sexMale      8.3453194
## reps2       -0.3684327
## reps3        2.6411388
```

The solution to the random genetic effects, a, can be found using the following R command:

```
ahat <- (A*Va)%*% Vi %*% (y-X%*%bhat)
head(ahat)
```

```
##            [,1]
## 91 -0.001564943
## 92 -0.663690910
## 93  1.066507303
## 94  0.096965707
## 95 -1.303217779
## 96 -1.021420120
```

```
tail(ahat)
```

```
##          [,1]
## 1262 1.111041
## 1263 1.047981
## 1264 0.477426
## 1265 1.941591
## 1266 1.035109
## 1267 2.310096
```

Question 2: Make histogram for y and the estimated breeding values. What do you think about their distribution?

Answer:

Question 3: Make a scatter plot of y and the estimated breeding values. What do you think about their relationship?

Answer:

Question 4: Which of the sires has the highest breeding value for body weight? Which dam has the highest breeding value for body weight?

Answer:

Computing reliabilities for the estimated breeding values for traits in the mouse data:

The last step is to compute reliabilities to determine how well we have estimated the breeding value in relation to the true breeding value. The reliability (i.e., variances of prediction error are often expressed as a number going from 0 to 1. The general formula is:

REL = (Var(TBV) - Var(TBV-EBV))/(Var(TBV)

TBV=True Breeding Values EBV=Estimated Breeding Values

The standard error of prediction, or SEP, is the square root of the variance of prediction error.

Estimation of breeding values and reliabilities can also be done by solving the mixed model equation. Procedures for solving the mixed model equations are implemented in the `gsolve` function from the R package `qgg` introduced previously. The input to this function is y, X, A or G and estimates of the variance components (e.g., σ^2_a and σ^2_e).

```
fit <- gsolve(y=y,X=X, GRM=list(A=A), Ve=Ve, Va=Va)
```

We can use the `str`, `head` and `tail` functions to look at the output from the `gsolve` function:

```
str(fit)
```

```
## 'data.frame':    1267 obs. of  3 variables:
##  $ rel: num  0.734 0.734 0.736 0.736 0.734 ...
##  $ pev: num  0.0462 0.0462 0.0459 0.0459 0.0462 ...
##  $ sep: num  0.215 0.215 0.214 0.214 0.215 ...
```

```
head(fit)
```

```
##         rel        pev       sep
## 1 0.7335770 0.04624977 0.2150576
## 2 0.7335770 0.04624977 0.2150576
## 3 0.7356780 0.04588505 0.2142080
## 4 0.7356780 0.04588505 0.2142080
## 5 0.7337962 0.04621171 0.2149691
## 6 0.7337962 0.04621171 0.2149691
```

```
tail(fit)
```

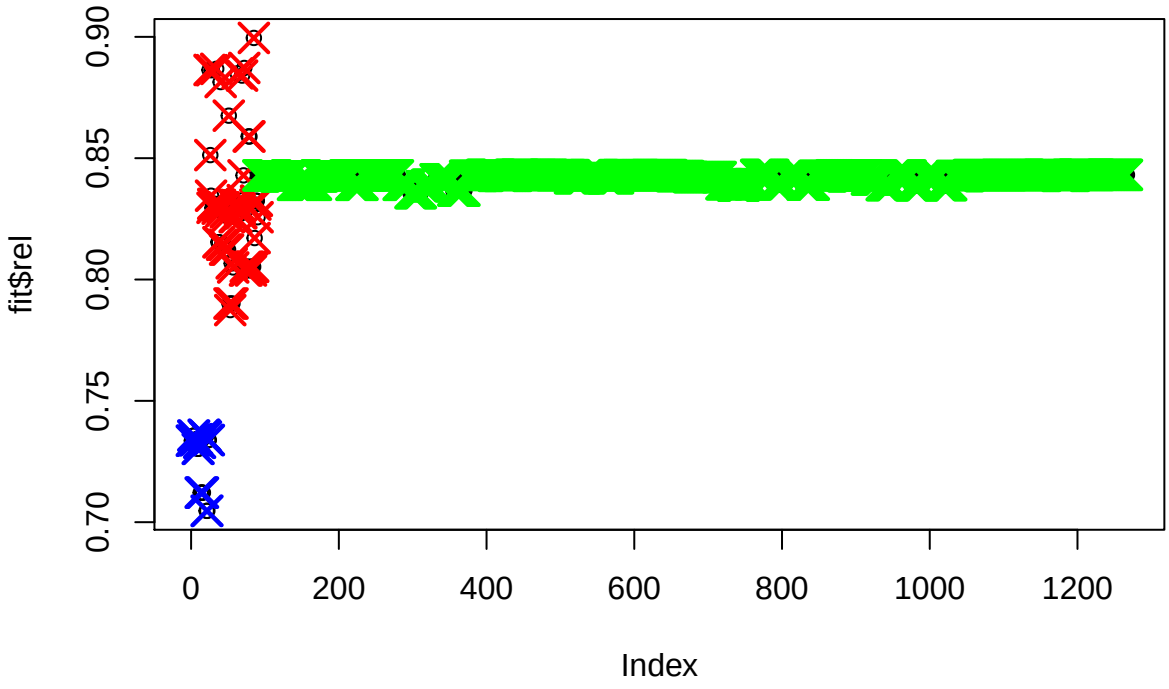
```
##            rel        pev       sep
## 1262 0.8431152 0.02723445 0.1650286
## 1263 0.8431152 0.02723445 0.1650286
## 1264 0.8431152 0.02723445 0.1650286
## 1265 0.8431152 0.02723445 0.1650286
## 1266 0.8431152 0.02723445 0.1650286
## 1267 0.8431152 0.02723445 0.1650286
```

Question 5: Which mouse has the hight reliability and what could be the reason for this?

Answer:

To further explore the reliabilities we can make a plot of them using the following command:

```
plot(fit$rel)
F0 <- pedigree$generation=="IC" | pedigree$generation=="M6"
F1 <- pedigree$generation=="F1"
F2 <- pedigree$generation=="F2"
n <- length(fit$rel)
points(x=(1:n)[F0],y=fit$rel[F0],col="blue", pch=4, cex=2, lwd=2 )
points(x=(1:n)[F1],y=fit$rel[F1],col="red", pch=4, cex=2, lwd=2 )
points(x=(1:n)[F2],y=fit$rel[F2],col="green", pch=4, cex=2, lwd=2 )
```



The reliabilities for generation F0 is blue, F1 is red and F2 is green. What we can observed is that mice from the F0 generation (i.e., IC and M6) has the lowest reliability.

Question 6: Could you explain this?

Answer:

To further explore the value of using phenotypic information from different relatives consider the general formula for reliability of estimated breeding value using different sources of information: {-}

$$r^2_{a,\hat{a}} = \frac{(a')^2 n h^2}{1 + (n-1)r} \tag{4}$$

where a' is the genetic relationship between the breeding individual and individuals with phenotypes, n is the number of phenotypic records, h^2 is the trait heritability, and r is correlation between individuals with observations ($r = a'' h2 + c2$, where a'' = genetic relationship between individuals with records and target, $c2$ = common environmental component).

We want to compare the reliabity of the estimated breeding value for an individual computed based on phenotypic observation on different types of relatives. Assume that the trait narrow sense trait heritability $h^2 = 0.35$ and that the common environmental component $c2 = 0$.

What is the reliability if we compute the breeding values based on:

1) Own
2) Mother
3) 50 paternal halfsibs (same father)
4) 20 offspring that halfsibs (different mothers)

Question 8: Which phenotypic information source give the highest reliability?

Answer:

Question 8: Which of the sires has the highest breeding value for blood glucose levels?

Answer: