Part E

Genomic selection with BLUP

1 Genomic selection using SNP-BLUP

In this practical you will perform genomic selection in a small data set using SNP-BLUP. The data set consists of a reference population of 325 bulls with daughter yield deviations (DYDs) for protein %. This phenotype is an accurate predictor of genotype, eg the heritability is close to one for the trait DYD. The bulls have been genotyped for 10 SNPs.

Then there are a set of 31 calves who are selection candidates for this years progeny test team. They are genotyped for the same 10 markers. Your task is to predict GEBV for these 31 selection candidates. To do this we will need to predict the effects of the 10 SNPs in the reference population, using the equations:

$$\begin{bmatrix} \mathbf{1}_n'\mathbf{1}_n & \mathbf{1}_n'\boldsymbol{X} \\ \boldsymbol{X}'\mathbf{1}_n & \boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I}\lambda \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\boldsymbol{\alpha}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_n'\boldsymbol{y} \\ \boldsymbol{X}'\boldsymbol{y} \end{bmatrix}$$

Where $\boldsymbol{\alpha}$ are the SNP effects, 1_n is a vector of ones (325×1), $\boldsymbol{X}$ is a design matrix allocating SNP genotype to records, μ is the overall mean. We will use R to solve these equations. The $\boldsymbol{X}$ matrix has already been built for you, and is contained in the file *xvec_day4.inp*. The $\boldsymbol{y}$ vector is contained in the file *yvec_day4.inp*.

What you need to do is write a small R script to solve the equations. This can be done by starting the script in notepad, then opening it in the R console.

The first lines should declare the parameters number of markers and number of records. A this point we will also specify the value of lambda as 10.

```
> nmarkers <- 10      #number of markers
> nrecords <- 325     #number of records
> lambda    <- 10      #value for lambda
```

Next we will read in the files. Change the path to the location where you have stored the files. Note that these statements should all be on one line. Have a look at these files before opening them.

```
> X <- matrix(scan("Data/PartE/xvec_day4.inp"), ncol = nmarkers, byrow = TRUE)
> y <- matrix(scan("Data/PartE/yvec_day4.inp"), byrow = TRUE)
```

So now we have the matrix $\boldsymbol{X}$, the vector $\boldsymbol{y}$. We still need a vector of ones and a identity matrix dimension number of markers $\times$ number of markers...

```
> ones <- array(1, c(nrecords))
> ident_mat <-diag(nmarkers)
```

The next step is to build the coefficient matrix. This can be done in blocks, e.g.

```
> coeff <- array(0, c(nmarkers + 1, nmarkers + 1))
> coeff[1:1, 1:1] <- t(ones) %*% ones
> coeff[1:1, 2:(nmarkers+1)] <- t(ones) %*% X
```

Question. You will need to build the other blocks. You will also need to build the right hand side of the equation.

```
> coeff[2: 2:(nmarkers+1), 1] <- t(X) %*% ones
> coeff[2:(nmarkers+1), 2:(nmarkers+1)] <- t(X) %*% X + lambda * ident_mat
> rhs = rbind(t(ones) %*% y, t(X) %*% y)
```

The solutions can be obtained easily by using the inbuilt function solve,

```
> solution_vec <- solve(coeff, rhs)
```

Question. Print out this vector of solutions (eg print(solution_vec)). What is the solution for the mean? Which SNP has the largest effect?

Next we want to print GEBV for the selection candidates. This is done with the equation:

$$\mathbf{GEBV} = X\hat{\boldsymbol{\alpha}}$$

Question.

- The $\hat{\boldsymbol{\alpha}}$ are the solutions for the SNP effects you have just solved. The xvector for the selection candidates is in the file *xvec_prog.inp*. Can you write a small R script to calculate the GEBV?
- Fours years later, all the selection candidates receive a phenotypic record from a progeny test. The results are in the file *yvec_prog.inp*. What is the correlation between your GEBV and the progeny test result (which is close ot the true breeding value)? (Don't expect this to be too high with only 10 SNPs).

```
> Xprog <- matrix(scan("/Data/PartE/xvec_prog.inp"), ncol = nmarkers, byrow = TRUE)
> GEBV = Xprog %*% solution_vec[-1]
> yprog <- matrix(scan("Data/PartE/yvec_prog.inp"), ncol = 1, byrow = TRUE)
> cor(GEBV, yprog)
```

2 Genomic selection using GBLUP

In this second exercise, we will analyse the same data as previously, but using the GBLUP (genomic relationship matrix) approach. The mixed model is

$$\boldsymbol{y} = \mathbf{1}_n\boldsymbol{\mu} + \boldsymbol{Zg} + \boldsymbol{e}$$

Where terms are as above, and $\boldsymbol{Z}$ is a design matrix allocating records to individuals, and $\boldsymbol{g}$ is a vector of (genomic) breeding values. The $\boldsymbol{g}$ are random effects, assumed to be distributed $\mathcal{N}(0, \boldsymbol{G}\sigma^2_g)$, where Θ is the genomic relationship matrix.

The solutions to the mixed model equations are

$$\begin{bmatrix} \hat{\boldsymbol{\mu}} \\ \hat{\boldsymbol{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_n'\mathbf{1}_n & \mathbf{1}_n'\boldsymbol{Z} \\ \boldsymbol{Z}'\mathbf{1}_n & \boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{G}^{-1}\frac{\sigma_e^2}{\sigma_g^2} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1}_n'\boldsymbol{y} \\ \boldsymbol{Z}'\boldsymbol{y} \end{bmatrix}$$

Question. The first step in building the mixed model equations is to build $\boldsymbol{G}$.

$\boldsymbol{G}$ is constructed as

$$\boldsymbol{G} = \boldsymbol{W}\boldsymbol{W}'/2\sum_{j=1}^{m} p_j(1-p_j),$$

where m is the number of markers, p_j is the allele frequency of the 2nd allele for SNP j, and $w_{ij} = x_{ij} - 2p_j$. Remember in the GBLUP, $\boldsymbol{G}$ (and $\boldsymbol{Z}$) have to include all individuals, including the individuals with no phenotypes of their own that we wish to predict. So the $\boldsymbol{X}$ matrix has to include all individuals, including the progeny.

You can create this new $\boldsymbol{X}$ from the original X matrices in the first practical using the rbind (join by row command in R):

```
> Xall <- rbind(X, Xprog)
> nanims = nrow(Xall)
```

We next construct $\sum_{j=1}^{m} p_j(1-p_j)$

```
> # p_j for all SNPS
> p = colMeans(Xall)/2
> # sump
> sump <- 0
> for(j in 1:nmarkers)
+     sump <- sump + 2*p[j]*(1-p[j])
```

The sump is the sum of the heterozygosities,

$$2\sum_{j=1}^{m} p_j(1-p_j)$$

and now we calculate $\boldsymbol{W}$

```
> W = matrix(0, nrow = nanims, ncol = nmarkers)
> for(i in 1:nanims){
+   for(j in 1:nmarkers){
+         W[i,j] = Xall[i,j] - 2*p[j]
+   }
+ }
>
> #note: both code can be merged to limit the number of loops (on j)
```

Then $\boldsymbol{G}$, and its inverse can be constructed as

```
> G = W%*%t(W)/sump
> # The next line adds a small amount to the diagonal of G,
```

```
> # otherwise G is not invertable in this small example!
> G <- G + diag(nanims)*0.01
> Ginv <- solve(G)
```

The only other matrix we do not have is $\boldsymbol{Z}$, which is has dimensions nrecords $\times$ nanims. $\boldsymbol{Z}$ is a diagonal matrix for those animals with records, and a block of zeros for those animals without records (as there are no records to allocate these animals to).
$\boldsymbol{Z}$ can be constructed as

```
> Z1 <-diag(nrecords)
> Z2 <-matrix(0, 325, 31)
> Z <- cbind(Z1, Z2)
```

Question.

- Now go ahead and build the mixed model solution equations

$$\begin{bmatrix}\hat{\boldsymbol{\mu}}\\ \hat{\boldsymbol{g}}\end{bmatrix} = \begin{bmatrix}\mathbf{1}_n'\mathbf{1}_n & \mathbf{1}_n'\boldsymbol{Z}\\ \boldsymbol{Z}'\mathbf{1}_n & \boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{G}^{-1}\frac{\sigma_e^2}{\sigma_g^2}\end{bmatrix}\begin{bmatrix}\mathbf{1}_n'\boldsymbol{y}\\ \boldsymbol{Z}'\boldsymbol{y}\end{bmatrix}$$

  For $\frac{\sigma_e^2}{\sigma_g^2}$ use a value of 1.

- What is the accuracy of the genomic predictions for the 31 selection candidates from the GBLUP, eg $r(\hat{\boldsymbol{g}}, \text{yprog})$? How does this compare with the accuracy of SNP-BLUP?

```
> # coeff
> coeff <- array(0, c(nanims + 1, nanims + 1))
> coeff[1:1, 1:1] <- t(ones) %*% ones
> coeff[1:1, 2:(nanims+1)] <- t(ones) %*% Z
> coeff[2: 2:(nanims+1), 1] <- t(Z) %*% ones
> coeff[2:(nanims+1), 2:(nanims+1)] <- t(Z) %*% Z + Ginv
> rhs = rbind(t(ones) %*% y, t(Z) %*% y)
```

```
> gblup <- solve(coeff, rhs)
```

```
> # the genomic prediction for the 31 selection candidates is
> yprog_pred = gblup[-c(1:326)]
```

```
> # the accuracy is
> cor(yprog, yprog_pred)
```

3 Comparison between SNP-BLUP and GBLUP

Question. If you plot GEBV for the SNP-BLUP prac against $\hat{\boldsymbol{g}}$, do you get a regression line with a slope of 1, indicating these are equivalent models? Why, or why not (hint, did we centre and standardise the X matrix to the w matrix in SNP-BLUP)? If you use the same genotype matrix ($\boldsymbol{W}$) in both SNP-BLUP and GBLUP are the genomic predictions identical?

```
> plot(GEBV,yprog_pred, xlab = "GEBV", ylab = "SNP-BLUP")
> lm(yprog_pred ~ GEBV)
> abline(lm(yprog_pred ~ GEBV))
```

```
> Wprog = W[-c(1:325),] #only W for the progeny
> blup_W = Wprog %*% solution_vec[-1]
> plot(GEBV, blup_W, xlab = "GEBV", ylab = "SNP-BLUP with W")
> lm(blup_W ~ GEBV)
```

```
Writing to file Additional Files For Students/Rcode/PartD - BLUP.R
```