Laboratory 2

Consider the dataset in Table 1, modified from p110 Ben Hayes course notes. We will use this dataset to explore some alternative models for fitting SNP effects. The columns include the allele calls at each marker locus (M1, M2 and M3), followed by the covariate that represent the number of A (A1, A2 and A3) or B (B1, B2 and B3) alleles at each locus.

Animal	phenotype	M1	M2	M3	A1	B1	A2	B2	A3	B3
1	9.68	BB	AB	AA	0	2	1	1	2	0
3	2.29	AB	BB	BB	1	1	0	2	0	2
20	0.81	AA	AB	AB	2	0	1	1	1	1
4	3.42	AA	AB	AA	2	0	1	1	2	0
2	5.69	BB	BB	BB	0	2	0	2	0	2
5	5.92	AB	AA	AA	1	1	2	0	2	0
6	2.82	AB	AB	BB	1	1	1	1	0	2
7	5.07	BB	AB	BB	0	2	1	1	0	2
8	8.92	BB	BB	AA	0	2	0	2	2	0
9	2.4	AA	BB	AB	2	0	0	2	1	1
10	9.01	BB	BB	AA	0	2	0	2	2	0
11	4.24	AB	AB	AB	1	1	1	1	1	1
12	6.35	BB	AA	AB	0	2	2	0	1	1
13	8.92	BB	AB	AA	0	2	1	1	2	0
14	-0.64	AA	BB	BB	2	0	0	2	0	2
15	5.95	AB	AA	AA	1	1	2	0	2	0
16	6.13	AB	AB	AA	1	1	1	1	2	0
17	6.72	AB	AB	AA	1	1	1	1	2	0
18	4.86	AB	AB	AB	1	1	1	1	1	1
19	6.36	BB	BB	BB	0	2	0	2	0	2
21	9.67	BB	AB	AA	0	2	1	1	2	0
22	7.74	BB	AB	AB	0	2	1	1	1	1
23	1.45	AA	BB	AB	2	0	0	2	1	1
24	1.22	AA	AB	AB	2	0	1	1	1	1
25	-0.52	AA	BB	BB	2	0	0	2	0	2

This data first needs to be read into Julia. The command ;pwd will show the working directory. The datafile needs to be located in the working directory. You could either copy it there, or change the working directory using the ;cd "dirname" command, where dirname is the path to the working directory. The command ;ls will show the files in the working directory.

A simple R script will be provided with the following commands to read the datafile.

```
genomicdata =readdlm("BenHayesp110.txt",header=true)
```

will read the text file into an array object in Julia. The first element of the array is genomicdata[1] and is the table of data, whereas genomicdata[2] is the header. The columns of the table e.g. a1 in column 6, can be accessed as genomicdata[1][:,6]. For example,

```
ytmp = float(genomicdata[1][:,2])
Ztmp = float(genomicdata[1][:,6:11])
```

will read in a potential y vector and Z matrix.

We will be fitting some models where rank is an issue for certain analyses. For example, in least squares models, we need to have at least as many animals as we have effects. This is typically not an issue if the fitted effects are treated as random. However, for equivalent models that fit animal effect using SNP genotypes to form relationships, the genomic relationship matrix will not be full rank unless there are at least as many SNP effects fitted as there are animals. For this reason, in different models we will use different subsets of the complete **y** and **Z** vector. The variable nanim sets the number of animals to be used. The following lines will set up the example to use the first thirteen animals in the datafile.

```
nanim  = 13;
y =ytmp[1:nanim];
X = ones(nanim,1);
Z = Ztmp[1:nanim,:];
neffects = size(Z,2);
nfix     = size(X,2);
nloci    =neffects/2;
istart   =nfix+1;    #these are pointers to assist in extracting subvectors
iend     =nfix+neffects;
```

Example 1: Fitting both alleles at the three loci as random effects using GLS.

The GLS equation(s) for the model we discussed in the lecture are

$$\hat{\mathbf{b}}^{0} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}\right), \text{ for } \mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}.$$

These equations are useful as **V** is typically full rank, but are not practical in many situations where **V** is large. In this example with just the mean fitted as the only fixed effect, the GLS equation will be a scalar form.

In order to form **V**, you will need to know **G** and **R**. Suppose the residuals are homogeneous and uncorrelated. We will use a residual variance of 1. **R** can be formed using the diag command.

```
sigmaSqE=1;
R  =diagm(fill(sigmaSqE,nanim))
```

The incidence matrix **Z** has 6 columns – one for each of the allelic effects. Suppose the three loci have different variance – say 2, 4 and 3, respectively. Create a **G** matrix of order 6 with columns corresponding to the columns in **Z**.

G=diagm([2,2,4,4,3,3])

Form and inspect **V** and compute $\mathbf{V^{-1}}$.

Be sure to save all your steps so you can immediately repeat your calculations with a modified dataset or different parameters.

Estimate the fixed effects by solving the GLS equations. Print out the result(s). The BLUPs of the random effects can then be obtained from selection index principles, but adjusting the phenotypic records with the GLS estimates of the fixed effects (rather then the true values as is required in selection index). That is, solve

$$\hat{\mathbf{a}} = \mathbf{GZ'V^{-1}}\left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}^{0}\right).$$

Note that the estimates of the allelic effects sum to zero, even though no such constraint was actively used. This is a feature of mixed models in certain circumstances.

Calculate the substitution effects by forming a contrast vector (**k**) with order equal to the order of $\hat{\mathbf{a}}$, that contains all zeros except elements 1 and -1 corresponding to the first and second allele at a locus, and then compute the linear function $\mathbf{k'}\hat{\mathbf{a}}$. Record the results. You can use vcat() or hcat() as appropriate to stack matrices or vectors vertically or horizontally. Place the three contrast vectors into a matrix **K** whose first row is the **k** vector given above and the second and third rows are the corresponding vectors for computing substitution effects at the second and third loci respectively. In that case, the matrix-vector product **Ka** will compute all three substitution effects at once.

Example 2: Shrinkage of substitution effects.

Modify the three pairs of diagonal elements of **G**, or equivalently, modify the single diagonal element of the nanim by nanim matrix **R** in order to modify the variance ratio lambda, of residual to genetic variance. In an animal model, lambda is $(1-h^2)/h^2$ which will be 0 if h^2 is 1 and a large number if h^2 is small. For a heritability of 0.25, lambda is 3. In genomic prediction models, the genetic variance is partitioned among all the loci. If there are hundreds of loci, the lambda ratio for each locus will be large. You can simulate this effect by making the diagonal elements or **R** say 10 or 100 times larger than **G**. Compare the estimated substitution effects for varying values of residual variance (in relation to additive variance). Shrinkage is related to the magnitude of the ratio of residual to additive variance. If residual variance is

small this ratio will be reduced and the estimates will approach least squares. Inspect the variance ratio for each scenario you attempt.

If order to compute the least squares estimate you will need to form the least squares equations treating allelic effects as fixed. To do this, you need to form a new incidence matrix for fixed effects that includes the old fixed effects (eg the overall mean) as well as the allelic effects. You can do this using hcat(X,Z) to augment the columns of the two incidence matrices. However, this new matrix will not have full column rank so the least squares equations will not be full rank. You should be able to constrain the new equations to full rank by limiting the augmented matrix to include only one column of allelic effects for each locus.

For example, $\mathbf{X}_{\text{new}}$ <- hcat(X,Z[:,[1, 3, 5]]) will use only those three columns. Then the least squares solutions can be obtained from solving the following full rank equations. The first effect in these equations will be an intercept rather than a mean, unless you center the covariates in the **Z** matrix by subtracting 1.

$$\left[\mathbf{X}_{new}'\mathbf{X}_{new}\right]\left[\hat{\mathbf{b}}^{0}\right] = \left[\mathbf{X}_{new}'\mathbf{y}\right]$$

Modify the constant nanim to alter the number of animals in the datafile that will be used in the calculation. Try larger and smaller values.

What do you conclude about the importance of treating SNP effects as random in terms of shrinkage of estimated effects?

Before continuing, you will want to reset the genetic and residual variances back to their original values.

Example 3: Fitting both alleles at the three loci as random effects using MME.

An alternative approach to estimate random effects is to use the mixed model equations. Rather than requiring the inverse of **V**, the typical form of the mixed model equations requires the inverse of **G** and the inverse of **R**. Its general form is as follows

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{bmatrix}\begin{bmatrix} \hat{\mathbf{b}}^{0} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}.$$

In simple cases where **R** is a scaled identity, only the inverse of **G** is required as the scalar residual variance can be factored out by multiplication. Remember that the inverse of the coefficient matrix will need to be scaled by the residual variance to

compute the correct prediction error variances or reliabilities when you use this modified form. Form and solve these simpler mixed model equations, as follows

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} + \sigma_e^2\mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}}^0 \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix}.$$

You can use the commands hcat () and/or vcat() to join two matrices of conformable order by column or by row respectively or form these directly as a matrix in Julia as we demonstrated in the first exercises.

Compare the solutions for the fixed effects and the six random allelic effects to the GLS solutions. They should be identical. If not, check your equations before you proceed.

Extract the prediction error variance-covariance (PEV) matrix ($\text{var}(\hat{\mathbf{a}} - \mathbf{a}) = \mathbf{C22}\sigma_e^2$) of the fitted allelic effects, where **C22** is that submatrix of the inverse of the mixed model equations corresponding to the rows and columns representing random effects (ie $\mathbf{Z'Z} + \sigma_e^2\mathbf{G}^{-1}$ portion of the inverse). Compute $\text{var}(\hat{\mathbf{a}}) = \mathbf{G} - \mathbf{C22}\sigma_e^2$ by subtracting the PEV matrix from the genetic variance-covariance matrix. The reliability of the predictions (squared correlation between true and predicted merit) are obtained by dividing the diagonal elements of $\mathbf{G-C22}\,\sigma_e^2$ by the diagonal elements of **G**. You might find the Julia function diag() useful for this purpose. Reliability is used in some industries (eg dairy) to convey the information content in estimated breeding values (EBVs).

Compute the substitution effects by forming relevant contrast vectors as in the previous question.

From the viewpoint of genomic prediction rather than QTL detection, we will be more interested in linear functions of the estimated SNP effects, such as $\mathbf{Z}\hat{\mathbf{a}}$. Compute that linear function for all animals. You could also compute the correlation with phenotype using the cor() function. You may want to plot that estimate of genetic merit against the phenotype using the plot() command.

```
using Gadlfy  # You may need Pkg.add("Gadfly")
plot(x=y,y=Za)
```

If you are using the Julia notebook the plot will appear in the notebook, if you are using the Julia terminal, the plot will appear in your browser.

We typically have to compute reliabilities of estimated breeding values. The reliability for any arbitrary contrast **k**, can be calculated as linear function of the **G** and **C22** matrices as follows

$$r^2_{\mathbf{k'\hat{a}}} = \frac{\mathbf{diag}\left[\mathbf{k'}\left(\mathbf{G} - \mathbf{C22}\sigma^2_e\right)\mathbf{k}\right]}{\mathbf{diag}\left[\mathbf{k'Gk}\right]}.$$

In mixed models, any linear combination of random effects is estimable, so conformable **k** can contain any elements. One meaningful choice of **k'** is the elements of a row of **Z**, as that contrast estimates the linear combination of random contributions relevant to a particular animal. The reliabilities of all animals can be simultaneously predicted using the entire **Z** matrix in place of **k'** in the above equation. Compute the breeding values of all the animals and their corresponding reliabilities.

Example 4: Directly fitting animal effects using genomic relationships.

Rather than estimating allelic effects at every locus, an equivalent model can be derived that directly solves the animal effects in the appropriate mixed model equations. This formulation of the problem in the usual representation of the mixed model equations will only work when the genomic relationship matrix is full rank. The genomic relationship matrix will not be full rank if there are more animals than loci or if any two animals have identical genotypes.

Reduce nanim to 3 and recompute the quantities in example 3. The animals in the original Hayes datafile have been reordered so that the genomic relationship matrix is full rank for the first three animals.

Form the genomic relationship matrix as **ZGZ'**, and invert it using inv(). Form and solve the mixed model equations, and compute the reliabilities for each animal. In computing the reliabilities, note that the matrix you previously used for **G** should now be replaced by **ZGZ'**. To fit animal effects directly, use the mixed model equations in the form below where the previous incidence matrix for the random effects has been replaced by the matrix **I** of appropriate order.

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'} \\ \mathbf{X} & \mathbf{I} + \sigma^2_e\left[\mathbf{ZGZ'}\right]^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}}^0 \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{y} \end{bmatrix}$$

Compare your results to the answers (for the animals) you obtained in example 3. They should be identical if you use the same animals in predicting the SNP effects.

Example 5: Alternative parameterizations fitting substitution effects rather than allelic effects.

Modify the **Z** matrix by reading only columns 1, 3 and 5 (or 2, 4 and 6). This allows you to fit substitution effects rather than both allelic effects. You will also need to appropriately alter the order of **G** and double the genetic variance for substitution

effects for each locus compared to allelic effects because $\text{var}(\alpha) = \text{var}(a_1 - a_2) = \text{var}(a_1) + \text{var}(a_2) = 2\,\text{var}(a)$. If you don't recode the new **Z** matrix, you have effectively modified the overall mean and the estimated breeding values will all be altered by a constant compared to the previous questions. This is no problem in real life, as breeding values are typically rescaled to a consistent base after computation and prior to publication of the results.

You may want to further experiment by subtracting 1 from every element of **Z**, so each SNP is coded -1, 0 and 1 rather than 0, 1 or 2.

For the modified incidence matrices, repeat example 1, fitting the GLS equations, example 2, fitting the mixed model equations for substitution effects and example 3, fitting the genomic relationship matrix. These three models are equivalent to each other and should give the same solutions to each other for this parameterization. You should also find that the solutions for substitution effects or animals are the same as you obtained in examples 1-3 except the breeding values may differ by a constant depending upon your parameterization. The fixed effects solutions will not be the same, neither will the prediction error variances or reliabilities of predicted random effects be typically identical.