

ALA 7.3**Name:** Sarah Yeo**Score:****Files:**

ALA 7.3_mrkrA.csv
ALA 7.3_mrkrB.csv
ALA 7.3_phenoA.csv
ALA 7.3_phenoB.csv
GS demo.R

Goal:

1. Determine prediction accuracies of RRBLUP for two distinct breeding projects.
2. Determine the optimal training set size for Genomic Prediction using rrBLUP

Useful R commands

- rm()
- attach()
- mixed.solve

Keywords: Genomic Prediction, GBLUP, RRBLUP, Heritability, prediction accuracy, MSE**References:**

Class notes on Linear Mixed Model Equations
Chapter 11, Bernardo
Meuwissen. pdf
Introduction to Genomic Selection in R.pdf

Activity:

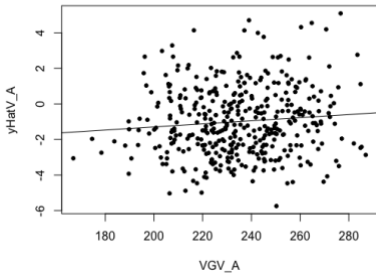
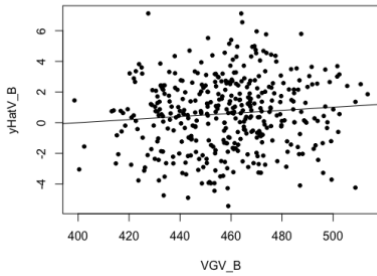
In ALA 7.1 and 7.2 we learned that it is possible to predict using an LMM to obtain useful predictions, particularly with fewer plots and with data from unbalanced data sets. However, we were still constrained to include every line at least once among the field trials. Genomic Selection has enabled the ability to predict performance of lines even if they have not been evaluated in any field trials.

Consider four data sets, two with genetic scores obtained using a 6k SNP chip (ALA 7.3_mrkrA.csv and ALA 7.3_mrkrB.csv) and two consisting of 2000 BLUP values for a phenotypic trait exhibiting reliability of 0.7 (ALA 7.3_phenoA.csv and ALA 7.3_phenoB.csv). Both marker and phenotypic data are from 2000 testcrossed lines derived in the F₅ generation from 20 crosses of homozygous lines in breeding projects A and B. Currently, the breeding projects are evaluating the lines in single rep tests with 1/2 of the lines evaluated at each location, similar to what you did in ALA 7.2. The SNP scores have been translated to -1,0,1 scores and have no missing data at ~4300

markers. Unlike the demonstration of how to use RRBLUP, we do not know the true genotypic values of these lines.

- Using rrBLUP and a training set of 1600 of the 2000 lines, determine the prediction accuracy for each of the two breeding projects.

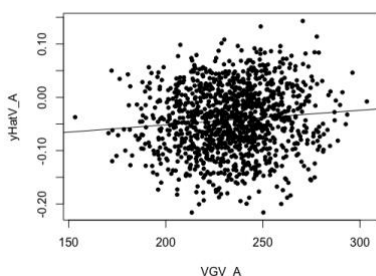
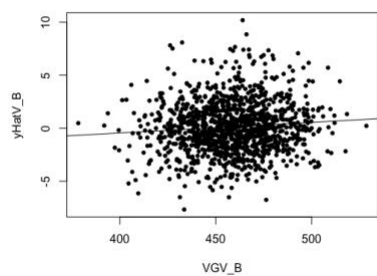
Using 80% of the data for training (sample.split function used to split sample) I got the following results via R

Marker A	Marker B
 <p>Correlation = 0.0970609 mse = 56058.42</p>	 <p>Correlation = 0.09339572 mse = 209770.9</p>

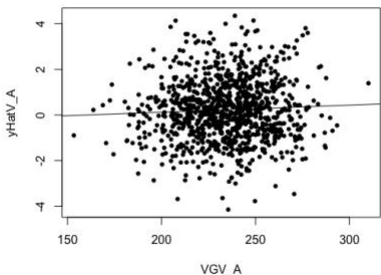
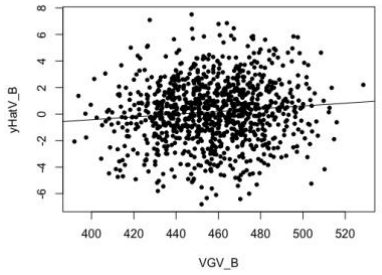
As we can see, both models produce a low prediction accuracy (correlation) but marker A is a slightly better model in comparison.

- Beginning with 40% of the total number of lines determine the optimal training set size for each of the breeding projects (see page 36 of “Introduction to Genomic Selection in R.pdf” for a demonstration of how to do this.

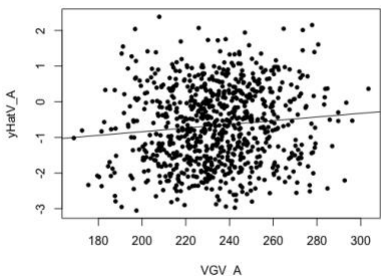
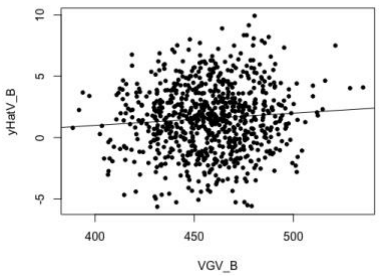
Using 40% of the data for training (sample.split function used to split sample) I got the following results via R

Marker A	Marker B
 <p>Correlation = 0.09863667 mse = 55112.9</p>	 <p>Correlation = 0.08839376 mse = 210518</p>

Using 50% of the data for training (sample.split function used to split sample) I got the following results via R

Marker A	Marker B
 <p>Correlation = 0.05107964 mse = 55314.51</p>	 <p>Correlation = 0.09902934 mse = 209341.1</p>

Using 60% of the data for training (sample.split function used to split sample) I got the following results via R

Marker A	Marker B
 <p>Correlation = 0.1059665 mse = 55495.24</p>	 <p>Correlation = 0.08086955 mse = 208497.5</p>

If we compare correlations to what we got in question 1 we get the following table and can see that this is not a great model using marker A or B.

	A-cor	B-cor	A-mes	B-mse	A-vg	B-vg	A-i ²	B-i ²
40%	0.09863667	0.08839376	55112.9	210518	1,239.76	1101.1038	0.4499	0.1046
50%	0.05107964	0.09902934	55314.51	209341.1	1,162.40	1,078.54	0.4203	0.1030
60%	0.1059665	0.08086955	55495.24	208497.5	1,218.46	1,132.89	0.4391	0.1087
80%	0.0970609	0.09339572	56058.42	209770.9	1,071.68	1,136.80	0.3823	0.1084

Prediction Accuracy (r_{MG}):

Comparison:

	A	B
40%	14.71%	27.33%
50%	7.88%	30.85%
60%	15.99%	24.53%
80%	15.70%	28.37%

	A	B
80%	15.697%	28.369%
40%	-0.991%	-1.039%
50%	-7.818%	2.481%
60%	1.294%	-3.837%

As we can see, the best model out of these choices for marker A is the one generated with 60% of the data and the best model out of these for marker B is with 50% of the data.

3. Make a recommendation to management about whether breeding decisions can be made with fewer lines evaluated in field plots.

I suggest that models like this have potential value for decision-making with fewer lines in field plots, but a better model needs to be developed before any such approach has any real value. This can be accomplished through more representative training sets and data collection in relation to the lines you are trying to assess. In theory, you could get a better model simply by chance from running the code again to generate a new random set of lines to generate your model with. Still, you would not expect to see a significant improvement that way since our reliability is low.