Comparison of marker-based Predictions using RRBLUP and Bayesian LASSO

By Anita Bhandari Sharma *Supervisor: Dr. Zheng Xu*

ABSTARCT:

The main objective of this project was to compare the predictions from RRBLUP and Bayesian LASSO. Markers were used to predict two traits- BMI and body length related to obesity. The gender of the mice was used as a covariate in predictions. The dataset was divided into training and validation sets for cross-validation. The Bayesian LASSO run with 1000 iterations showed better prediction than RRBLUP. The prediction accuracy was measured through root mean squared errors (RMSE) and correlation between the predicted and observed values of the response variables in the validation set.

INTRODUCTION:

Genomic predictions have emerged as a powerful tool to improve the selection of individuals before phenotyping in plants and livestock. The reason behind this is that it enables breeders to make selections at an early stage to ensure rapid genetic gain and reduce phenotyping costs.

Thus, prediction accuracy is what we are looking for to achieve the goal of the best selections. It would be of interest to compare different models or approaches to get the best predictions. Several methods, such as ridge regression, ABLUP, GBLUP, etc., are available in the literature for making predictions. This project only aims to compare two widely used methods for the predictions based on markers.

METHODS AND MATERIALS:

The “Mice” data used are available in R under the BGLR package. The mice data comes from the Wellcome Trust (http://gscan.well.ox.ac.uk), an experiment carried out to detect and locate QTLs for complex traits in a mice population (Valdar et al. 2006a; 2006b). There are 1814 individual mice, each genotyped for 10,346 polymorphic markers. The dataset used in the study already has SNPs with minor allele frequency (MAF) smaller than 0.05 removed and missing marker genotypes imputed with the corresponding average genotype calculated with estimates of allele frequencies derived from the same data. The traits related to obesity are body mass index (BMI) and body length. The data also contains additional information about gender, litter, cage density, etc. Litter denotes which litter the mouse came from (‘‘litter’’; e.g., ‘‘3’’ means the animal came from his parents’ third litter). However, for simplicity, only gender was taken as the covariate for this study. Gender was coded as numeric 0 and 1 for female and male mice. Genomic predictions have emerged as a powerful tool to improve the selection of individuals before phenotyping in plants and livestock. Thus, prediction accuracy is what we are looking for to achieve the goal of the best selections. Our objective is to compare RRBLUP and Bayesian lasso models or approaches to get the best predictions.

Ridge regression (RR), which is equivalent to best linear unbiased prediction (BLUP) in the context of mixed models (Whittaker et al., 2000; Meuwissen et al., 2001) is one of the first popular methods proposed for genomic selection. The ridge regression–best linear unbiased prediction (RR–BLUP) assumes that all marker effects are normally distributed with marker effects having identical variance (Meuwissen et al. 2001) and equal shrinkage to zero for all marker effects. The basic RR-BLUP model is

$$\boldsymbol{y}|\boldsymbol{u}\sim \boldsymbol{N}(\ \boldsymbol{X\beta}\ +\boldsymbol{Zu}\ ,\boldsymbol{I}\sigma_e^2)$$

$$\boldsymbol{u}\sim N(0,\boldsymbol{I}\sigma_a^2)$$

$$\boldsymbol{e}\sim N(0,\boldsymbol{I}\sigma_e^2)$$

Where X is a matrix of covariates taken as fixed effect, and Z refers to a design matrix for SNP effects, rather than an incident matrix and u refers to SNP effects. The above mixed- model, in which there is a single variance component other than the residual error, has a close relationship with ridge regression where, the ridge parameter $\lambda = \frac{\sigma_e^2}{\sigma_u^2}$, is the ratio between the residual and marker variances. The BLUP solution for the marker effects can be obtained by solving mixed model equation (i) as $\hat{u} = GZ'V^{-1}(y - X\hat{\beta})$, G being the relationship matrix not the genotype and V being the variance of y.

$$\begin{bmatrix} \boldsymbol{X^TX} & \boldsymbol{X^TZ} \\ \boldsymbol{Z^TX} & \boldsymbol{Z^TZ}+\boldsymbol{I}\lambda \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X^Ty} \\ \boldsymbol{Z^Ty} \end{bmatrix} \qquad \ldots\ldots.(i)$$

Bayesian LASSO is a popular model that combines model selection with prediction. The basic linear model is given by:

$$y|\beta,\sigma^2\sim\ N(X\beta, I_n\sigma^2)$$

with the conditional Laplacian prior to β_i as

$$\beta_i|\sigma^2 \sim \frac{\lambda}{2\sigma} e^{-\frac{\lambda|\beta_i|}{\sigma}}$$

where conditioning on $\sigma^2$is important to guarantee a unique posterior model. The model estimate is given by,

$$\hat{\beta}_{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \epsilon_i^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

The BMI and body length traits were analyzed by using RRBLUP and Bayesian lasso models described above, and prediction accuracies are compared using the prediction error and correlation between observed and predicted phenotype. For validation, holdout validation approach is used. For this, the dataset is divided into training and validation set and then, use the results from fitted model on training set to predict the phenotype for the validation set. All analyses were performed in R using rrBLUP and BGLR packages.

RESULTS AND DISCUSSION:

Genomic selection increases the rate of genetic improvement and reduces cost of progeny testing by allowing breeders to preselect organism that inherited chromosome segments of greater merit (Meuwissen et al., 2001; Schaeffer, 2006). Single nucleotide polymorphism (**SNP**) markers can now cover the genome with high density and are inexpensive to obtain. Thus, markers were utilized to make predictions using RRBLUP and Bayesian LASSO.

The dataset was split into 60% and 40% as training and validation set. The holdout strategy was used for cross-validation. That is, the results from model with training data set were used to predict for the validation set. The accuracy of fit was tracked using correlation between predicted and the observed responses and the root mean squared predicted errors. For Lasso 1000 iterations were performed. The plots in fig1 and 2 show the predicted values against the observed values of BMI and body length respectively in the validation set. Both RRBLUP and Bayesian LASSO method show similar predictions for the traits.

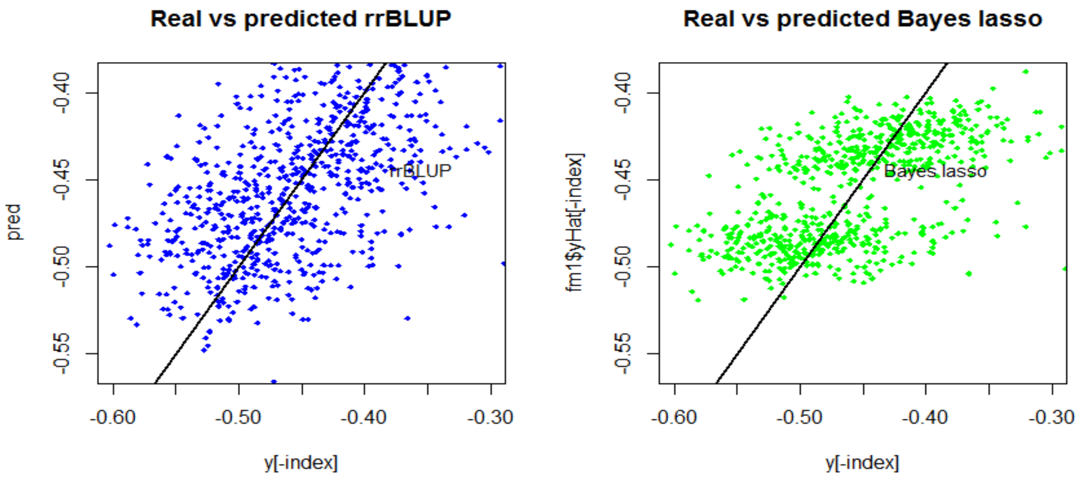

Fig-: Plots for BMI

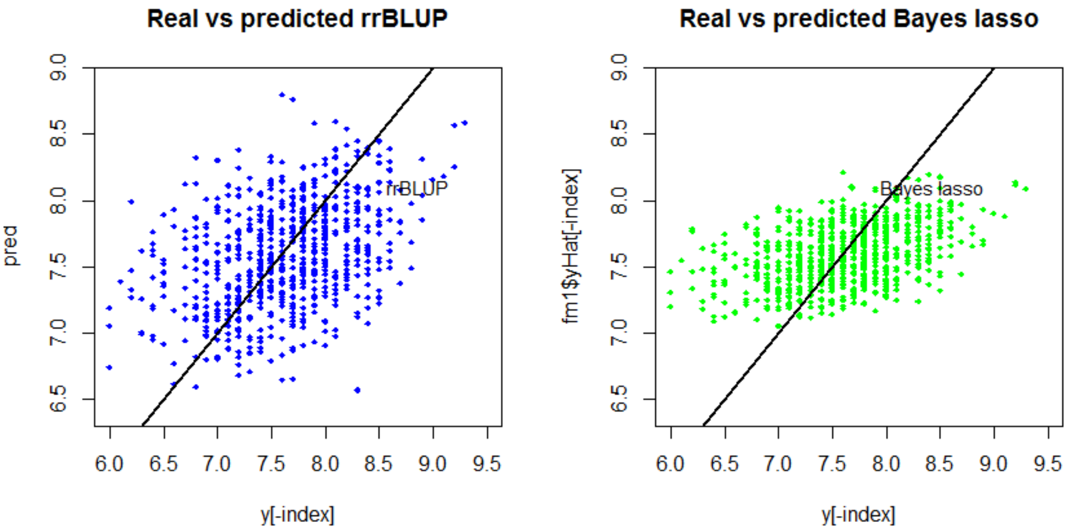

Fig-2:Plots for body length

The correlation between the predicted and observed BMI from the two methods are given in table 1. The predictions from Bayesian lasso has higher correlation to the observed values. Also, the root mean squared prediction errors show the same result.

For BMI	Correlation	RMSE
RRBLUP	0.5154682	0.05359368
BAYES LASSO	0.5661756	0.04709157

Table-1

The correlation between the predicted and observed body length from the two methods are given in table 2. Again, the predictions from Bayesian lasso has higher correlation to the observed values and the root mean squared prediction errors show the better fit for lasso. The predictions for body length are closer than that for BMI.

For Body length	Correlation	RMSE
RRBLUP	0.3905411	0.5493247
BAYES LASSO	0.4298184	0.4784702

Table-2

CONCLUSION:

The objective of this study was to compare two prediction methods-RRBLUP and Bayesian LASSO, based on markers. The BMI and body length of mice related to obesity were used as traits in both models. Bayesian LASSO outperformed RRBLUP in the prediction of both traits. However, the predictions were closer for body length compared to BMI. One possible reason for the good performance of the Bayesian LASSO model might be the fact that LASSO does model selection along with ridging, whereas no such implications were done in the RRBLUP model.

REFERENCES:

- Valdar, W., L. C. Solberg, D. Gauguier, S. Burnett, P. Klenerman et al., 2006a Genome-wide genetic association of complex traits in heterogeneous stock mice. Nat. Genet. 38:879-887.
- http://pbgworks.org/sites/pbgworks.org/files/Introduction%20to%20Genomic%20Selection%20in%20R.pdf
- http://pbgworks.org/node/1440
- https://cran.r-project.org/web/packages/rrBLUP/rrBLUP.pdf
- https://cran.r-project.org/web/packages/BGLR/BGLR.pdf
- http://genomics.cimmyt.org/BGLR-extdoc.pdf