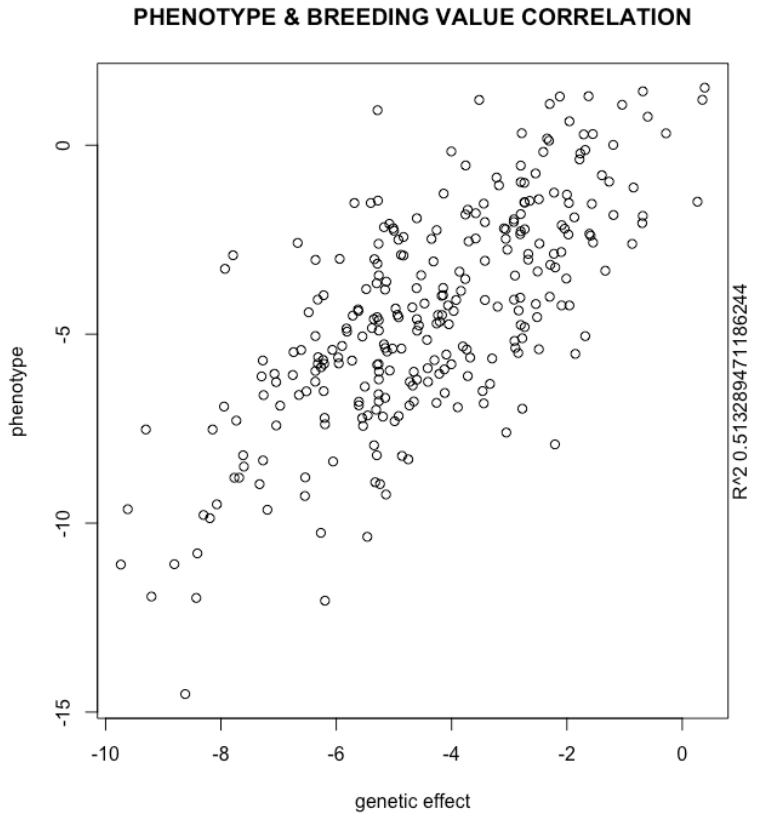
Samuel Revolinski
Hwk #6
Crops 545

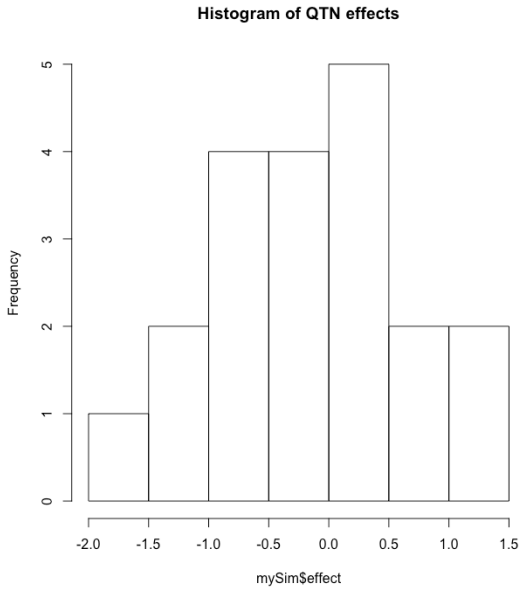
(1)
Hypothesis: The effects of the of the QTN will follow a normal distribution and it will be apparent on the histogram. The breeding values and actual phenotype will be correlated but not perfectly correlated as the breeding values are only the additive effects and does not take into account interaction of genes (epistasis) or dominance or unknown variables influencing it in the environment a correlation between 0.6 and 0.9 is likely as additive effects much of the time take up a large fraction of the variance explaining a trait.
Methods: Gapit functions were used with genetic data from the zzlab website to simulate a phenotype with 20 QTN controlling the trait and a heritability of 0.5 percent. A histogram was made of the QTN effects and the breeding value was plotted against the phenotype.

Figure 1. plot of phenotype against breeding value with R^2.



In figure 1 it can be seen that there is a reasonable R^2 of 0.51(correlation is 0.71) between the breeding value which is the additive effects added up basically and the phenotype. The breeding value could do a reasonable job of predicting the phenotype in this case.

Figure 2. Histogram of effects of the simulated QTN.



In figure 2 it appears that the simulated QTN effects follow a normalish distribution. Both parts of the hypothesis were not disproven. This shows that for traits following a normal distribution with 0.5 heritability and 20 QTN that using the breeding value for selection may be useful for genetic gains.

(2)

Hypothesis: This cross validation will show the mean of the correlations to be at least 0.4 for the phenotype and super high (0.9) for the breeding value because farmCPU is a pretty good model for genomic selection these days and the breeding value will be similar. The standard deviation will be below 0.10 because the model has multiple genes and according the central limit theorem the more samples we have (effect of a gene) the closer it should get to the mean effect.

Methods: Using the simulated effects in from the gapit function in problem 1 FARMCPU was implemented with the genetic map and 3 PCA's generated from the gapit function on the data. The 20 most significant snp were found using the farm CPU. Then for 30 replications the populations were split in two even parts, then one half (training population) was used to come up with the effects for each gene and then those effects were used to predict the other half (testing), this was done for both the phenotype and the breeding value and the correlation was collected from each rep. The means and standard deviations were then taken for phenotype correlations and breeding value correlations. Gapit used to get effects of snp

Presentation: ***Table 1. Mean and SD of phenotype and BV correlations (separately)***

	Mean	Standard Deviation
Phenotype Correlation	0.557058	0.05443024
Breeding Value Correlation	0.5617456	0.0650307

Results: In table 1 is can be noticed that the Breeding Value Correlation and phenotype correlations in the cross validation are not that different which surprised me as I expected the breeding values to be highly correlated because I assumed it would get the same thing for the 20 genes but apparently not as much as I thought. The hypothesis was not proven incorrect for

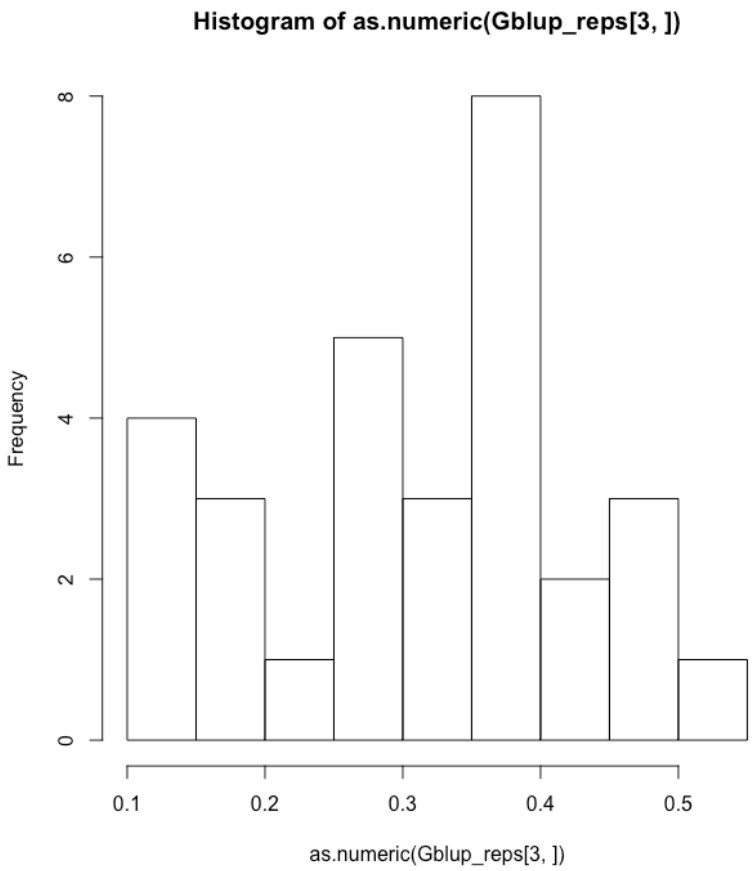
the phenotype correlations while is was lower then expected for the breeding values which means the breeding values are not static between sections of the population. The standard deviations did not disprove the hypothesis as they were both below 0.10.
(3)
Hypothesis: Shuffling around the values will cause the correlations for the phenotype and the breeding value to drop near zero because the effects will not be lined up with the genes correctly so their wont be a very big correlation.
Methods: All the rows were sampled by index randomly, then basically the same methods from problem two were used to get the correlations for phenotype and breeding value.
Presentation: ***Table 2. Mean and SD of phenol and BV correlations with shuffled data***

	Mean	Standard Deviation
Phenotype Correlation	0.004765755	0.004718135
Breeding Value Correlation	0.03777792	0.04644885

Results: basically all the correlations were zero because genetic data did not line up correctly to the phenotypic data. The hypothesis was not disproven. I will say though the mean correlation of the breeding value is about 10x higher then the phenotype, I think it could be that way because the breeding values it calculates are similar because the genes are the same genes it uses even though the phenotype is off. The correlations were much lower then in problem 2 because of the shuffling.
(4)
Hypothesis: Since Gblup is a well used prediction model and 80% of the population is being used for training the effects the correlation will be high like 0.8 or 0.9 for both the breeding values and the phenotypes in the training populations while high but still lower (0.6 or 0.7 ish) in the testing population. The testing should have higher sd for correlation because it is not on the set it was trained on.
Methods: inside the 30 replications a random sample of indexes for 80% of the population is made and this is the training data while 20% is saved to be the testing data. Then gapit gblup is trained on the training data, then the effects of the markers are used to predict the breeding value and phenotype then those are compared to get the correlations between the testing predictions and training population. After the 30 replications the mean and standard deviations of the correlations are calculated.
Presentation: ***Table 3. Mean and SD of correlations from testing against training population***

	Mean	Standard Deviation
PhenTrain corr	0.9301768	0.01394302
PhenTest corr	0.3136942	0.1165891
BVTrain corr	0.7097225	0.01502776
BVTest corr	0.3911833	0.1055311

Figure 3. Histogram of correlations between training population and testing population using Gblup



Results: In table 3 you can see that for the both the phenotype and BV the correlations in the training is 0.9 or 0.7 which was expected by the hypothesis and doesn't disprove it while the testing population correlations for both phenotype and breeding value is about 0.3 which is much lower then expected. The standard deviations are higher in the testing then the training which does not disprove the hypothesis and makes sense because the testing population is not the one the models were trained to. Figure 3 shows that the correlations follow something that looks like and under simulated normal distribution where not enough samples have been taken to fill all the gaps but in general the highest point is in the middle with the lowest on the out side. This means that the correlations with the testing population for many trails may follow a standard normal distribution but more work needs to be done.

(5)

Hypothesis: Ridge regression should perform better then the gBlup used in for the phenotype and BV of testing population with trained model. Ridge Regression is better for less complex traits then gblup and with only 20QTN simulated the trait is not super complex and rrBlup should give higher correlations. The 5 fold method should yield lower SD's for testing pop correlations then problem 4 because the mean in 5 fold is a mean of means because there are 5 replications within each replication over all, one for each fold.

Methods: For 30 replications the total population was cut into 5 equal groups then for each group that group was used as training population and used to calculated predictions of

phenotype and breeding values for the rest of the populations including the one used as training. The correlations were averaged for the 5 pops with each group, then those 30 averages were collected. Means and standard deviations were then calculated from those averages.

Presentation: ***Table 4. Means and sd's of correlations from 5 fold validation on rrBLUP.***

	Mean	sd
Phenotype correlation	0.3439959	0.03484124
Breeding Value Correlation	0.4420499	0.02827126

Results:

The mean testing population phenotype correlation is slightly higher in problem 5 then problem 4 so the hypothesis was not disproven, but the sd in problem 5 for both phenotypes and breeding values is much lower then in problem 4. Overall the rrBlup seems to have performed slightly better then gblup but it really is only a couple percent better. The hypothesis that rrBLUP would perform better and more consistently was not disproven.