**Homework 6**

**Statistical Genomics**

**CROPS 545, Spring 2020**

Professor: Zhiwu Zhang

Due on May 1, 2020, Friday, 3:10PM, PST

**Data files**: You can use either the same dataset you used in homework 2, or switch to a different dataset with same requirements.

**Hand in:** Email your report (PDF, limited to five page) and R source code (text file) with subject of “CROPS545 HW6” to Zhiwu.Zhang@WSU.edu. Name your files as following:

Homework6\_ firstname\_lastname.pdf and Homework6\_ firstname\_lastname.R

**Grade components**: 1) Hypothesis or statement; 2) Results; 3) Methods; 4 presentation; 5) R source code (clarity, simplicity and documenting comments)

**Objectives**: 1) validation; 2) invalidate validation; 3) cross validation; 4) gBLUP; and 5) ridge regression.

1. Use GAPIT.Phenotype.Simulation function to simulate phenotypes with heritability of 75% controlled by 10 QTNs having effects with standard normal distribution. Display the QTN effects. Create 3D plots on scatter plots of total genetic effects (breeding values), residual effects, and phenotypes of individuals (5 points).
2. Conduct GWAS with BLINK on the entire population using phenotypes dependent variable. Identify SNPs that pass 1% threshold (after Bonferroni multiple test correction). Plot estimated effects of the associated SNPs against their true effects. (5 points).
3. Randomly sample 80% of individuals from the entire population as the training population and the rest as the testing population. Conduct GWAS with BLINK on the training population using phenotypes dependent variable. Identify SNPs that pass 1% threshold (after Bonferroni multiple test correction). Plot estimated effects of the associated SNPs against their true effects. (15 points).
4. Use the effects estimated in (3) for SNPs passed 1% threshold (after Bonferroni multiple test correction) to predict the phenotypes of the individuals in the training population and testing population for Marker Assisted Selection (MAS). Calculate the correlation between the observed and predicted phenotype for training and training population and testing population separately. Explain the difference between the training and testing population (15 point).
5. Conduct Ridge regression on the training to estimate all marker effects. Use the estimated marker effects to predict the phenotypes of the individuals in the training population and testing population. Calculate the correlation between the observed and predicted phenotype for training and training population and testing population separately. Compare Ridge regression with MAS. (15 point).
6. Conduct gBLUP on both training population and testing population with the phenotypes masked in the testing population. Calculate the correlation between the observed and predicted phenotype for training and training population and testing population separately. Compare gBLUP with Ridge regression. (15 points).
7. Redo (6) with the SNPs identified in (3) as fixed effects for GWAS assisted GS (15 points).
8. Redo (6) with the SNPs identified in (2) as fixed effects for GWAS assisted GS. Explain the difference from (7) (15 points).

**Extra credit**

1. Repeat (3-8) at least 30 times and compare prediction accuracy based on their averages and standard errors (30 points).
2. Redo (2) using residual instead of phenotypes as dependent variable and redo (9) (30 points).