**Homework 6**

**Statistical Genomics (545)**

**Spring 2021**

Professor: Zhiwu Zhang

TA: Zhou Tang

Due on Friday, April 30, 2021, 3:10PM PST

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

**Hand in:** Email your R RMD and Knit (HTML, PDF, or Word) with subject of “StaGen545 HW6” to Zhiwu.Zhang@WSU.edu. Name your files as

“Homework6\_ firstname\_lastname”.

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

1. Use the 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 phenotype as the dependent variable. Identify SNPs that pass the 1% threshold (after Bonferroni multiple test correction). Plot the 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 phenotype as the dependent variable. Identify SNPs that pass the 1% threshold (after Bonferroni multiple test correction). Plot the estimated effects of the associated SNPs against their true effects (15 points).
4. Use the effects estimated in (3) for SNPs that passed the 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 phenotypes for training population and testing population separately. Explain the difference between the training and testing populations (15 points).
5. Conduct ridge regression on the training population 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 phenotypes for training population and testing population separately. Compare ridge regression with MAS (15 points).
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 phenotypes for 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 phenotype as the dependent variable and redo (9) (30 points).