STAT 847: Analysis Assignment 2

DUE: Saturday March 23, 2024 by 11:59pm Eastern

Your assignment must be submitted by the due date listed at the top of this document, and it must be submitted electronically in .pdf format via Crowdmark. Organization and comprehensibility is part of a full solution. Consequently, points will be deducted for solutions that are not organized and incomprehensible. Furthermore, if you submit your assignment to Crowdmark, but you do so incorrectly in any way (e.g., you upload your Question 2 solution in the Question 1 box), you will receive a 5% deduction (i.e., 5% of the assignment's point total will be deducted from your point total).

There are a total of 30 marks.

Each genetic variable represents a single nucleotide polymorphism (SNP). That is, a single letter A, C, T, or G in the genetic code of that specimen. Specifically, it's a gene that isn't the same across all specimens, which makes it useful for statistical analysis.

This dataset contains all five phenotypes and the first 10,000 SNPs from a Genome-Wide Association Study of the species Arabidopsis thaliana, a plant.

Variable	Description
See:	https://easygwas.biochem.mpg.de/data/public/dataset/view/42/
Perimeter_Growth	The response variable
SNP_ABCD	The explanatory variable, gene number ABCD

Note: These have been coded into 0, 1, 2, or 3, so, while treating these continuous variables isn't the correct thing to do, we're going to do it anyways because a our methods will be able to pick out some of the important genes even with the mispecification.

Use the following code to load and split the data

```
library(randomForest)

dat = read.csv("F1-Hybrids_Pheno_10000genes.csv")

genes = dat[,9:10008]
pheno = dat[,1:8]
dat = NULL
```

1. (4 points) Using the `randomForest` function in `library(randomForest)`, make five random forests, each one using one of the phenotype variable `Perimeter_Growth` as a response y variable. The forest should use all 10,000 of the gene variables (These are the 9th, . . . , 10,008th variables). Give your forest 500 trees, have each tree use 300 gene variables, and set a minimum node size of 1. Sample with replacement. Report the percentage of variance explained by the forest using `print()`.

2. (2 points) Get a `hist()` of the `$importance` values from your random forest model of perimeter growth (not the MPH). Use this to comment on the relative importance of some genes over others in determining perimeter growth. Use 100 bins for the histogram.

3. (0 marks) Use the following code to make a new dataset that only includes perimeter growth and the most important 50 genetic variables from random forest for perimeter growth. `mod2` is the name of the `randomForest()` output in this case.

```r
# mod2 is
# Set a cutoff of the 50th most important variable
cutoff = rev(sort(mod2$importance))[50]

# Keep only those 50 variables
idx = which(mod2$importance >= cutoff)
genes_imp = genes[,idx]

dat_imp = cbind(pheno$Perimeter_Growth, genes_imp)
names(dat_imp)[1] = "Perimeter_Growth"
```

4. (4 points) Using `rpart`, and this new dataset `dat_imp` (or `genes_imp`) of the 50 most important variables for perimeter growth, create a single regression tree of perimeter growth. Plot the tree with `prp` in the `rpart.plot` package.

5. (4 marks) Using `regsubsets` in the `leaps` package, and the new dataset `dat_imp` (or `genes_imp`), use best subsets regression with the Adjusted R-squared criterion. Report the variables of the best model, their coefficients, and the adjusted r-squared of the model.

Hints:

To get the adjusted r-squared values, use `summary(regsubsets())`

To get a particular model, see https://stats.stackexchange.com/questions/193204/picking-a-particular-model-from-regsubsets

6. (4 marks) Run a PCA on the 50 important variables in `genes_imp`. Report the total (cumulative) variance explained by the first 10 principal components. Plot a scree plot.

7. (4 marks) Build a linear model of the response variable `Perimeter_Growth` using the first ten PCA dimensions from the previous question, and nothing else. Report the `summary(lm())`. Comment on the difference between this model's adjusted R-squared and the

The adjusted r-squared values for the top 10 PCs and the best subsets model are about the same.

8. (4 marks) Describe briefly one advantage and one disadvantage of the PCA-based model over the best subsets model. (There are several correct answers, but only the first two will be marked).

9. (4 marks) The variance inflation factor of an explanatory variable in a model is a function of how collinear that variable is with the over explanatory variables in the model are. The higher the number, the more collinear and the most the variance estimates of the slopes are being inflated by including that variable. We can find the variable inflation factor with `vif(lm())`, where `vif` is found in the `car` package.

Find the `vif()` of both the PCA-based model and best-subsets model.

Report the VIFs for both models and briefly explain why the PCA-based model has such low inflation factors (1 is the lowest possible).