STAT 425 **Applied Regression and Design** **Spring 2020**

Homework No. 8

1 Instructions

Due Date: Monday May 11th, 11:59 pm on Compass
Homework presentation should be neat. You should submit your homework through Compass. R codes should be submitted jointly with your report. **Please submit your digital files in .pdf or .html format**, so it is easier for the graders to open your files. If you feel it would help, you are encouraged to discuss Homework results with other students, but you have to present assignments individually using your own words. **Undergraduate students should attempt problems marked as UG. Graduate students should attempt problems marked as GR.**

2 Problems

1. **Problem 1 (GR)**: Four hundred and three African Americans were interviewed in a study to understand the prevalence of obesity, diabetes and other cardiovascular risk factors in central Virginia. Data is presented in diabetes. In this question we want to build a regression tree-based model for predicting glycosolated hemoglobin (glyhb) in term of the other relevant variables.

   (a) Plot the response against each of the predictors and comment on the apparent strength of the relationship observed.

   (b) Investigate the pattern of missing values in the data. By eliminating a combination of rows and columns, produce a reduced dataset that contains no missing values.

   (c) Fit the default tree. From the output answer the following questions: How many observations had $stab.glu < 158$? What was the mean response for these observations? What characterizes the largest terminal node in the tree? What is the mean response of this node?

   (d) Make a plot of the tree. What feature of the plot reveals the most important predictor?

   (e) Plot the residuals against the fitted values for this tree. Comment.

   (f) Select the optimal tree using Cross-validation. What would be the smallest tree that could be reasonably used?

2. **Problem 2 (UG)**: The data set *wbca* comes from a study of breast cancer in Wisconsin. There are 681 cases of potentially cancerous tumors of which 238 are actually malignant.

   (a) Fit a binary regression with *Class* as the response and the other nine variables as predictors. Report the residual deviance and associated degrees of freedom. Can this information be used to determine if the models fits the data? Explain.

(b) Use AIC as the criterion to determine the best subset of variables. (Use the *step* function).

(c) Suppose that a cancer is classified as benign if $p > 0.5$ and malignant if $p < 0.5$. Compute the number of errors of both types (false positives and false negatives) that will be made if this method is applied to the current data with the reduced model.

Note: All data sets are from the faraway library in R