**Due Date: April 1, 2020**

**PART I Reading Assignment: Chapter VIII of “An Introduction of Statistical Learning”**

**Answer the following questions based on your reading.**

**Problem 1 (2 Points)** Suppose that the data for study is where is a p-dimensional prediction vector. In regression analysis, we assume that the model has the form . (Hint: Page 314)

**Problem 2 (2 Points)** Suppose that the data for study is where is a p-dimensional prediction vector. In regression tree analysis, we assume that the model has the form . (Hint: Page 314)

**Problem 3 (2 Points) (True/False)** Even if the true function form of the model is well approximately by a linear model, the model built using decision trees is still better than the model built using regression.

False

**Problem 4 (10 Points)** Based on the textbook, Decision Trees for either classification or regression have four advantages over the traditional regression and logistic regression methods:

1. **Trees are very easy to explain to people. In fact, there are even easier to explain than linear regression.**
2. **Some people believe that decision trees more closely mirror human decision-making than other analytical methods.**
3. **Tree can be displays graphically and are easier to interpreted even by a non-expert.**
4. **Trees can easily handle quantitative predictors without the need to create dummy variables.**

* However, it has one major disadvantage that is **Trees have lower prediction accuracy than other tools .**

**Problem 5 (2 Points)** Suppose that be a random sample from a normal population with mean and variance . The variance of the mean is given by σ2/n. . This means that we can reduce the variance of almost all statistics through averaging.

**PART II Programming (12 Points)**

Data: The data set used is “ASS06\_DATA”.

**Problem 1 (6 Points)** Bagging for regression

DO K = 1 to 20;

STEP 01: Randomly select a sample of size 1,460 with replacement.

STEP 02: Build a **regression** model to predict the “HousePrice”

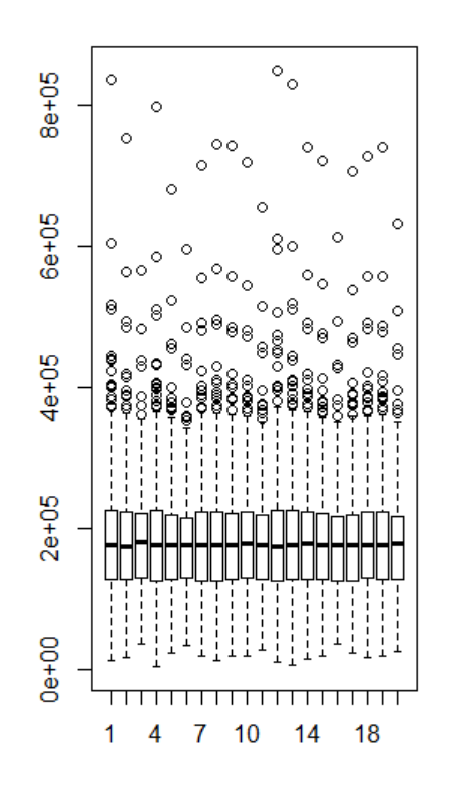
STEP 03: Use the model built in STEP 02 to score all 1,460 observation and keep all the predicted values for each observation

END DO;

STEP 04: Calculate the bagging estimator using the formula

STEP 05: Calculate the error estimator for the bagging estimator in “STEP 04” using the formula .

Produce a box plot for all 20 bagging predictions with “predicted” “HousePrice” as the Y-axis.

****

**Problem 2 (6 Points)** Bagging for decision trees

DO K = 1 to 20;

STEP 01: Randomly select a sample of size 1,460 with replacement.

STEP 02: Build a **decision trees** model to predict the “HousePrice”

STEP 03: Use the model built in STEP 02 to score all 1,460 observation and keep all the predicted values for each observation

END DO;

STEP 04: Calculate the bagging estimator using the formula

STEP 05: Calculate the error estimator for the bagging estimator in “STEP 04” using the formula .

Produce a box plot for all 20 bagging predictions with “predicted” “HousePrice” as the Y-axis.

