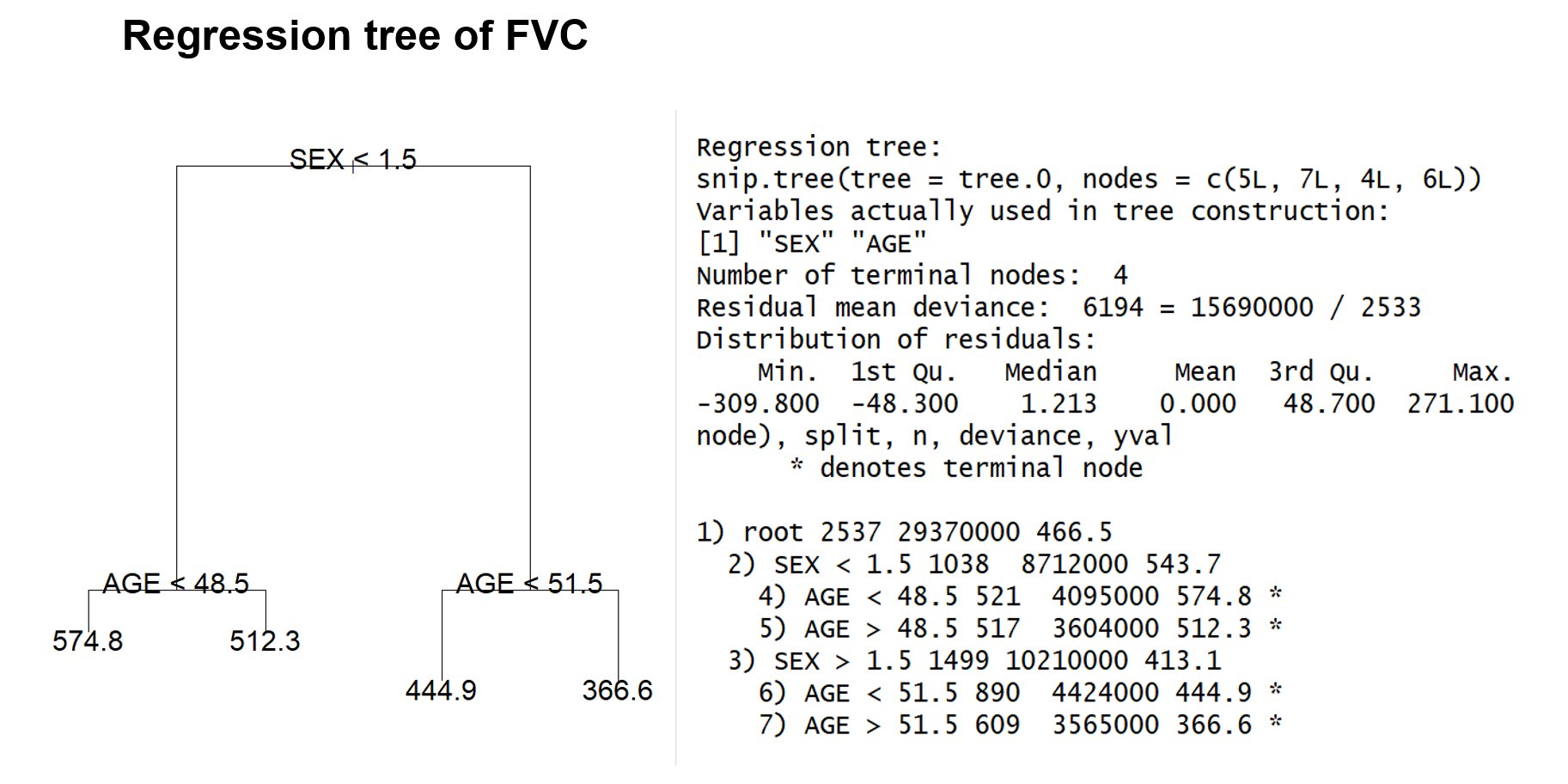
# **BS806 --- Homework 7**

**Reading**. Chapter 8 of (ISLR) Introduction to statistical learning

**Question 1**. (**70 pts**) The data set “**FHS\_data.csv**” includes variables SEX (1=male, 2 = female), Age, Smoke (=1 for current smoker), FVC (score of pulmonary function), SPF (systolic blood pressure), T2D (1=history of diabetes) and other variables. The data were used to generate the regression tree in the figure below.



1. (**20 pts**) Write down the 4 rules described by the 4 branches of the tree (i.e. the rules defining each group and their prediction).
2. (**5 pts**) What is the estimate of the variance of FVC from this tree?
3. (**10 pts**) Simulate a sample of **FVC values** from the displayed tree above (i.e. use the estimated mean, sd and the structure information. Let’s assume the sd is the same for all four leaves so you may use the information from previous question) using the existing variable values SEX and AGE in the data set “FHS\_data.csv” as predictors. Use the same sample size (i.e. we will generate 2,537 FVC values from this simulation) reported in the original tree/branch, and assume that FVC follows a normal distribution (i.e. use ***rnorm()*** function).
4. (**10 pts**) Compare the distribution of the simulated data for FVC and the real FVC values in the data set.
5. (**20 pts**) Use the simulated data for FVC and the real data for the covariates SEX, AGE, Smoke, SPF, T2D to construct a new regression tree, and control the depth of the tree using the option control=tree.control(nobs=nrow(data.frame), mindev = 0.001) where “data.frame” is the data frame created from the data set “FHS\_data.csv”. Use the **cv.tree()** function to select the best regression tree. Build the best regression tree and plot it. Include the plot in your write-up.
6. (**5 pts**)Compare the tree used to simulate the data (i.e. the displayed tree above) with the tree generated using your simulated data.

Include in your write up the R log with the details of the simulation and the analysis of the simulated data.

**Question 2.** The data set “FHS\_data.csv” includes variables SEX (1=male, 2 = female), Age, Smoke (=1 for current smoker), FVC (score of pulmonary function), SPF (systolic blood pressure), T2D (1=history of diabetes), **Death (occurred within 20 years** after measurements of all other variables).

1. (**10 pts**) Use this dataset to build a classification tree to predict death within 20 years, using the function tree() with default parameters and plot it. How many branches are represented by the tree?
2. (**5 pts**) What is the probability of death within 20 years for a person aged > 60 years based on the tree you constructed in (a)?
3. (**5 pts**) What is the predicted value for a person aged > 60 years based on the tree you constructed in (a)?
4. (**10 pts**) Generate a classification tree to predict death within 20 years using a training set of 2/3 of the observations and then predict the outcome in the test set comprising the remaining 1/3 observations. What is the misclassification error of the prediction? (Hints: using predict() function with type =”class”)

**Optional Question 3**. (**No point! For your own practice if interested in learning rpart()**) Use the data of question 2 to generate a classification tree using the rpart() function

tree.0 <- rpart( as.factor(dth) ~ SEX+AGE+Smoke+FVC+SPF+T2D, method="class", data=***data***)

rpart.plot(tree.0)