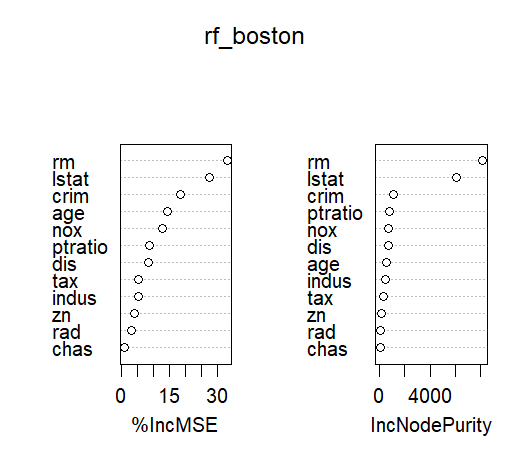
## Assignment 1 Part 3 (total 160 points)

## Problem 9-10 are for Topic 3.4: Various Tree Method

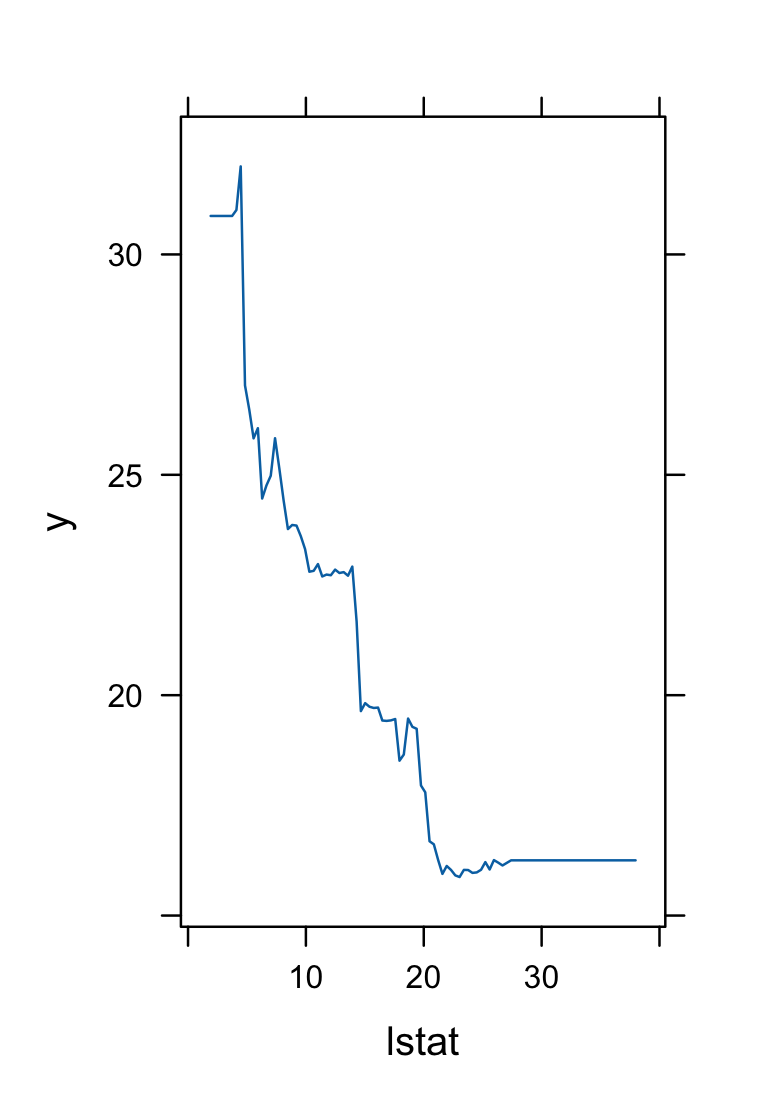
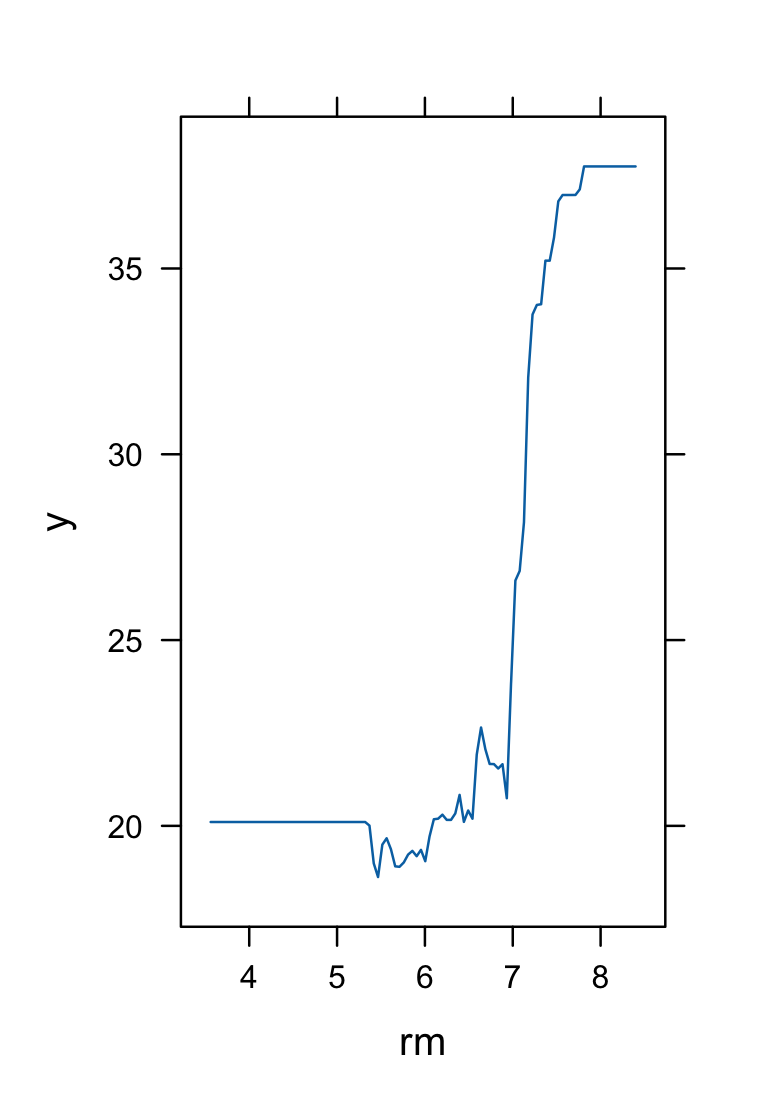
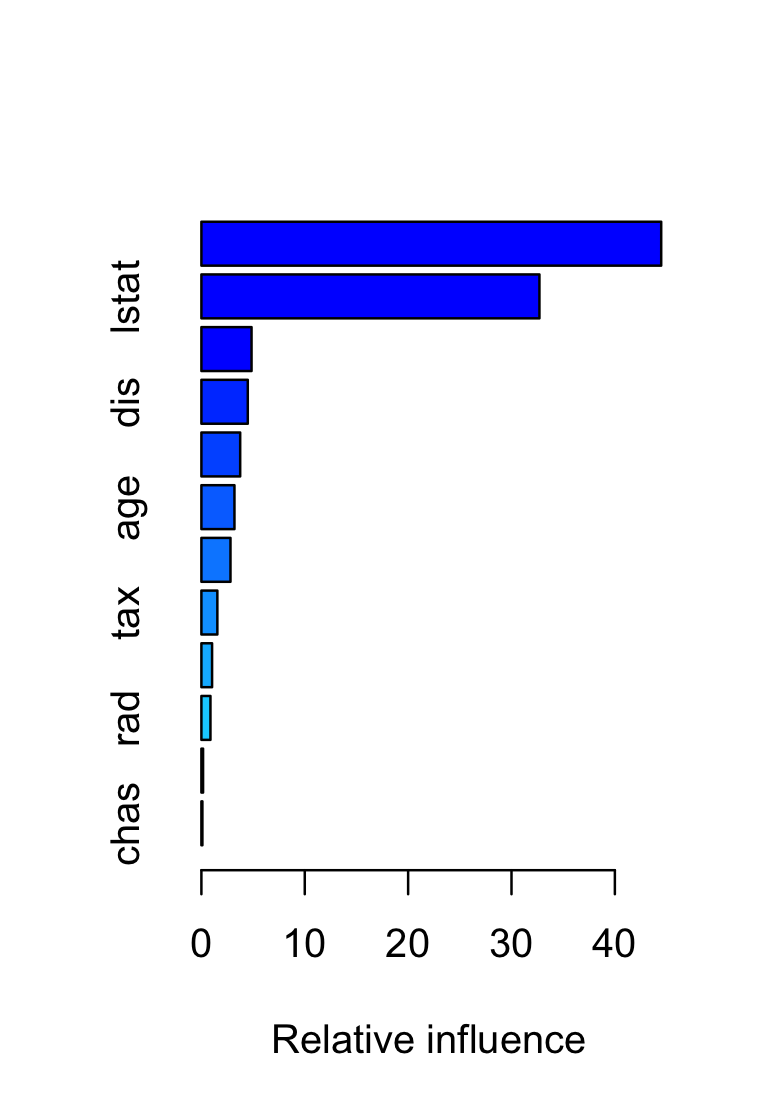
**Problem 9.** **(80 points total): various tree methods for regression tree**

Use **Boston** dataset, follow Lab 8.3.3-8.3.5. With the dataset, we wish to predict the median house prices of greater Boston area on the basis of various statistics associated with the house price. “medv” is the response variable. We practice various tree methods.

1. (40 points) **Bagging & Random Forest:** Use set.seed(1), randomly separate the dataset into half training dataset and half validation/test dataset. Call library(randomForest) and use the randomForest() function to build the trees
2. By default, randomForest() build 500 trees with Bagging method. Note you need to set mtry= the total number of the variables (12 for the Boston dataset). Use the predict() function to forecast with the test data from the trees you just built. What is the test MSE of the bagging method? MSE= 23.40359
3. Set ntree=25 to build 25 trees by Bagging method, Use the predict() function to forecast with the test data from the trees you just built. What is the test MSE of this Bagging method? Comparing to (a), is the performance better or worse? Why? MSE(25)= 24.59162 - it is worse for here as it is bigger compared to the previous one
4. Set mtry=6, which is smaller than the total number of the variables. Bagging becomes Random Forest. By default, randomForest() build 500 trees. Use the predict() function to forecast with the test data from the trees you just built. What is the test MSE of the Random Forest method? Comparing the results in (a) and (b), is it better or worse, why? MSE = 20.17751 and this one is the best oout of 3 we have seen
5. Use the importance() function to view and the varImpPlot() function plot the importance of each variable from the trees that you built in (c). What are your observations? Interpret the results.



1. (30 points) **Boosting:** (30 points) Boosting: Use set.seed(1), randomly separate the dataset into half training dataset and half validation/test dataset. Call library(gbm) and use the gbm() function to build the trees.
2. Set the tree size limit to 4, build 5000 trees for boosting. By default, shrinkage =0.001. Use summary() function to produce a relative influence statistics. Use plot() to produce partial dependence plot. interpret your plots.



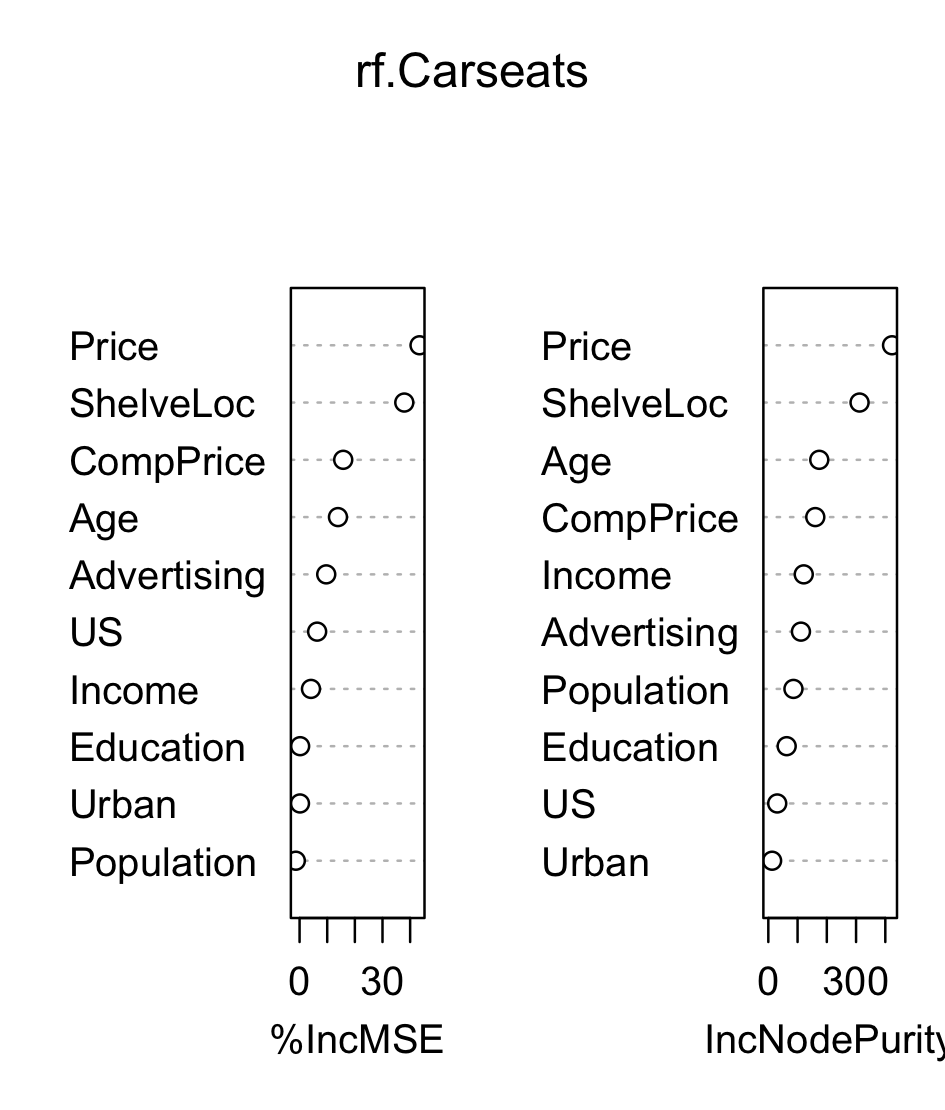
1. use the boosted model to predict “medv” on the test set, and check the test error MSE. It is 18.39057
2. Set shrinkage = 0.2, redo (a) and (b). How does shrinkage influence the model performance? 16.54778 and influence of the shrinkage is high
3. (10 points) BART: Use set.seed(1), randomly separate the dataset into half training dataset and half validation/test dataset. Call library(BART) and use the gbart() function to build the trees. use the boosted model to predict “medv” on the test set, and check the test error MSE

15.94718 and this one is the best out of all MSEs

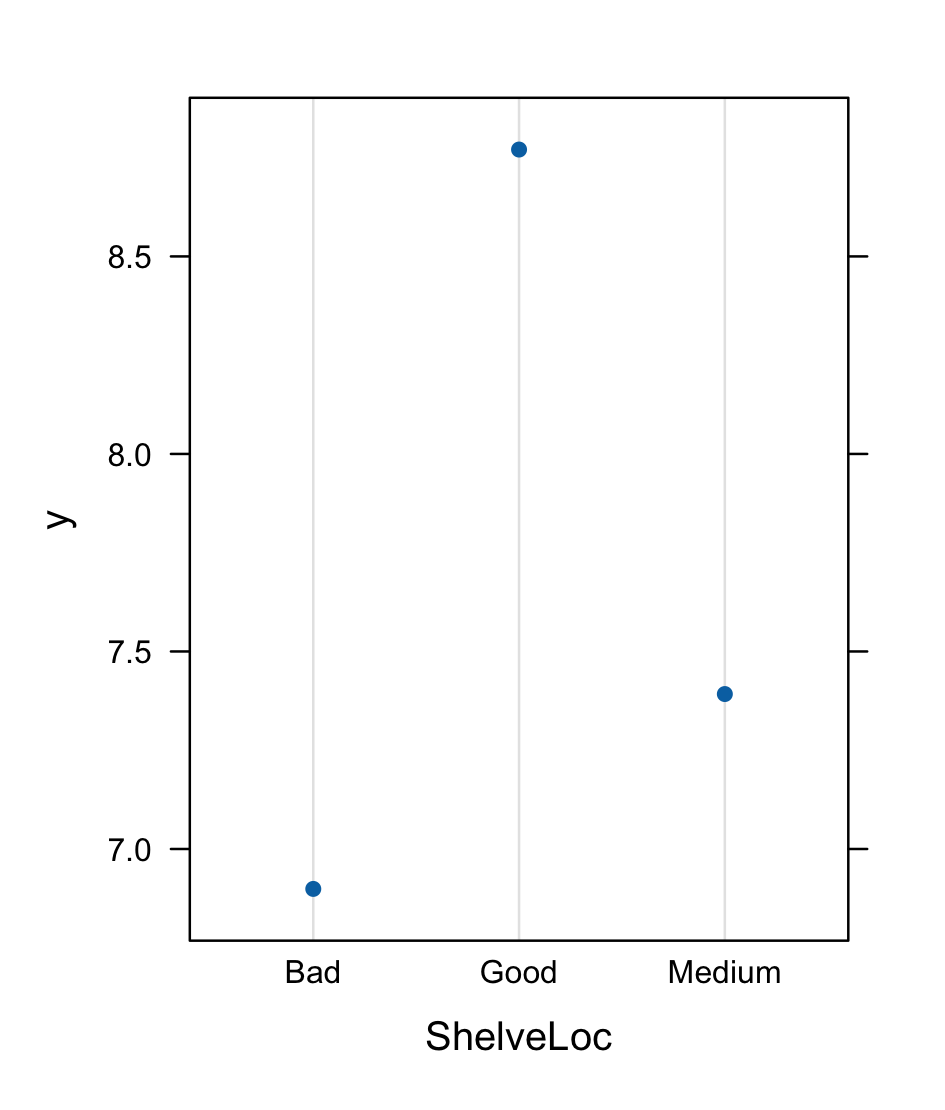
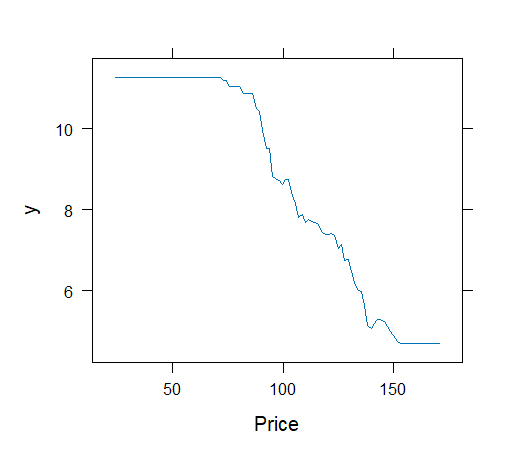
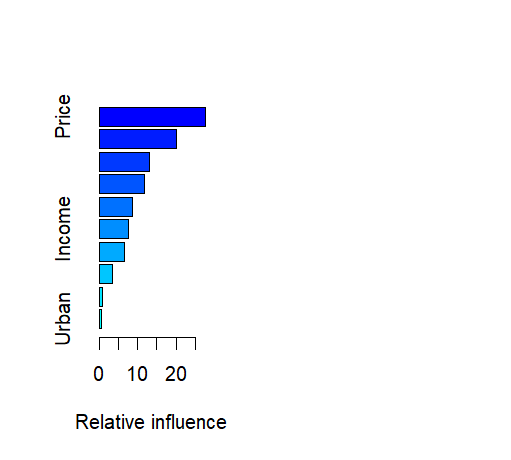
**Problem 10.** **(80 points total): Fit a regression tree**

Use **Carseats** dataset. With the dataset, we wish to predict the “Sales” of car seats on the basis of various statistics associated with the sale data in the previous year. We practice various tree methods.

1. (40 points) **Bagging & Random Forest:** Use set.seed(1), randomly separate the dataset into half training dataset and half validation/test dataset. Call library(randomForest) and use the randomForest() function to build the trees
2. By default, randomForest() build 500 trees with Bagging method. Note you need to set mtry= the total number of the variables. Use the predict() function to forecast with the test data from the trees you just built. What is the test MSE of the bagging method?
   1. 2.623527
3. Set ntree=25 to build 25 trees by Bagging method, Use the predict() function to forecast with the test data from the trees you just built. What is the test MSE of this Bagging method? Comparing to (a), is the performance better or worse? Why?
   1. 2.87522 which is worse because its a bigger number
4. Set mtry=4, which is smaller than the total number of the variables. Bagging becomes Random Forest. By default, randomForest() build 500 trees. Use the predict() function to forecast with the test data from the trees you just built. What is the test MSE of the Random Forest method? Comparing the results in (a) and (b), is it better or worse, why?
   1. 2.819087 which is the better than previous one because its the smaller value
5. Use the importance() function to view and the varImpPlot() function plot the importance of each variable from the trees that you built in (c). What are your observations? Interpret the results.
   1. What we notice is that Price and shelvloc are the most important variables in our model for predicting sales



1. (30 points) **Boosting**: Use set.seed(1), randomly separate the dataset into half training dataset and half validation/test dataset. Call library(gbm) and use the gbm() function to build the trees.
   * 1. Set the tree size limit to 4, build 1000 trees for boosting. By default, shrinkage =0.001. Use summary() function to produce a relative influence statistics. Use plot() to produce partial dependence plot. interpret your plots.



* + 1. use the boosted model to predict “Sales” on the test set, and check the test error MSE
       1. 1.899762
    2. Set shrinkage = 0.1, redo (a) and (b). How does shrinkage influence the model performance?
       1. 2.065894, it increases the MSE noticeably

1. (10 points) BART: Use set.seed(1), randomly separate the dataset into half training dataset and half validation/test dataset. Call library(BART) and use the gbart() function to build the trees. use the boosted model to predict “medv” on the test set, and check the test error MSE
   1. 0.184202