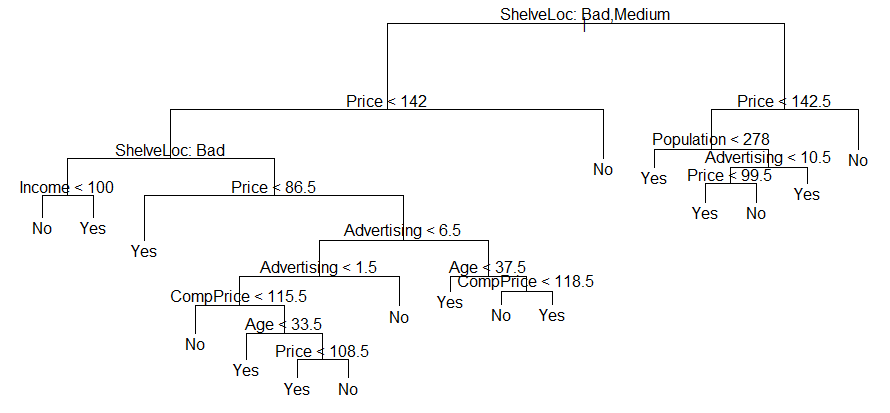
**Andy Lathrop | PREDICT 422 Programming Assignment 8: Tree Models**

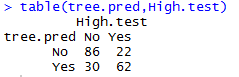
# ISLR Text: Lab Section 8.3: Decision Tree

Base code from textbook site:<http://www-bcf.usc.edu/~gareth/ISL/Chapter%208%20Lab.txt>

Type the commands yourself as you read through the lab.

In this lab we perform classification on high (response level = “Yes”) versus low sales volume (“No”) of child car seats, using a variety of predictor variables. The final, pruned tree following cross-validation is shown below. The prediction accuracy is 77%, an improvement from the initial error rate of 71.5% found with the baseline model using all data for the training set.

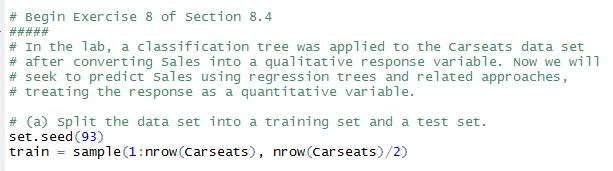
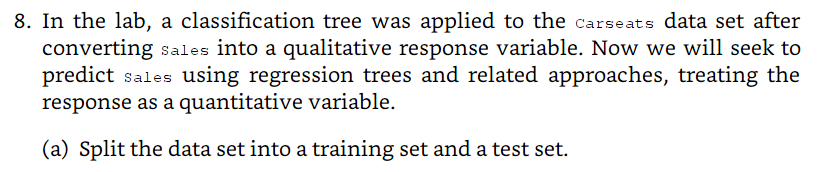




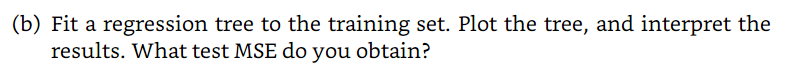
# ISLR Text: Exercise 8 of Section 8.4

Base code from: <https://github.com/asadoughi/stat-learning/blob/master/ch8/8.Rmd>

## Part a

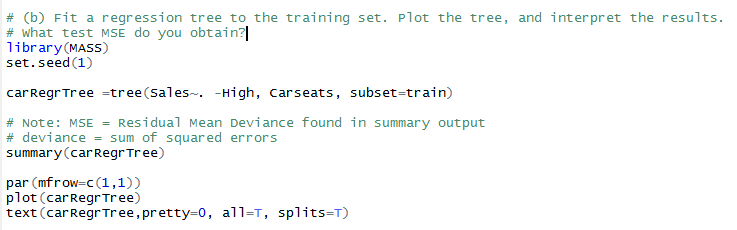


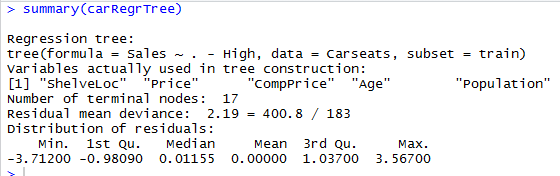
## Part b

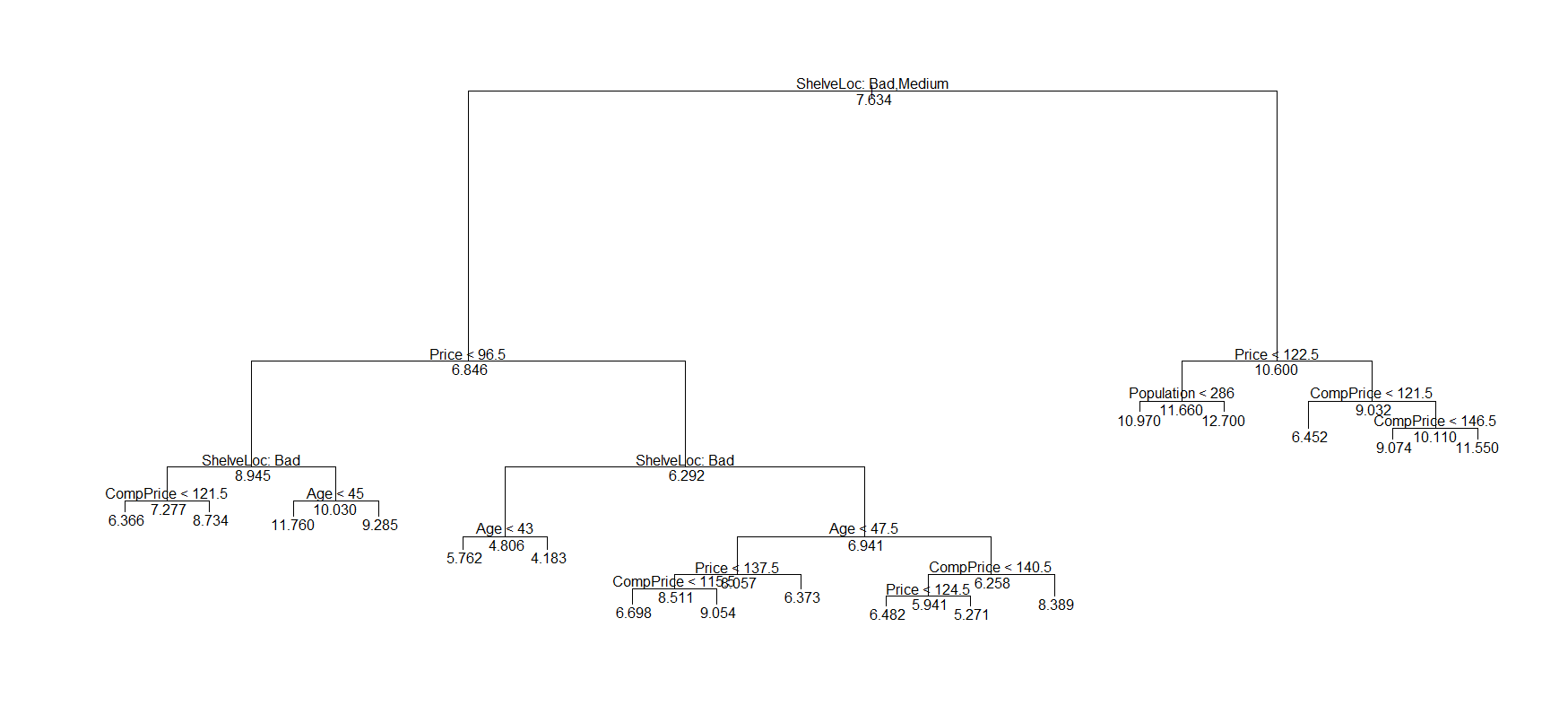


Interpretation: The variables used to create the tree are ShelveLoc, Price, CompPrice, Age, and Population. I’ve labeled all the nodes – so for example the value of 7.634 at the root (top) of the tree is the overall mean value for Sales (for the training set), prior to any splitting. ShelveLoc is the most important predictor of Sales, followed by price. The labeled value at each node is the mean response value prior to splitting at that node. The variable labels with their splitting criteria (i.e. ShelveLoc: Bad, Medium) indicate the data subset for the LEFT branch of the split. In the example of the root, the left branch is the data subset that includes ‘Bad’ and ‘Medium’ values for ShelveLoc .

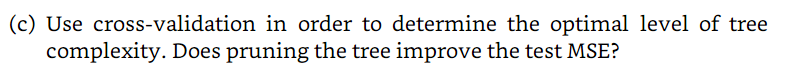
The MSE is the residual mean deviance from the summary output. For this model, the value is **2.19**





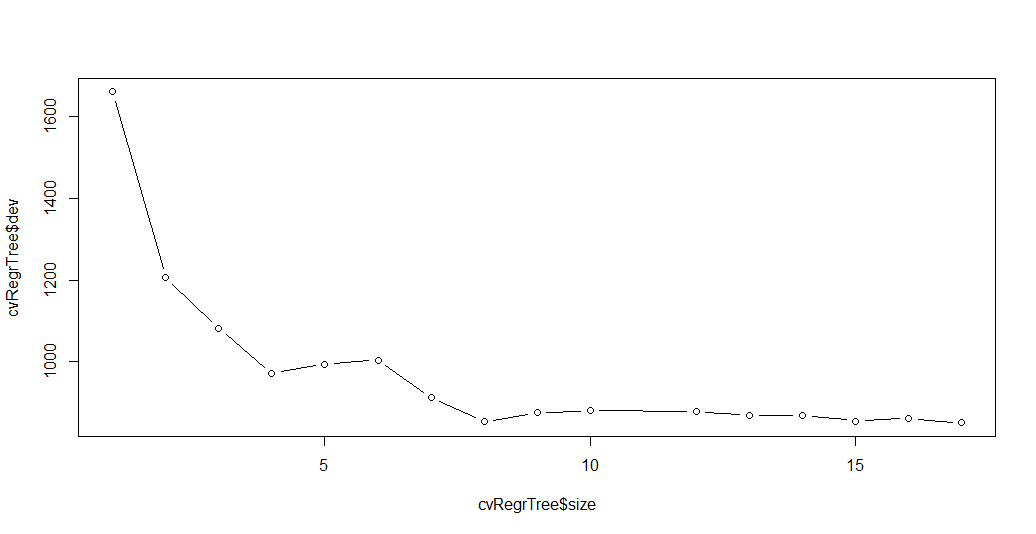


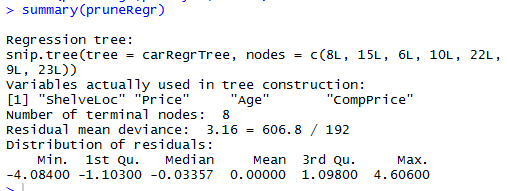
## Part c

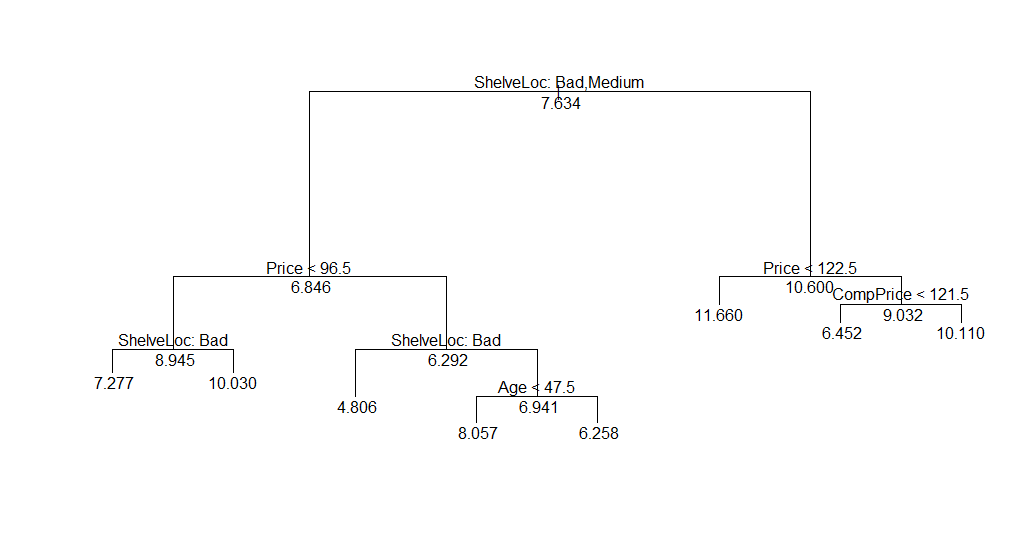


The cross validation shows that a simpler tree can be produced with 8 terminal nodes (compared to 17 in the original model), with only a slight increase in MSE (3.16 vs 2.19). This might be useful for explanatory purposes. However, the best MSE is obtained with the 17 node model.

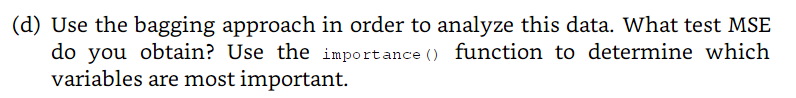




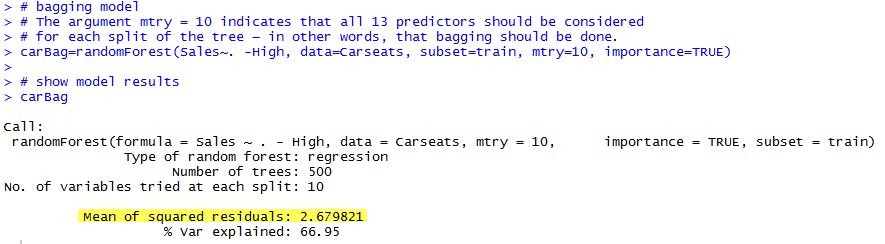


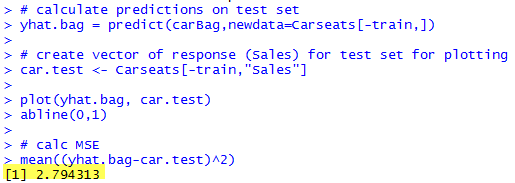


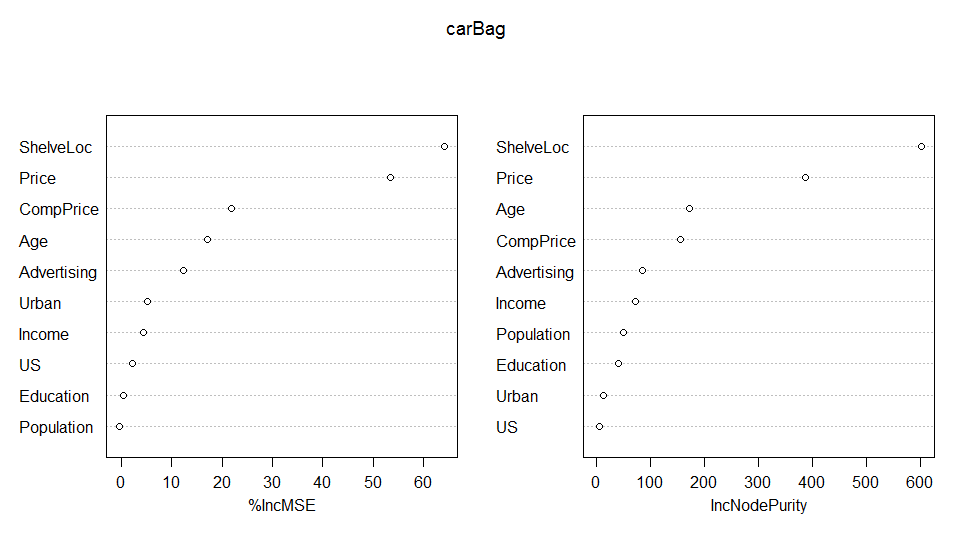
## Part d

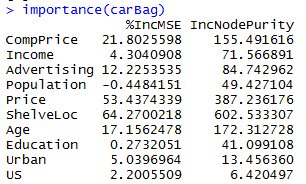


The training MSE I obtained was 2.68. The test MSE was 2.79. The 2 most important variables are ShelveLoc and Price. The remainder of the variables can be seen in the table and plot below.

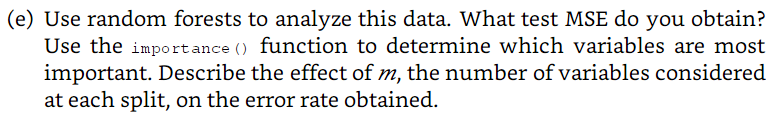








## Part e

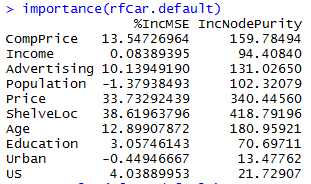
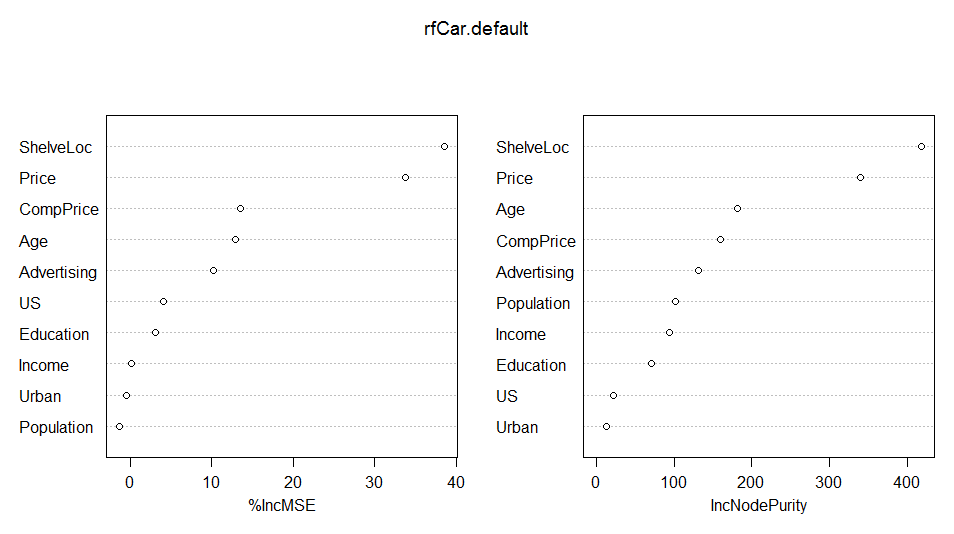


For my first random forest model, I did not specify a value for the mtry parameter, using the default p/3 (in this case, it used mtry=3). The training MSE was 3.01. Manually setting mtry to 4, the training MSE was 2.79. Using the tuneRF function, the best mtry value was 10 (m=p, the value used for bagging), with a training MSE of 2.32 (see plot below).

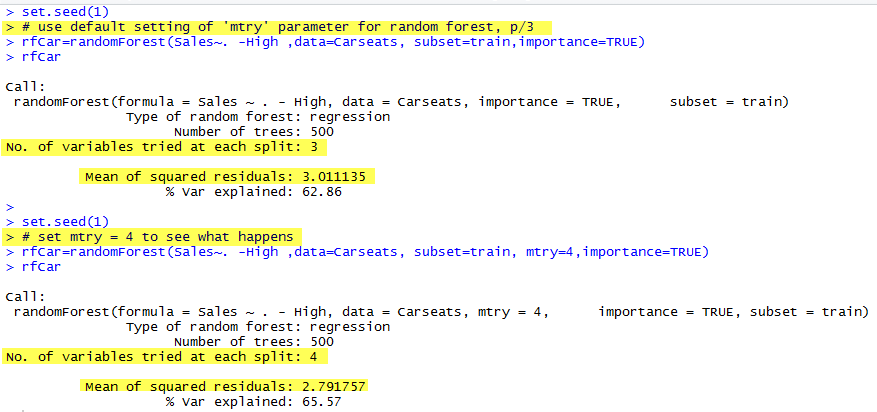
The test MSE for the default mtry value (mtry=3) was 3.21. ShelveLoc and Price were again the 2 most important variables.



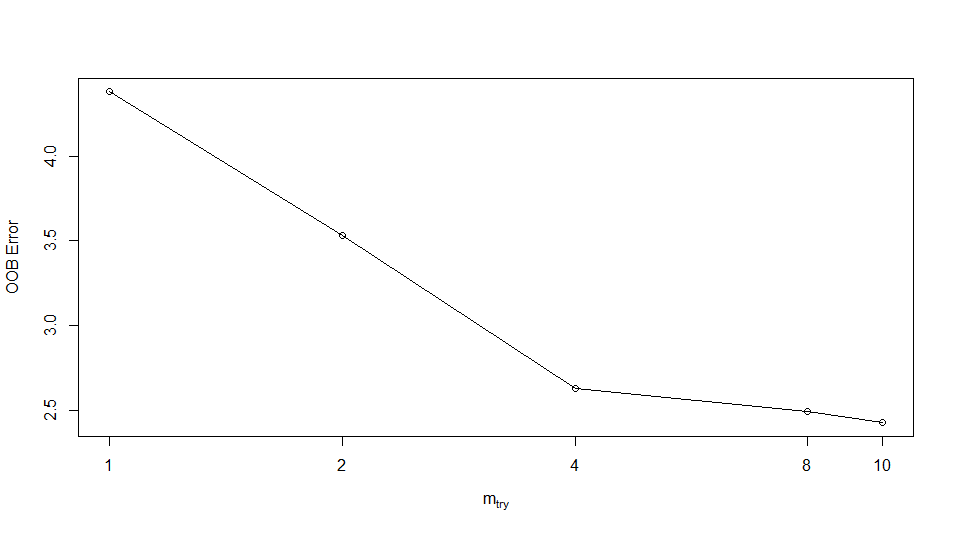
Variable importance

Output for testing various values of mtry:



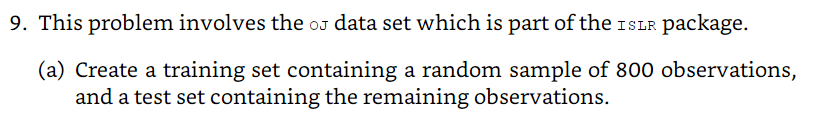
Plot of tuneRF function results to compare MSE vs mtry



# ISLR Text: Exercise 9 of Section 8.4

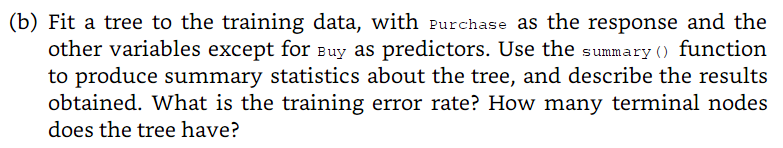
Base code from <https://github.com/asadoughi/stat-learning/blob/master/ch8/9.Rmd>

## Part a

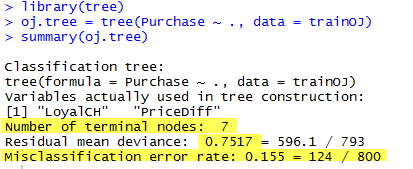




## Part b

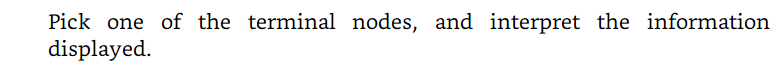


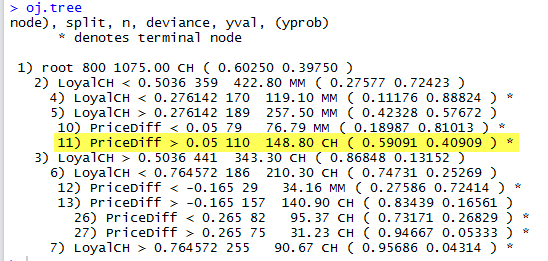
The tree uses variables LoyalCH and PriceDiff. The training error rate is 0.155. The number of terminal nodes is 7.



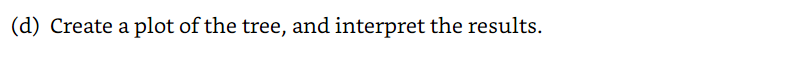
## Part c



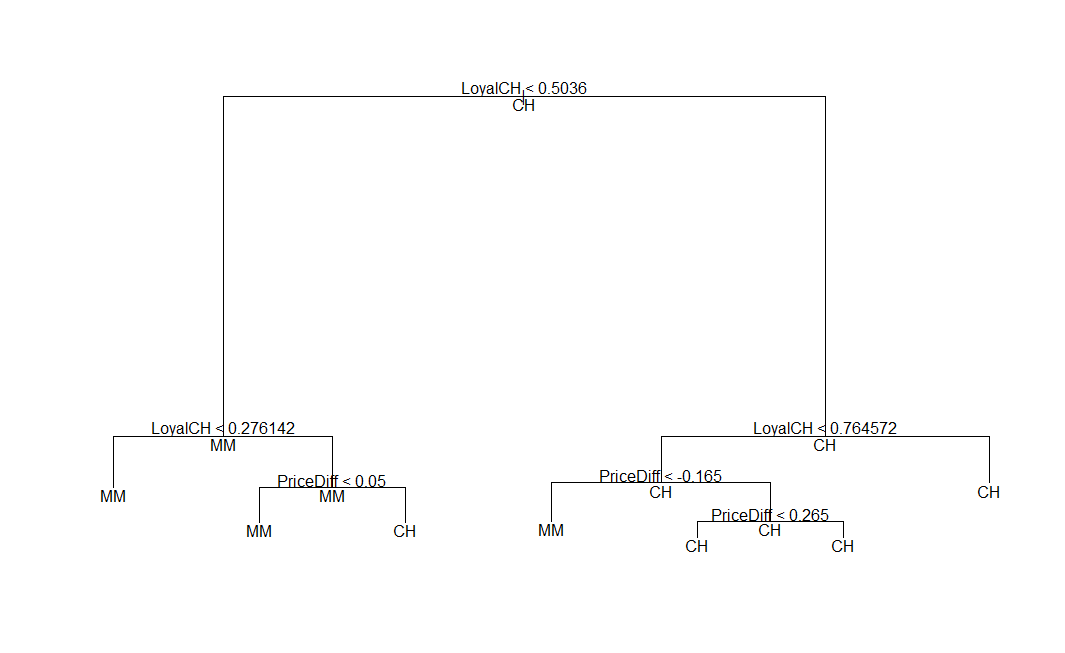
 For the terminal node labeled "11": The splitting variable at this node is PriceDiff > 0.05. The data definition of the variable PriceDiff is ‘Sale price of MM less sale price of CH’, meaning that this split occurs for prices where MM (Minute Maid) is greater than CH (Citrus Hill) by more than 5 cents. There are 110 points in the subtree below this node. The deviance (sum of squared errors) for all points contained in region below this node is 148.80. A \* in the line denotes that this is a terminal node. The prediction at this node is Sales = CH. About 60% points in this node have CH as the value of Sales}. The remaining 40% points have a value of MM.



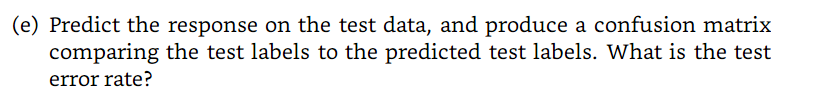
## Part d



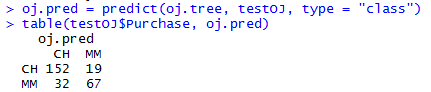
LoyalCH is the most important variable of the tree. It is used in the top 3 nodes. The data definition of LoyaltyCH is ‘Customer brand loyalty for CH,’ but it is not exactly clear how this value is measured. The breakpoints on the tree visualization, along with the detailed output, can be used to see how the purchase decision prediction can be determined. In this tree, the only 2 variables used are LoyalCH and PriceDiff. It is important to remember when following the tree structure that the LEFT branch follows the decision criteria indicated on the node. For example, the root node split criteria is LoyalCH **<** 0.5036. The LEFT branch shows the splitting that follows after this criteria is met. The RIGHT branch shows the splitting criteria when LoyalCH **>** 0.5036.



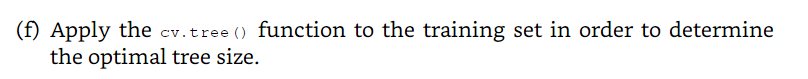
## Part e



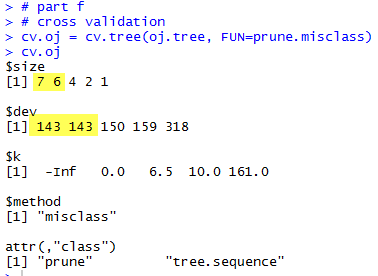
Confusion matrix below. Test error rate is (19+32)/270 = 18.9%



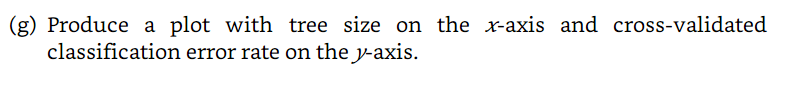
## Part f

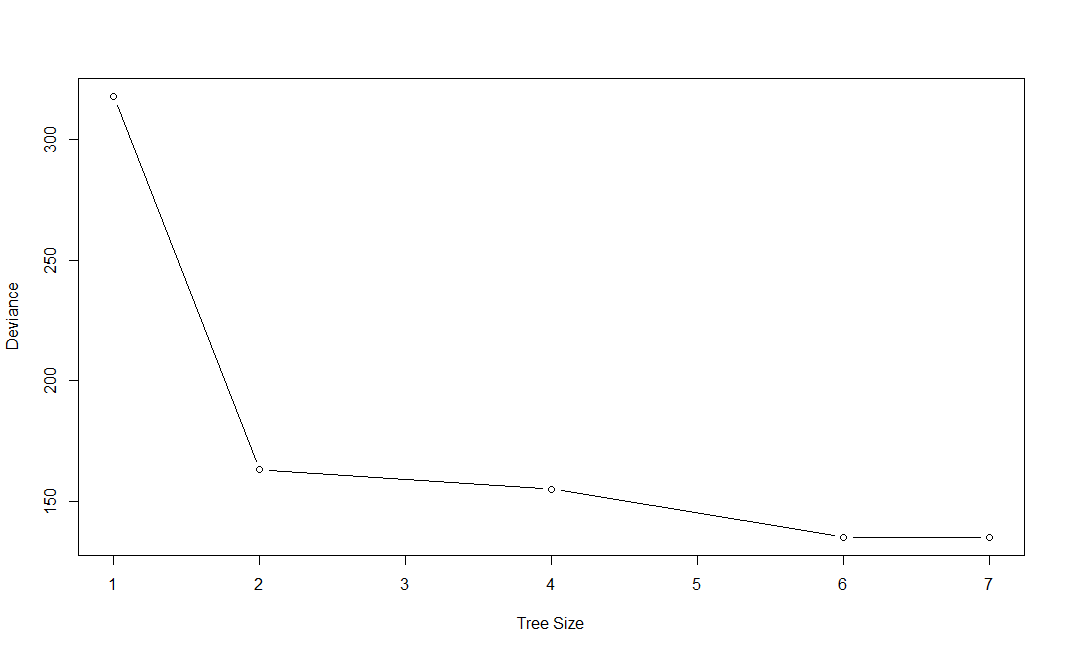


Tree size 7 and 6 both yield the same deviation as measured by misclassification error.

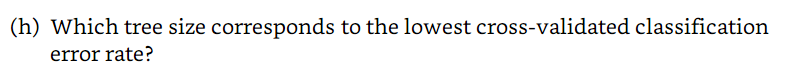


## Part g



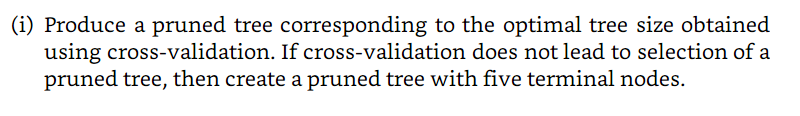


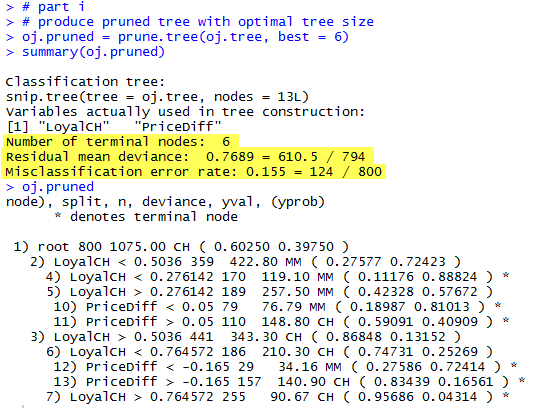
## Part h



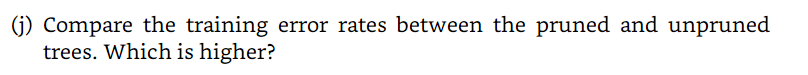
Tree size of 6 gives lowest CV error

## Part i



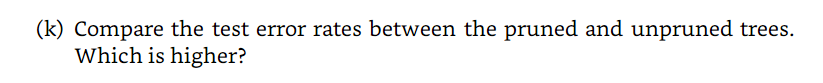


## Part j

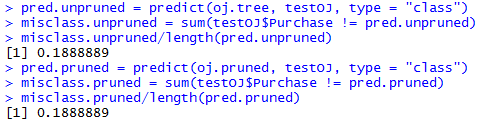


Unpruned = 0.155, Pruned = 0.155. They are the same

## Part k



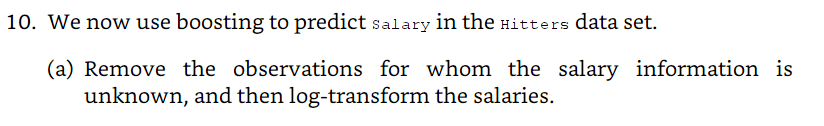
They are the same at 0.189

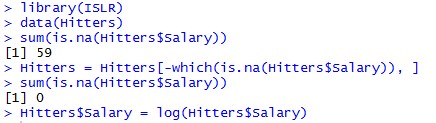


# ISLR Text: Exercise 10 of Section 8.4

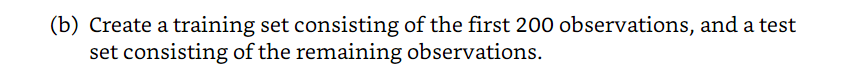
Base code from <https://github.com/asadoughi/stat-learning/blob/master/ch8/10.Rmd>

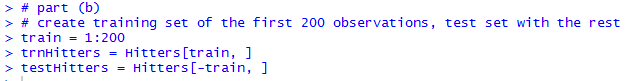
## Part a



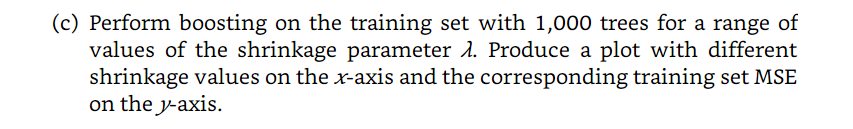


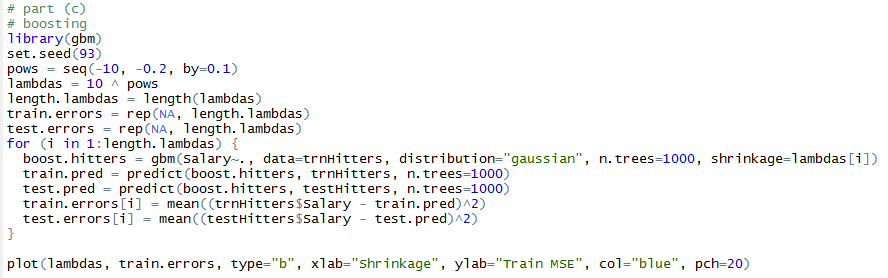
## Part b

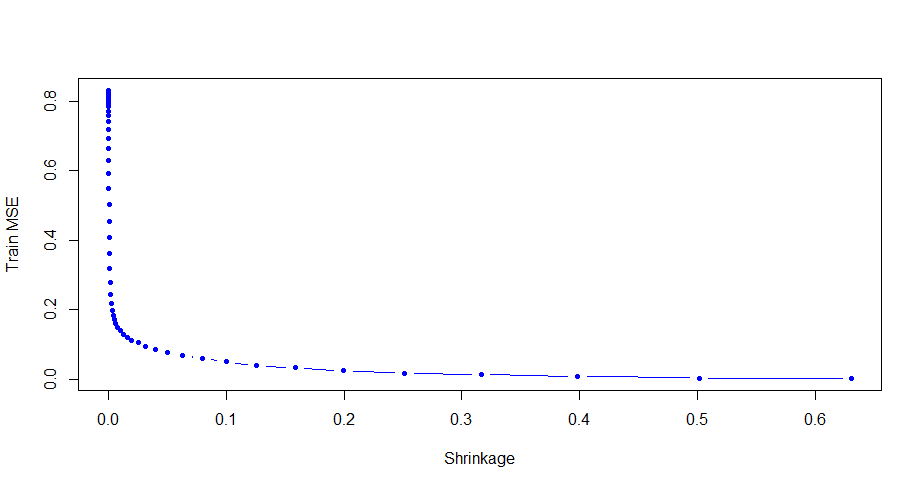




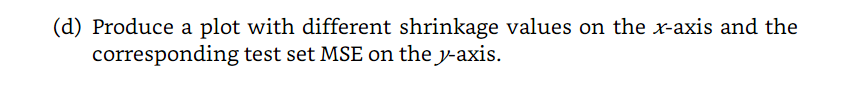
## Part c



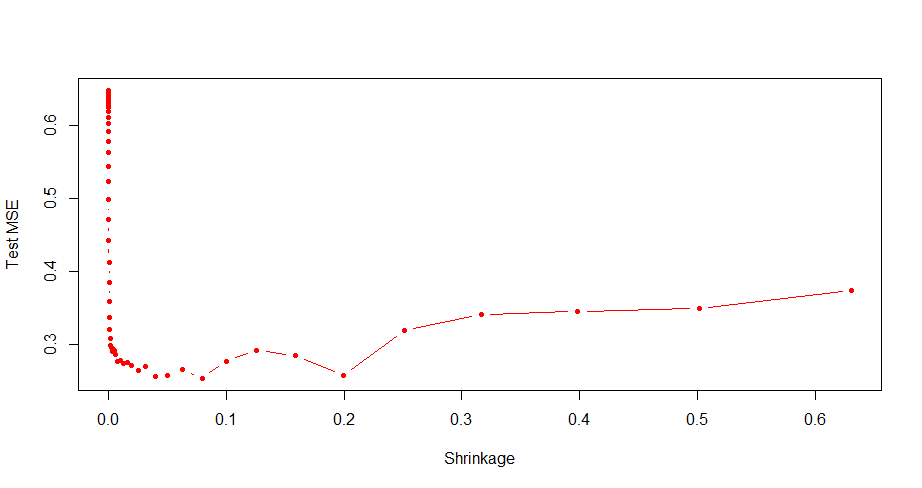
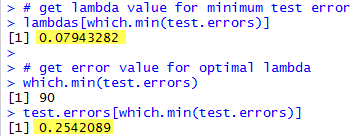




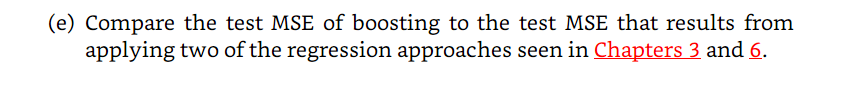
## Part d



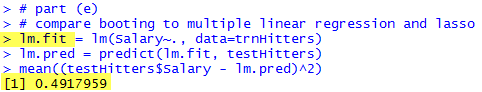
Lowest test MSE is 0.25, which occurs at lambda=0.079

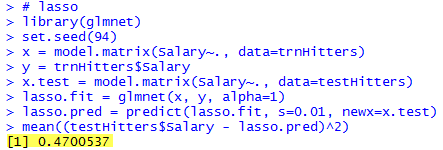


## Part e



Both multiple linear regression and lasso model have a higher test MSE



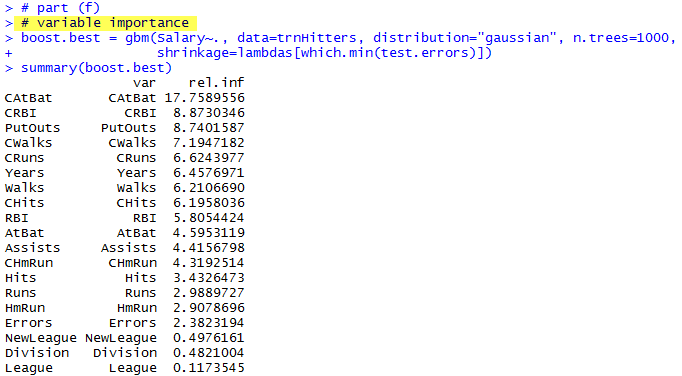


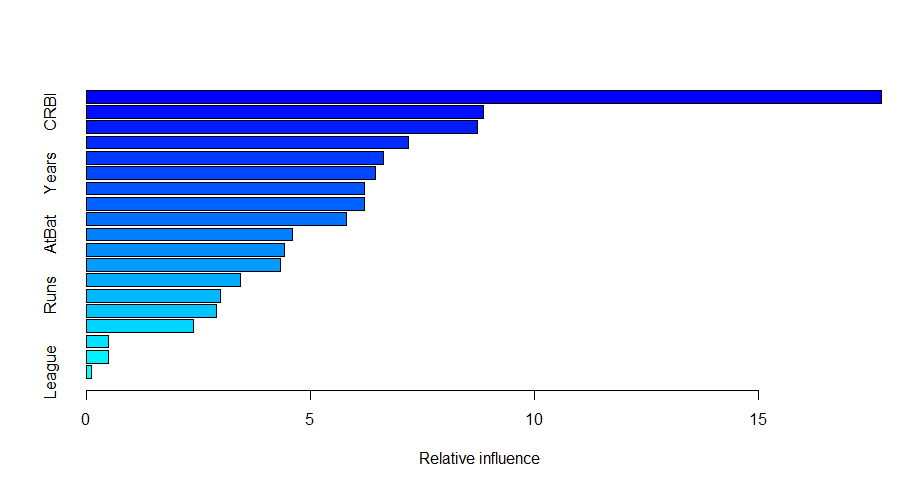
## Part f





3 most important variables are CAtBat, CRBI, and PutOuts.





## Part g

