9. This problem involves the OJ data set which is part of the ISLR2 package.

1. Create a training set containing a random sample of 800 observations, and a test set containing the remaining observations.



1. Fit a tree to the training data, with Purchase as the response and the other variables as predictors. Use the summary() function to produce summary statistics about the tree, and describe the results obtained. What is the training error rate? How many terminal nodes does the tree have?

一張含有 文字 的圖片

自動產生的描述

The tree have 9 terminal nodes and have the error rate at 15.88%. It only pick 5 variable as predictor.

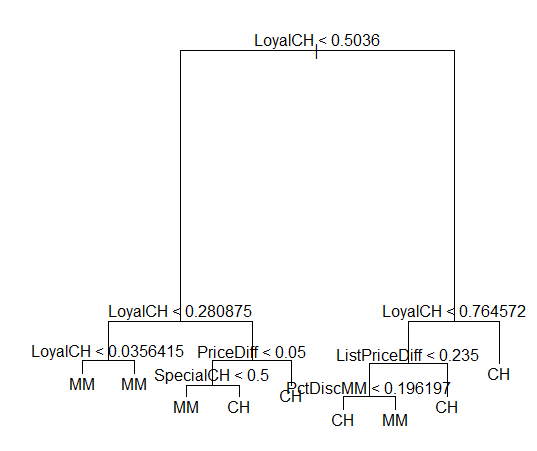
1. Type in the name of the tree object in order to get a detailed text output. Pick one of the terminal nodes, and interpret the information displayed.

One of the result is

13) ListPriceDiff > 0.235 102 65.43 CH ( 0.90196 0.09804 ) \*

The node no.13 which having the variable “ListPriceDiff” > 0.235. this branch have 102 observation, and deviance is 65.43. the prediction will be “CH” and the proportion of “CH” in these 102 observation is 90.196% otherwise, the proportion of “MM” is 9.804%.

1. Create a plot of the tree, and interpret the results.



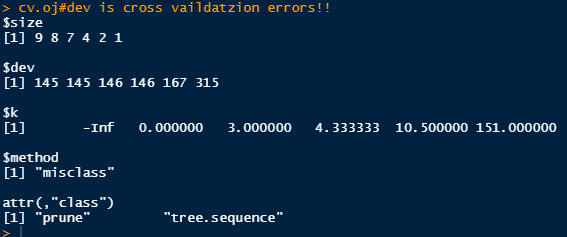
This tree have 9 terminal nodes the first variable that this tree consider is “LoyalCH”. For example the observation which “loyal CH” is greater than 0.76572 will be classify as “CH”, the one which “loyal CH” is less than 0.0356 will be classify as “MM”.

1. Predict the response on the test data, and produce a confusion matrix comparing the test labels to the predicted test labels. What is the test error rate?  
   一張含有 文字 的圖片

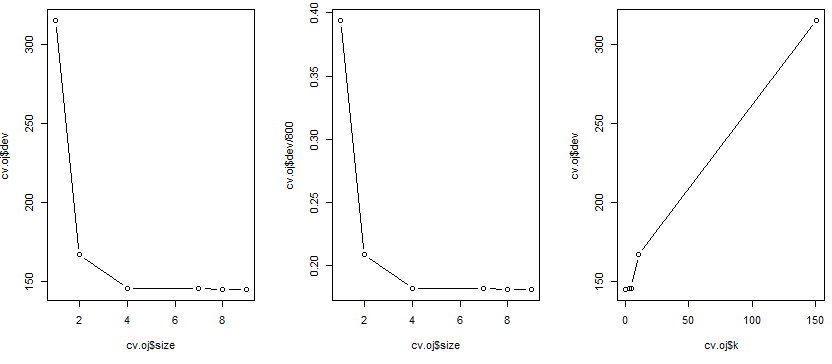
   自動產生的描述

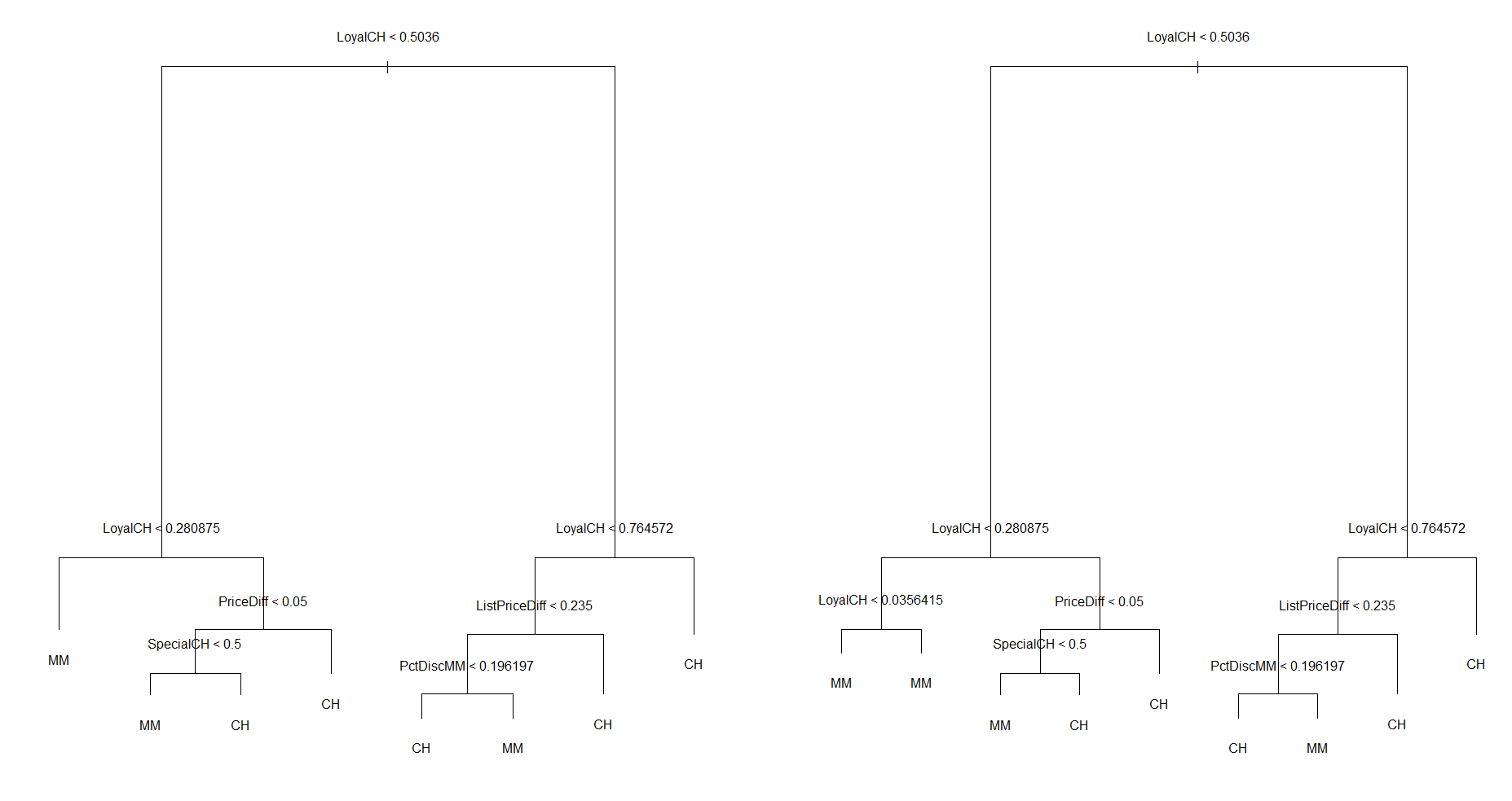
The confusion matrix is shown above, the error rate is (38+8)/270=17.03%

1. Apply the cv.tree() function to the training set in order to determine the optimal tree size.



According to the result above, the tree size 8 or 9 will have lowest misclassification rate.

1. Produce a plot with tree size on the x-axis and cross-validated classification error rate on the y-axis.   
   left: tree size and number of errors  
   middle: tree size and error rate.  
   right:by alpha and number of errors
2. Which tree size corresponds to the lowest cross-validated classification error rate?  
   By the plot above, the tree size 8 and 9 have lowest cross-validated classification error rate.
3. Produce a pruned tree corresponding to the optimal tree size obtained using cross-validation. If cross-validation does not lead to selection of a pruned tree, then create a pruned tree with five terminal nodes.



Left: tree size =8, right: tree size =9

Two trees are almost the same, although the tree which have nine terminal nodes have one more criterion of “loyalCH < 0.036”, but no matter it is True or False the prediction always be “MM”. Therefore this won’t effect the result.

1. Compare the training error rates between the pruned and unpruned trees. Which is higher?  
   一張含有 文字 的圖片

   自動產生的描述  
     
   Because the size of the tree is as same as the tree I create in part (a), the training error is still 15.88%.
2. Compare the test error rates between the pruned and unpruned trees. Which is higher?  
   一張含有 文字 的圖片

   自動產生的描述  
   Because the size of the tree is as same as the tree I create in part (a), the testing error is still (38+8)/270=17.03%.

We now use boosting to predict Salary in the Hitters data set.

1. Remove the observations for whom the salary information is unknown, and then log-transform the salaries.

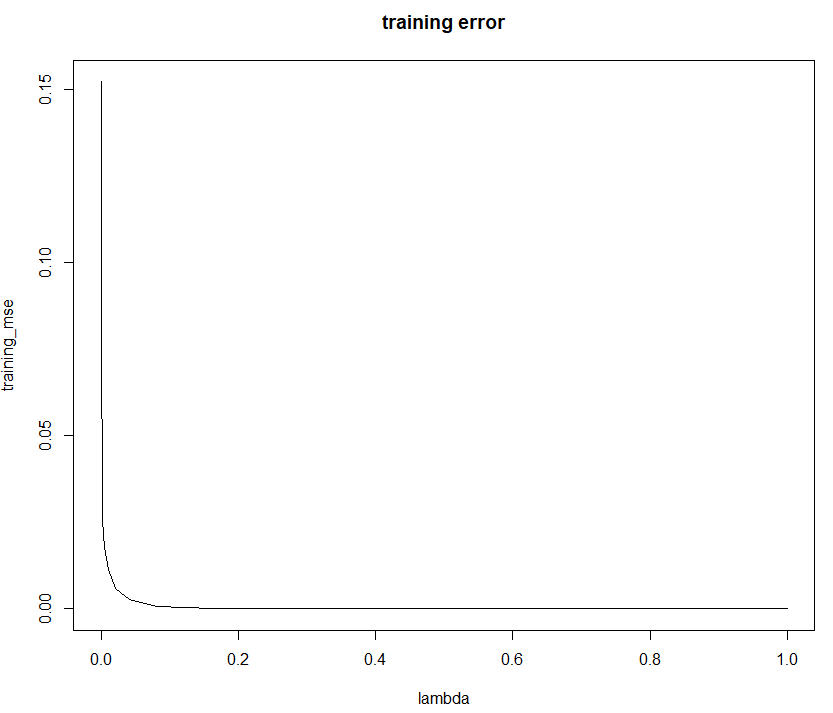
The new data have 263 rows, and the salaries was been log-transformed under the base 10.

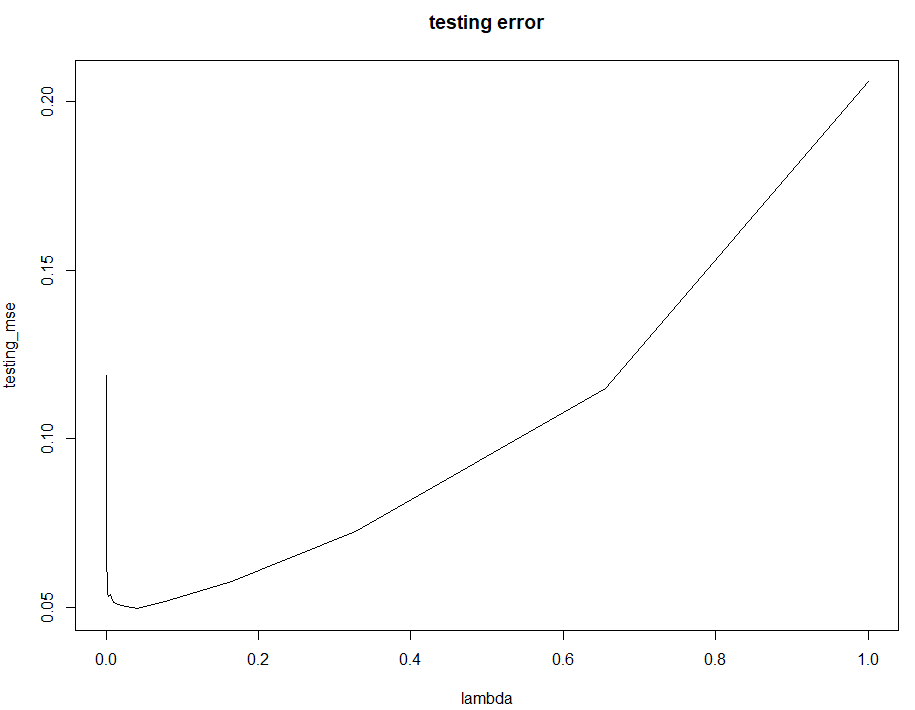
1. Create a training set consisting of the first 200 observations, and a test set consisting of the remaining observations.  
   The training set have 200 observations, and test set have 63 observations.
2. 一張含有 文字 的圖片

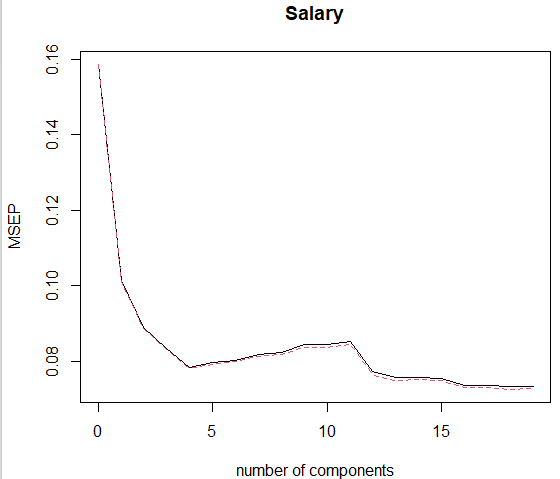
   自動產生的描述Perform boosting on the training set with 1,000 trees for a range of values of the shrinkage parameter λ. Produce a plot with different shrinkage values on the x-axis and the corresponding training set MSE on the y-axis.

I found that when λ is bigger than 1 the testing MSE will increase significantly

so I redo this section by smaller λ as followed, 

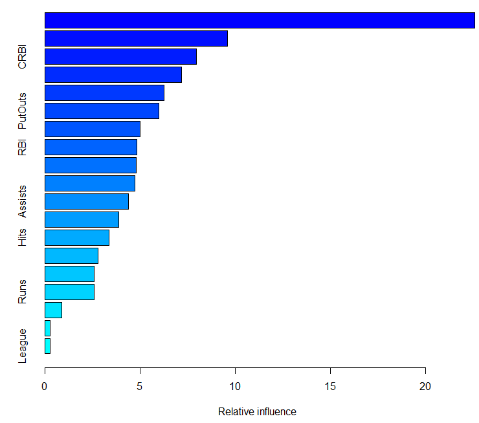
We can see that as λ increase the training MSE get smaller when λ=0.65536 the training MSE is smallest (2.907\*10^-15)

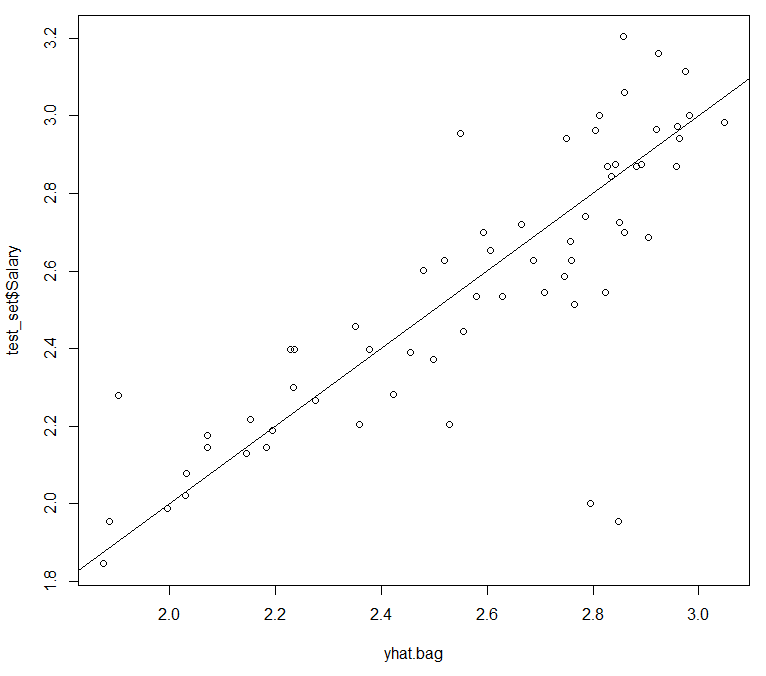
1. Produce a plot with different shrinkage values on the x-axis and the corresponding test set MSE on the y-axis.  
     
   Things become different when I calculate the testing MSE.  
   As λ increase the testing MSE also increase, and the lowest testing MSE happened in λ=0.00256, the Testing MSE is 0.05336.
2. Compare the test MSE of boosting to the test MSE that results from applying two of the regression approaches seen in Chapters 3 and 6.  
    I use the best subset method and the principle components regression(PCR) to predict.

The best subset method recommends to use 5, 8 and 13 predictor, according to what I choose for the model selection. The testing MSE is 0.094, 0.0884 and 0.094, when having 13, 8 and 5 predictors.

The PCR method have local minimum when M=4, and have global minimum when M=18.

When M=4, the MAPE is 0.0974, and when M=18, the MAPE is 0.09412

1. Which variables appear to be the most important predictors in the boosted model?   
   By using the default λ the most important predictor is “CAtBat”

1. Now apply bagging to the training set. What is the test set MSE for this approach?

The scatter plot of the real value and the predict value is shown in the left.

And the MSE 0.04343