Homework 4

Yanhao Li

Contents

```
library(tidyverse)
library(ISLR)
library(lasso2)
library(ISLR)
library(rpart)
library(rpart.plot)
library(ranger)
library(caret)
library(gbm)
```

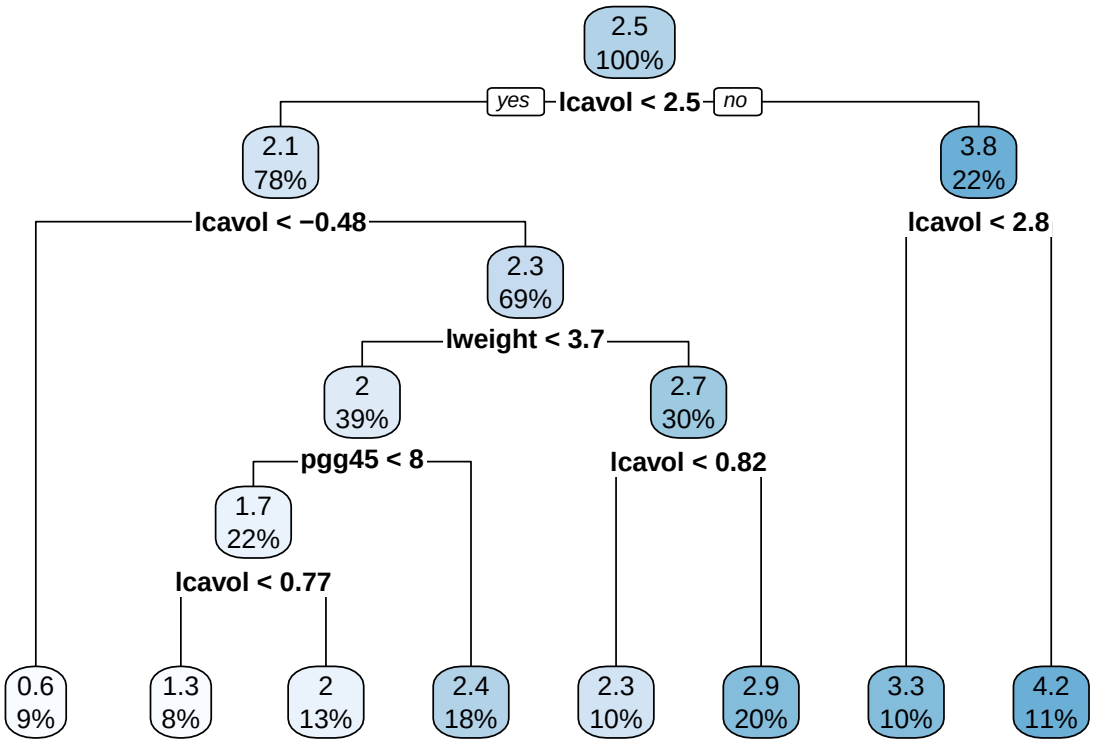
Question 1

Load, clean, and tidy data

```
data("Prostate")

prostate = Prostate %>%
  janitor::clean_names()
```

Question a

```
set.seed(1)

tree1 = rpart(formula = lpsa ~ .,
              data = prostate)

rpart.plot(tree1)
```

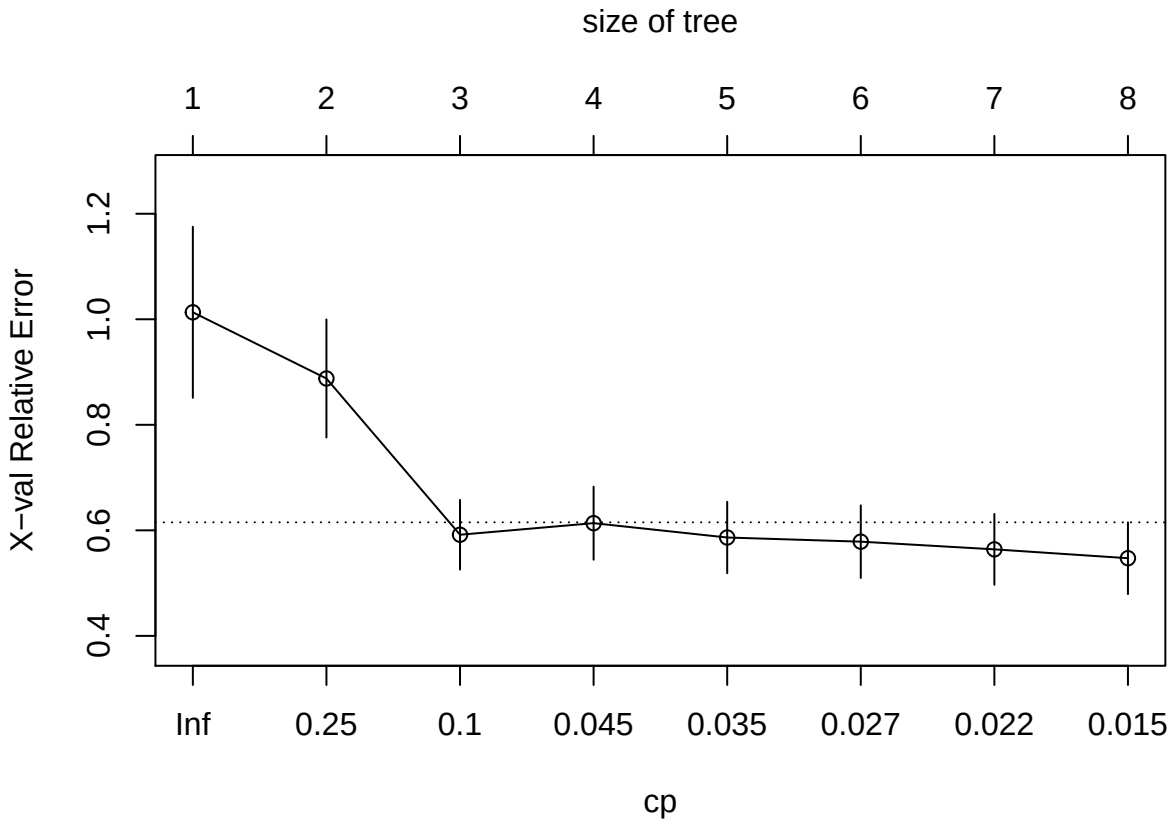



```
printcp(tree1)
```

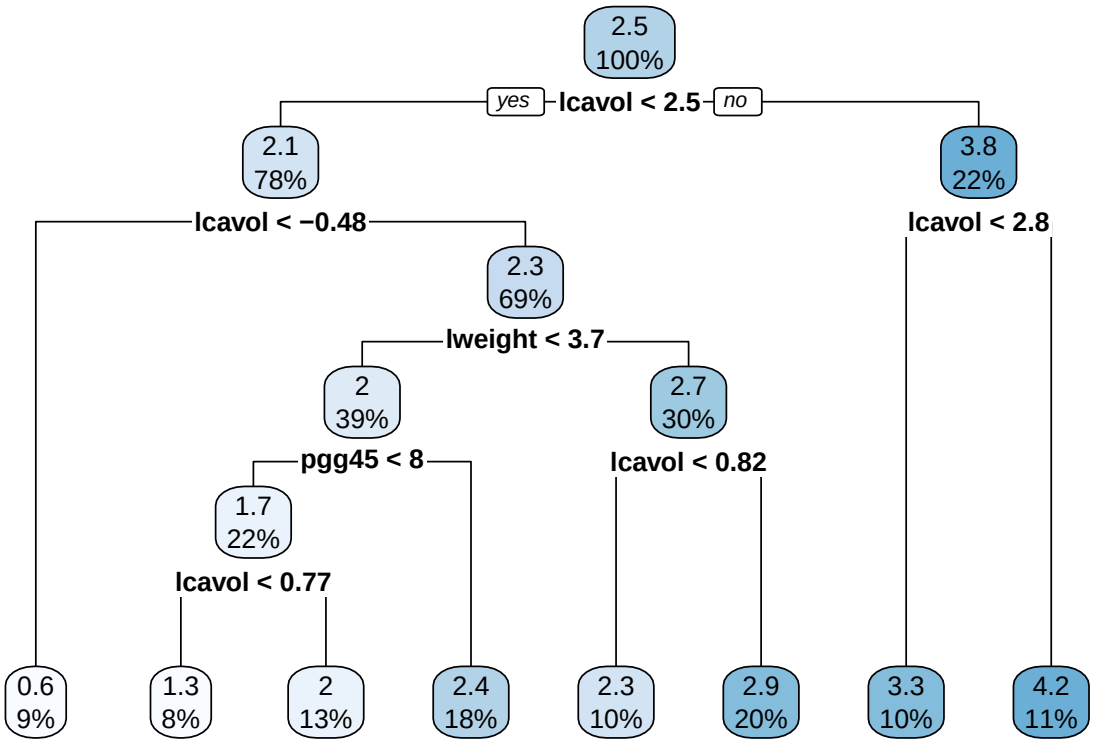
```
##
## Regression tree:
## rpart(formula = lpsa ~ ., data = prostate)
##
## Variables actually used in tree construction:
## [1] lcavol  lweight pgg45
##
## Root node error: 127.92/97 = 1.3187
##
## n= 97
##
##         CP nsplit rel error  xerror     xstd
## 1 0.347108      0   1.00000 1.01323 0.162162
## 2 0.184647      1   0.65289 0.88779 0.111915
## 3 0.059316      2   0.46824 0.59168 0.066102
## 4 0.034756      3   0.40893 0.61359 0.069269
## 5 0.034609      4   0.37417 0.58640 0.067630
## 6 0.021564      5   0.33956 0.57853 0.068772
## 7 0.021470      6   0.31800 0.56398 0.067155
## 8 0.010000      7   0.29653 0.54721 0.068034
```

```
cpTable <- tree1$cptable
```

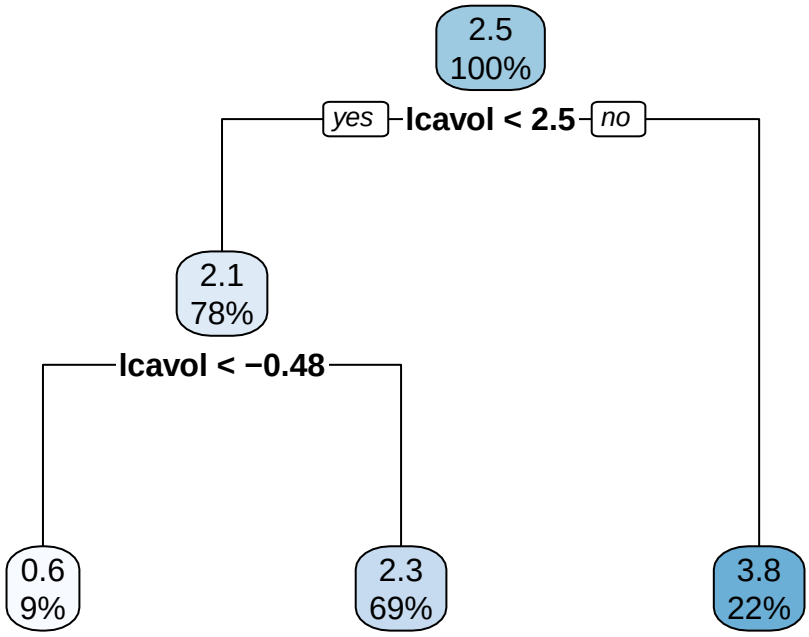
```r
plotcp(tree1)
```




```r
# minimum cross-validation error
minErr <- which.min(cpTable[,4])

tree2 <- prune(tree1, cp = cpTable[minErr,1])

rpart.plot(tree2)
```




```
# 1SE rule
tree3 <- prune(tree1, cp = cpTable[cpTable[,4] < cpTable[minErr,4] + cpTable[minErr,5],1][1])

rpart.plot(tree3)
```



Tree size corresponds to the lowest cross-validation error is 8. It is different from the tree size obtained using the 1 SE rule, which is 3.

Question b

Question c

Question d

Question e

Question f

Question 2

Load, clean, and tidy data

Question a

Question b

Question c