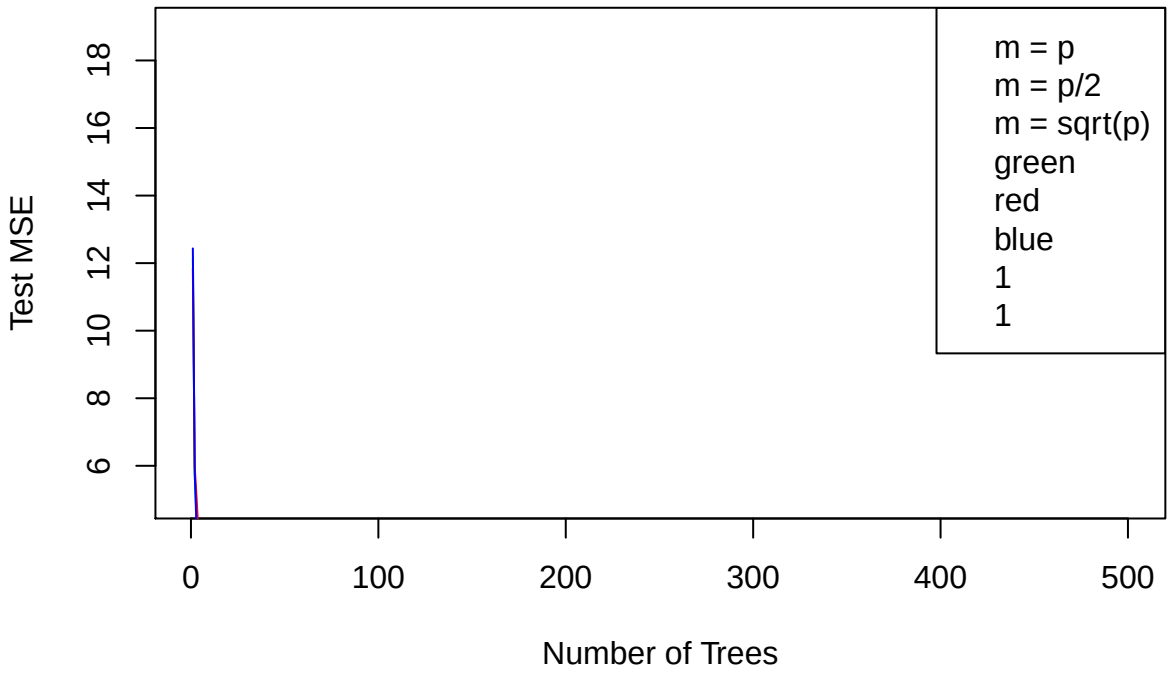
homework5

Collin

5/28/2021

#Question 1

```
data = Boston
train = sample(dim(data)[1],dim(data[1]/2))

x.train = data[train, -14]
x.test = data[train, -14]
y.train = data[train,14]
y.test = data[train,14]


p = dim(x.train)[2]
p.2 = p/2
p.sq = sqrt(p)


rf.p = randomForest(x.train, y.train,
                    xtest = x.test, ytest = y.test,
                    ntree = 500, ntry = p)


rf.p.2 = randomForest(x.train, y.train,
                    xtest = x.test, ytest = y.test,
                    ntree = 500, ntry = p.2)

rf.p.sq = randomForest(x.train, y.train,
                    xtest = x.test, ytest = y.test,
                    ntree = 500, ntry = p.sq)

#plot
plot(1:500, rf.p$test$mse, col = "green", type = "l",
     xlab = "Number of Trees", ylab = "Test MSE",
     ylim = c(5,19))
lines(1:500, rf.p.2$test$mse, col = "red", type = "l")
lines(1:500, rf.p.sq$test$mse, col = "blue", type = "l")
legend("topright", c("m = p", "m = p/2", "m = sqrt(p)", col = c("green", "red", "blue",cex = 1, lty = 1
```



The test MSE Looks super high for smaller number of trees, but as the number of trees increases, the test MSE decreases.

Question 8 (2)

a)

```
set.seed(10101)
train = sample(dim(Carseats)[1], dim(Carseats)[1]/2)
carseats_train = Carseats[train,]
carseats_test = Carseats[-train,]
```

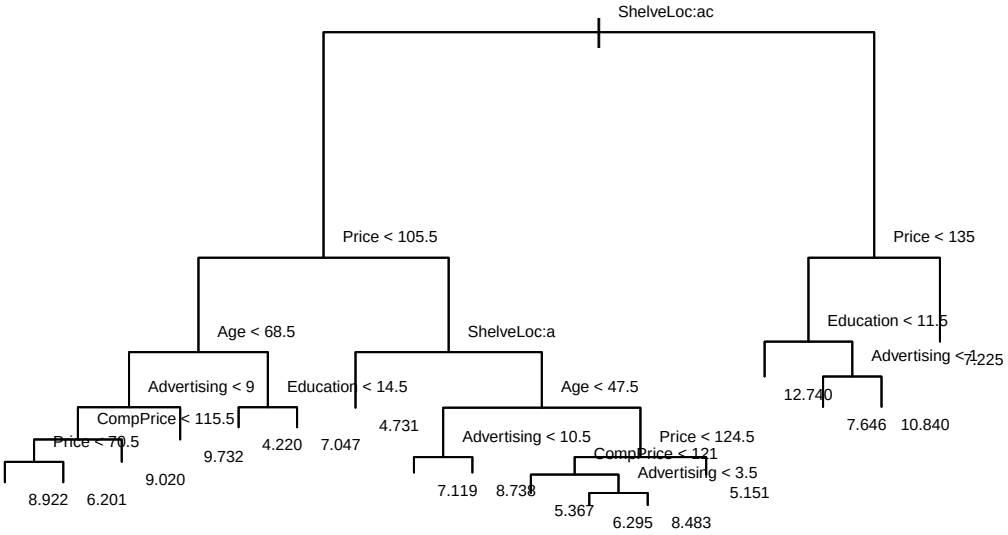
##b)

```
#Regression tree
tree_carseats = tree(Sales ~ ., data = carseats_train)


summary(tree_carseats)
```

```
##
## Regression tree:
## tree(formula = Sales ~ ., data = carseats_train)
## Variables actually used in tree construction:
## [1] "ShelveLoc"   "Price"       "Age"         "Advertising" "CompPrice"
## [6] "Education"
## Number of terminal nodes:  17
## Residual mean deviance:  2.027 = 370.8 / 183
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -3.64500 -0.94450 -0.08439  0.00000  0.91340  5.44300
```

```
#plot
plot(tree_carseats)
```

```
text(tree_carseats, pos = 4, cex = .5)
```

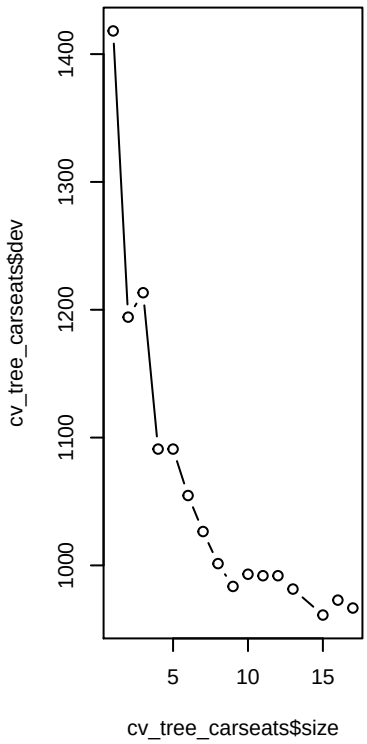


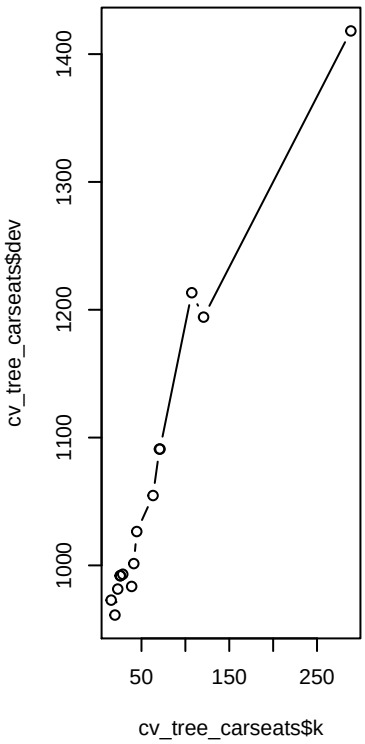

```
# OOS MSE

pred_carseats = predict(tree_carseats, newdata = carseats_test)
MSE = mean(  (carseats_test$Sales - pred_carseats  )^2   )
```

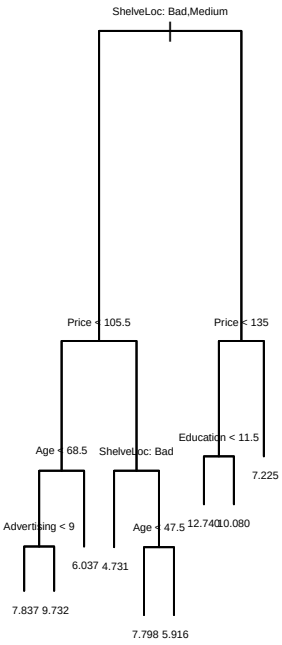
The MSE is 5.4043086

c)

```
#estimate cross-validated tree
cv_tree_carseats = cv.tree(tree_carseats, FUN = prune.tree)


#plot
par(mfrow =(c(1,3)))
plot(cv_tree_carseats$size,cv_tree_carseats$dev, type = "b")
plot(cv_tree_carseats$k,cv_tree_carseats$dev, type = "b")

#plot of pruned tree
pruned_carseats = prune.tree(tree_carseats, best = 9)
plot(pruned_carseats)
text(pruned_carseats,pretty = 0, cex = .5)
```







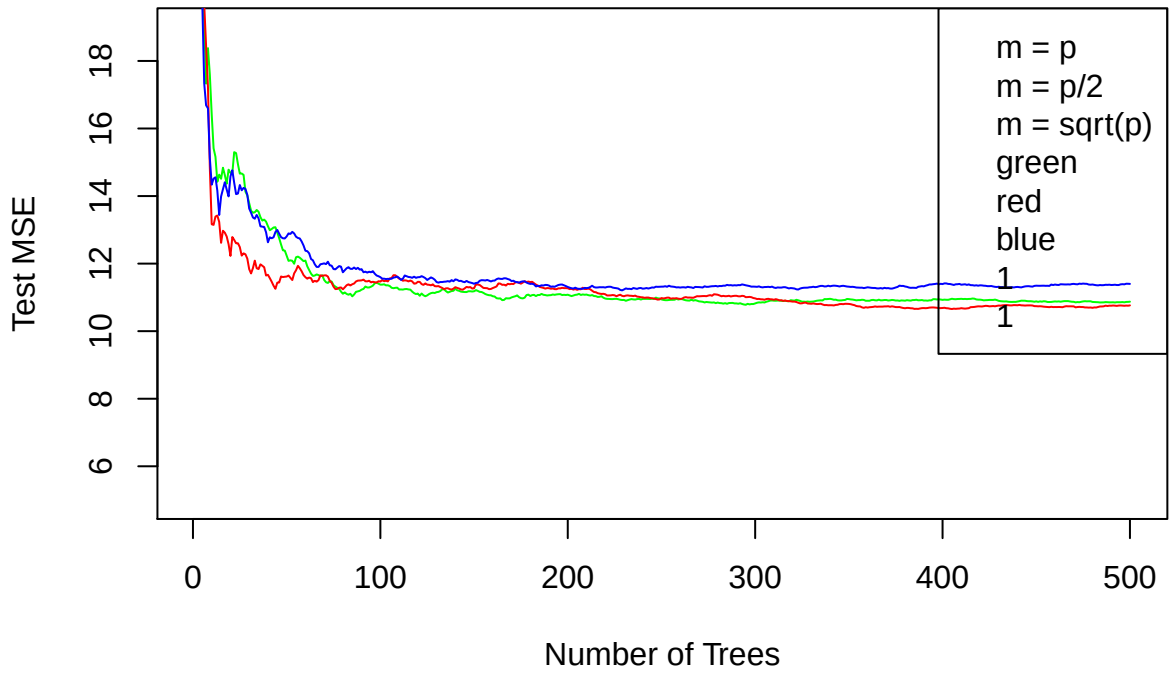

```
#OOS MSE

pred_pruned = predict(pruned_carseats, carseats_test)

MSE = mean( (carseats_test$Sales - pred_pruned)^2 )
```

the out of sample MSE is 4.8083061.

#Question 3 (Using CA SChools Data)

```
#predict math score
data = CASchools %>% drop_na()



train = sample(1:nrow(data), .5*(nrow(data)))
x.train = data[train, -14]
x.test = data[train, -14]
y.train = data[train,14]
y.test = data[train,14]



p = dim(x.train)[2]
p.2 = p/2
p.sq = sqrt(p)


rf.p = randomForest(x.train, y.train,
                    xtest = x.test, ytest = y.test,
```

```
                    ntree = 500, ntry = p)


rf.p.2 = randomForest(x.train, y.train,
                    xtest = x.test, ytest = y.test,
                    ntree = 500, ntry = p.2)

rf.p.sq = randomForest(x.train, y.train,
                    xtest = x.test, ytest = y.test,
                    ntree = 500, ntry = p.sq)

#plot
plot(1:500, rf.p$test$mse, col = "green", type = "l",
     xlab = "Number of Trees", ylab = "Test MSE",
     ylim = c(5,19))
lines(1:500, rf.p.2$test$mse, col = "red", type = "l")
lines(1:500, rf.p.sq$test$mse, col = "blue", type = "l")
legend("topright", c("m = p", "m = p/2", "m = sqrt(p)", col = c("green", "red", "blue",cex = 1, lty = 1
```



While the differences are marginal at best, the 3rd model, where the number of predictors that are sampled for splitting is the square root of the total number of predictors.

#Question #4

##a)

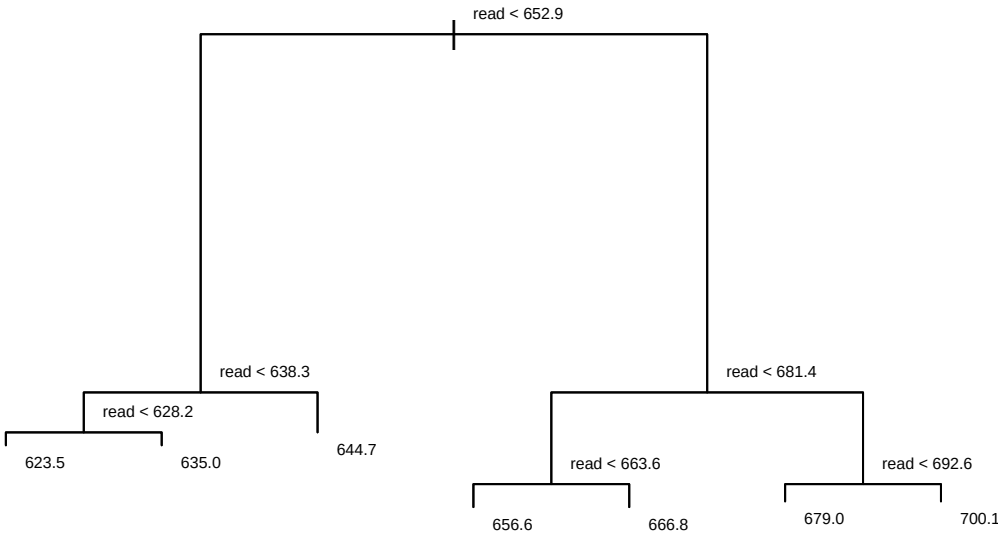
```
ca_schools = CASchools %>% drop_na()
set.seed(10101)
train = sample(dim(ca_schools)[1], dim(ca_schools)[1]/2) #split-half
ca_schools_train = ca_schools[train,]
ca_schools_test = ca_schools[-train,]
```

##b)

```
tree_ca_schools = tree(math ~ .- district - county - school, data = ca_schools_train)


summary(tree_ca_schools)
```

```
##
## Regression tree:
## tree(formula = math ~ . - district - county - school, data = ca_schools_train)
## Variables actually used in tree construction:
## [1] "read"
## Number of terminal nodes:  7
## Residual mean deviance:  50.57 = 10270 / 203
## Distribution of residuals:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -15.960  -4.930  -0.081   0.000   4.686  28.690
```

```
#plot
plot(tree_ca_schools)
text(tree_ca_schools, pos = 4, cex = .5)
```




```
# OOS MSE

pred_ca_schools = predict(tree_ca_schools, newdata = ca_schools_test)
MSE = mean(  (ca_schools_test$math - pred_ca_schools  )^2   )
```

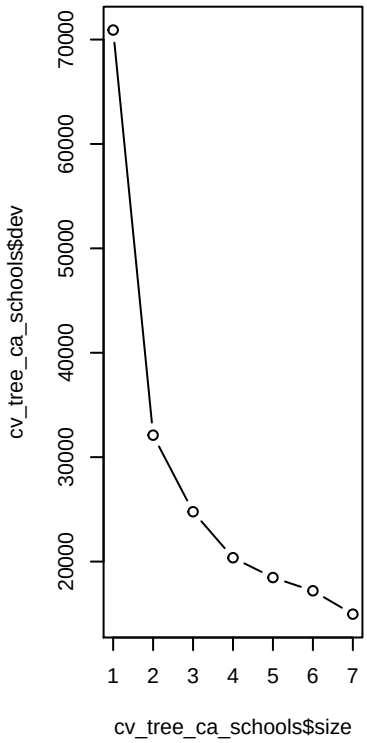
The out of sample MSE is 65.6557396

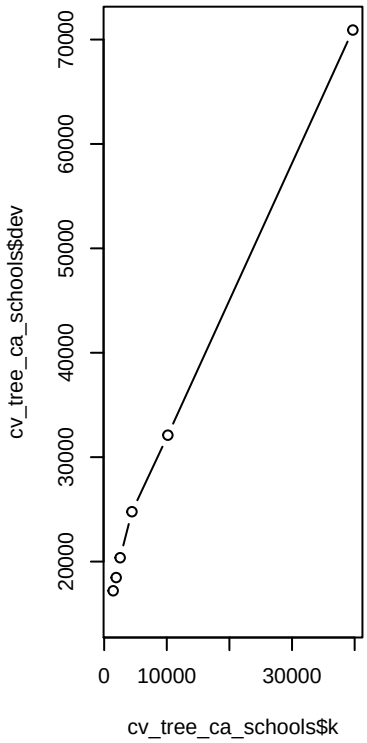
##c)

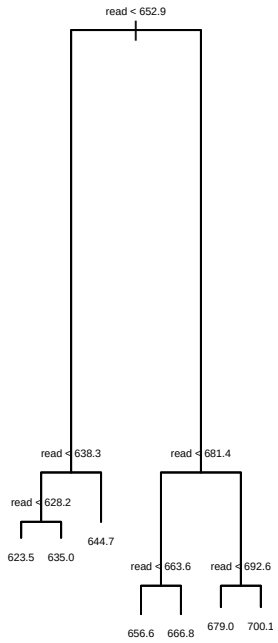
```
cv_tree_ca_schools = cv.tree(tree_ca_schools, FUN = prune.tree)


#plot
par(mfrow =(c(1,3)))
plot(cv_tree_ca_schools$size,cv_tree_ca_schools$dev, type = "b")
plot(cv_tree_ca_schools$k,cv_tree_ca_schools$dev, type = "b")

#plot of pruned tree
pruned_ca_schools = prune.tree(tree_ca_schools, best = 9)
```

```
## Warning in prune.tree(tree_ca_schools, best = 9): best is bigger than tree size
```

```
plot(pruned_ca_schools)
text(pruned_ca_schools,pretty = 0, cex = .5)
```








```
#OOS MSE

pred_pruned = predict(pruned_ca_schools, ca_schools_test)

MSE = mean( (ca_schools_test$math - pred_pruned)^2 )
```

The OOS MSE is 65.6557396.