Homework 6 - Predictive Modeling in Finance and Insurance

Dennis Goldenberg

2024-02-27

```
library(MASS)
library(leaps)
Boston$chas <- factor(Boston$chas)
```

1. Model Selection

a. Best Subset Selection

I perform the selection as intended:

```
bestSubset <- leaps::regsubsets(medv ~., data = Boston, method = "exhaustive",
                                nvmax = dim(Boston)[2] - 1)
summary(bestSubset)$outmat
```

```
##           crim zn  indus chas1 nox rm  age dis rad tax ptratio black lstat
## 1  ( 1 )  " "  " " " "   " "   " " " " " " " " " " " " "     " "   "*"
## 2  ( 1 )  " "  " " " "   " "   " " "*" " " " " " " " " "     " "   "*"
## 3  ( 1 )  " "  " " " "   " "   " " "*" " " " " " " " "*"     " "   "*"
## 4  ( 1 )  " "  " " " "   " "   " " "*" " " "*" " " " " "*"     " "   "*"
## 5  ( 1 )  " "  " " " "   " "   "*" "*" " " "*" " " " " "*"     " "   "*"
## 6  ( 1 )  " "  " " " "   "*"   "*" "*" " " "*" " " " " "*"     " "   "*"
## 7  ( 1 )  " "  " " " "   "*"   "*" "*" " " "*" " " " " "*"     "*"   "*"
## 8  ( 1 )  " "  "*" " "   "*"   "*" "*" " " "*" " " " " "*"     "*"   "*"
## 9  ( 1 )  "*"  " " " "   "*"   "*" "*" " " "*" "*" " " "*"     "*"   "*"
## 10  ( 1 ) "*"  "*" " "   " "   "*" "*" " " "*" "*" "*" "*"     "*"   "*"
## 11  ( 1 ) "*"  "*" " "   "*"   "*" "*" " " "*" "*" "*" "*"     "*"   "*"
## 12  ( 1 ) "*"  "*" "*"   "*"   "*" "*" " " "*" "*" "*" "*"     "*"   "*"
## 13  ( 1 ) "*"  "*" "*"   "*"   "*" "*" "*" "*" "*" "*" "*"     "*"   "*"
```

So , the first 6 variables that were selected were 1. lstat 2. rm 3. ptratio 4. dis 5. nox 6. chas. I show the c_p, BIC, and R^2 respectively for the first 6 models:

```
eStatdf<-data.frame(cbind(1:6, summary(bestSubset)$cp[1:6],
         summary(bestSubset)$bic[1:6],summary(bestSubset)$adjr2[1:6]))
colnames(eStatdf) <- c("Model #", "Cp", "BIC", "Adj. R Squared")
eStatdf
```

```
##   Model #        Cp       BIC Adj. R Squared
## 1       1 362.75295 -385.0521      0.5432418
## 2       2 185.64743 -496.2582      0.6371245
## 3       3 111.64889 -549.4767      0.6767036
## 4       4  91.48526 -561.9884      0.6878351
## 5       5  59.75364 -585.6823      0.7051702
## 6       6  47.17537 -592.9553      0.7123567
```

b. Forward and backward selection

I repeat the procedure for a, but doing forward and backward selection, and show the first 6 variables selected in each case in data frame format:

```
forSubset <- leaps::regsubsets(medv ~., data = Boston, method = "forward",
                               nvmax = dim(Boston)[2] - 1)
backSubset <- leaps::regsubsets(medv ~., data = Boston, method = "backward",
                               nvmax = dim(Boston)[2] - 1)
summary(forSubset)$outmat
```

```
##           crim zn  indus chas1 nox rm  age dis rad tax ptratio black lstat
## 1  ( 1 )  " "  " " " "   " "   " " " " " " " " " " " "     " "   "*"
## 2  ( 1 )  " "  " " " "   " "   " " "*" " " " " " " " " "     " "   "*"
## 3  ( 1 )  " "  " " " "   " "   " " "*" " " " " " " " " "*"     " "   "*"
## 4  ( 1 )  " "  " " " "   " "   " " "*" " " "*" " " " " "*"     " "   "*"
## 5  ( 1 )  " "  " " " "   " "   "*" "*" " " "*" " " " " "*"     " "   "*"
## 6  ( 1 )  " "  " " " "   "*"   "*" "*" " " "*" " " " " "*"     " "   "*"
## 7  ( 1 )  " "  " " " "   "*"   "*" "*" " " "*" " " " " "*"     "*"   "*"
## 8  ( 1 )  " "  "*" " "   "*"   "*" "*" " " "*" " " " " "*"     "*"   "*"
## 9  ( 1 )  "*"  "*" " "   "*"   "*" "*" " " "*" " " " " "*"     "*"   "*"
## 10  ( 1 ) "*"  "*" " "   "*"   "*" "*" " " "*" "*" " " "*"     "*"   "*"
## 11  ( 1 ) "*"  "*" " "   "*"   "*" "*" " " "*" "*" "*" "*"     "*"   "*"
## 12  ( 1 ) "*"  "*" "*"   "*"   "*" "*" " " "*" "*" "*" "*"     "*"   "*"
## 13  ( 1 ) "*"  "*" "*"   "*"   "*" "*" "*" "*" "*" "*" "*"     "*"   "*"
```

```
summary(backSubset)$outmat
```

```
##           crim zn  indus chas1 nox rm  age dis rad tax ptratio black lstat
## 1  ( 1 )  " "  " " " "   " "   " " " " " " " " " " " "     " "   "*"
## 2  ( 1 )  " "  " " " "   " "   " " "*" " " " " " " " " "     " "   "*"
## 3  ( 1 )  " "  " " " "   " "   " " "*" " " " " " " " " "*"     " "   "*"
## 4  ( 1 )  " "  " " " "   " "   " " "*" " " "*" " " " " "*"     " "   "*"
## 5  ( 1 )  " "  " " " "   " "   "*" "*" " " "*" " " " " "*"     " "   "*"
## 6  ( 1 )  " "  " " " "   " "   "*" "*" " " "*" " " " " "*"     "*"   "*"
## 7  ( 1 )  " "  " " " "   " "   "*" "*" " " "*" "*" " " "*"     "*"   "*"
## 8  ( 1 )  "*"  " " " "   " "   "*" "*" " " "*" "*" " " "*"     "*"   "*"
## 9  ( 1 )  "*"  " " " "   " "   "*" "*" " " "*" "*" "*" "*"     "*"   "*"
## 10  ( 1 ) "*"  "*" " "   " "   "*" "*" " " "*" "*" "*" "*"     "*"   "*"
## 11  ( 1 ) "*"  "*" " "   "*"   "*" "*" " " "*" "*" "*" "*"     "*"   "*"
## 12  ( 1 ) "*"  "*" "*"   "*"   "*" "*" " " "*" "*" "*" "*"     "*"   "*"
## 13  ( 1 ) "*"  "*" "*"   "*"   "*" "*" "*" "*" "*" "*" "*"     "*"   "*"
```

```
featSelect<-data.frame(1:6,cbind(c("lstat", "rm", "ptratio","dis","nox","chas"),
                                 c("lstat", "rm", "ptratio","dis","nox","black")))
colnames(featSelect) <- c("Model Number", "Var. forward", "Var. backward")
featSelect
```

```
##   Model Number Var. forward Var. backward
## 1            1        lstat         lstat
## 2            2           rm            rm
## 3            3      ptratio       ptratio
## 4            4          dis           dis
## 5            5          nox           nox
## 6            6         chas         black
```

c. Comparing Variable selections

The best Subset selection and forward selection algorithms selected the same 6 variables, and in the same order. The backward selection algorithm matched the other two up until model 6, where the 6th variable selected was black as opposed to chas.

```
BestFowModel <- lm("medv ~ lstat + rm + ptratio + dis + nox + chas",
                   data = Boston)
backModel <- lm("medv ~ lstat + rm + ptratio + dis + nox + black",
                   data = Boston)
print("Coefficients for Best Subset and forward model:")
```

```
## [1] "Coefficients for Best Subset and forward model:"
```

```
summary(BestFowModel)$coefficients
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  36.9226340 4.55908556   8.098693 4.291836e-15
## lstat        -0.5698442 0.04744883 -12.009657 2.305468e-29
## rm            4.1118117 0.40721667  10.097356 6.144302e-22
## ptratio      -1.0027463 0.11273664  -8.894591 1.078984e-17
## dis          -1.1445857 0.16671617  -6.865475 1.975595e-11
## nox         -18.7404327 3.22732486  -5.806801 1.134454e-08
## chas1         3.2443048 0.88324944   3.673147 2.654731e-04
```

```
print("Coefficients for Backward Model:")
```

```
## [1] "Coefficients for Backward Model:"
```

```
summary(backModel)$coefficients
```

```
##                 Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)  30.516970426 4.959607224   6.153102 1.560882e-09
## lstat        -0.545496912 0.048414974 -11.267111 2.165763e-26
## rm            4.354807129 0.410753352  10.602000 8.019446e-24
## ptratio      -1.012059411 0.112597327  -8.988308 5.194370e-18
## dis          -1.159602736 0.166618639  -6.959622 1.077921e-11
## nox         -15.842368174 3.278907022  -4.831600 1.805153e-06
## black         0.009577916 0.002677202   3.577584 3.806043e-04
```

The variables for the first 5 sig