week2 HW

ESC Sejung Kim

2021.7.26

1) forward stepwise selection

```
library(mlbench)

## Warning: package 'mlbench' was built under R version 4.0.5

data(BostonHousing)
sum(is.na(BostonHousing))

## [1] 0
```

# 결측치 개수 측정해보니 0 : 따로 처리해야 할 결측치 X

```
head(BostonHousing)

##      crim zn indus chas   nox    rm  age    dis rad tax ptratio      b lst
at
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.
98
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.
14
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.
03
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.
94
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.
33
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.
21
##   medv
## 1 24.0
## 2 21.6
## 3 34.7
## 4 33.4
## 5 36.2
## 6 28.7
```

미리 ?BostonHousing 함수를 통해 데이터에 관한 정보를 확인했고, 그 결과 medv 가 target variable 이라는 사실을 확인할 수 있었다 !

```
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 4.0.5
```

```
regfit.fwd = regsubsets(medv~., data = BostonHousing, nvmax=13, method = "for
ward")
summary(regfit.fwd)
```

```
## Subset selection object
## Call: regsubsets.formula(medv ~ ., data = BostonHousing, nvmax = 13, 
##     method = "forward")
## 13 Variables  (and intercept)
##         Forced in Forced out
## crim        FALSE      FALSE
## zn          FALSE      FALSE
## indus       FALSE      FALSE
## chas1       FALSE      FALSE
## nox         FALSE      FALSE
## rm          FALSE      FALSE
## age         FALSE      FALSE
## dis         FALSE      FALSE
## rad         FALSE      FALSE
## tax         FALSE      FALSE
## ptratio     FALSE      FALSE
## b           FALSE      FALSE
## lstat       FALSE      FALSE
## 1 subsets of each size up to 13
## Selection Algorithm: forward
##           crim zn  indus chas1 nox rm  age dis rad tax ptratio b   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 ) "*"  "*" "*"   "*"   "*" "*" "*" "*" "*" "*" "*"     "*" "*"  
```

dis 변수가 새롭게 선택된 것 확인.

```
reg.summary = summary(regfit.fwd)
coef(regfit.fwd,4)
```

```
## (Intercept)          rm         dis     ptratio       lstat 
##  24.4713576   4.2237922  -0.5519263  -0.9736458  -0.6654360
```

```
reg.summary$rsq[4]
```

```
## [1] 0.6903077
```

dis 변수의 추정 계수는 -0.5519263 이고, 이 모델의 R squared 는 약 0.69 !

#2. 7.5 of ESL

$W = E_y[Err_{in}] - E_y[\overline{err}]$

$= E_y[\frac{1}{N}\sum_{i=1}^{N} E_{y^0}[L(y_i^0, \hat{y}_i)]] - E_y[\frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{y}_i)]$

$= E_y[\frac{1}{N}\sum_{i=1}^{N} E_{y^0}[(y_i^0 - \hat{y}_i)^2]] - E_y[\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2]$ because loss function ⟶ squared error

$= E_y[\frac{1}{N}\sum_{i=1}^{N}(E_{y^0}[(y_i^0)^2] + \hat{y}_i^2 - 2E_{y^0}[y_i^0]\hat{y}_i] - E_y[\frac{1}{N}\sum_{i=1}^{N}(y_i^2 + \hat{y}_i^2 - 2y_i\hat{y}_i)$

여기서, $E_{y^0}[(y_i^0)^2] = E_y((y_i)^2)$ & $E_{y^0}[y_i^0] = E_y(y_i)$ because X is same ! (fixed)

$= \underset{\sim}{E_y}[\frac{1}{N}\sum_{i=1}^{N}(E_y(y_i^2) + \hat{y}_i^2 - 2E_y(y_i)\hat{y}_i - y_i^2 - \hat{y}_i^2 + 2y_i\hat{y}_i)$

$= \frac{2}{N}\Sigma \mathrm{cov}(\hat{y}_i, y_i)$

$\therefore W = \frac{2}{N}\Sigma \mathrm{cov}(\hat{y}_i, y_i)$