ESC 21-SUMMER week 2_Homework

W.J.Park

문제 1.

```r
install.packages("mlbench")
library(mlbench)
library(leaps)

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

m <- regsubsets(medv ~., data = BostonHousing, method = "forward")
summary(m)
coef(m, 4)
summary(m)$adjr2[4]


# 1. 단계별로 생성된 모델들 중 예측변수(또는 설명변수)가 4개인 모델에서, 새롭게 선택된 변수의 이름은 무엇인가요? dis
# 2. 그 변수의 추정된 계수의 값과 해당 모델의 결정계수 R^2는 무엇인가요? 0.6878351
```

```
ptratio      FALSE      FALSE
b            FALSE      FALSE
lstat        FALSE      FALSE
1 subsets of each size up to 8
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 ) " "  "*" " "   "*"   "*" "*" " " "*" " " " " "*"     "*" "*"
> coef(m, 4)
(Intercept)          rm         dis     ptratio       lstat 
 24.4713576   4.2237922  -0.5519263  -0.9736458  -0.6654360 
> summary(m)$adjr2[4]
[1] 0.6878351
```

문제 2.

Ex. 7.4 Consider the in-sample prediction error (7.18) and the training error $\overline{\text{err}}$ in the case of squared-error loss:

$$\begin{aligned}\text{Err}_{\text{in}} &= \frac{1}{N}\sum_{i=1}^{N} \text{E}_{Y^0}(Y_i^0 - \hat{f}(x_i))^2 \\ \overline{\text{err}} &= \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{f}(x_i))^2.\end{aligned}$$

Add and subtract $f(x_i)$ and $\text{E}\hat{f}(x_i)$ in each expression and expand. Hence establish that the average optimism in the training error is

$$\frac{2}{N}\sum_{i=1}^{N}\text{Cov}(\hat{y}_i, y_i),$$

as given in (7.21).

sol)

$$\begin{cases} op \equiv \text{Err}_{in} - \overline{\text{err}} \\ \omega \equiv E_y(op) \\ \quad = E_y(\text{Err}_{in} - \overline{\text{err}}) = \frac{2}{N}\sum_{i=1}^{N}\text{Cov}(\hat{y}_i, y_i) \end{cases}$$

$E_y(\text{Err}_{in} - \overline{\text{err}})$

$$= E_y\left(\frac{1}{N}\sum_{i=1}^{N} E_{Y_0}(Y_i^0 - \hat{f}(x_i))^2 - \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{f}(x_i))^2\right)$$

$$= \frac{1}{N}\sum_{i=1}^{N} E_y E_{Y_0}(Y_i^0 - \hat{f}(x_i))^2 - \frac{1}{N} E_y \sum_{i=1}^{N}(y_i - \hat{f}(x_i))^2$$

let $\hat{y}_i = \hat{f}(x_i)$

$$= \frac{1}{N}\sum_{i=1}^{N}\left[E_y E_{Y_0}(Y_i^0 - \hat{y}_i)^2 - E_y(y_i - \hat{y}_i)^2\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left[E_y E_{Y_0}(Y_i^{0\,2} - 2Y_i^0\hat{y}_i + \hat{y}_i^2) - E_y(y_i^2 - 2y_i\hat{y}_i + \hat{y}_i^2)\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left[\cancel{E_{Y_0}(Y_i^{0\,2})} - 2E_yE_{Y_0}(Y_i^0\hat{y}_i) + \cancel{E_y(\hat{y}_i^2)} - \cancel{E_y(y_i^2)} + 2E_y(y_i\hat{y}_i) - \cancel{E_y(\hat{y}_i^2)}\right]$$

$$\begin{cases} E_{Y_0}(Y_i^{0\,2}) = E_y(y_i^2) \\ E_{Y_0}(Y_i^0) = E_y(y_i) \end{cases}$$

$$= \frac{2}{N}\sum_{i=1}^{N}\left[E_y(y_i\hat{y}_i) - E_y(\hat{y}_i)E_y(y_i)\right] = \frac{2}{N}\text{Cov}(\hat{y}_i, y_i) \blacksquare$$