1. BostonHousing 데이터에 대해 'medv' 변수 Forward Stepwise selection 해보기

```r
library(mlbench)
```

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

```r
library(leaps)
```

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

library와 Data 불러오기

```r
data("BostonHousing")

sum(is.na(BostonHousing$medv))
```

```
## [1] 0
```

```r
dim(BostonHousing)
```

```
## [1] 506  14
```

```r
dim(na.omit(BostonHousing))
```

```
## [1] 506  14
```

결측치 유무 확인

```r
regfit.fwd=regsubsets(medv~., data=BostonHousing, method="forward")
summary(regfit.fwd)
```

Forward Stepwise Selection 진행

```
## Subset selection object
## Call: regsubsets.formula(medv ~ ., data = BostonHousing, 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 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 ) " "  "*" " "   "*"   "*" "*" " " "*" " " " " "*"     "*" "*"  
```

⇒ 설명변수 4개인 모델에서 새로 선택된 변수 : "dis"

```r
coef(regfit.fwd,4)
```

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

⇒ 'dis'의 추정된 계수 : -0.5519263

```r
summary(regfit.fwd)$rsq
```

```
## [1] 0.5441463 0.6385616 0.6786242 0.6903077 0.7080893 0.7157742 0.7221614
## [8] 0.7266079
```

$\Rightarrow R^2 = 0.6903077$

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).

목표

$$op \equiv \text{Err}_{in} - \overline{err}$$

$$w = E_y(op) = \boxed{E_y(\text{Err}_{in} - \overline{err})} = \frac{2}{N}\sum \text{Cov}(\hat{y}_i, y_i)$$

과정

(다음페이지)

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

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

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

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

$$= \frac{1}{N}\sum_{i=1}^{n}\left[E_{y^0} E_y (Y_i^{02} - 2Y_i^0 \hat{y}_i + \hat{y}_i^2) - E_y (y_i^2 - 2y_i\hat{y}_i + \hat{y}_i^2)\right]$$

$$= \frac{1}{N}\sum_{i=1}^{n}\left[E_{y^0}(Y_i^{02}) - 2E_{y^0}E_y(Y_i^0\hat{y}_i) + E_y(\hat{y}_i^2) - E_y(y_i^2) + 2E_y(y_i\hat{y}_i) - E_y(\hat{y}_i^2)\right]$$

$$= \frac{1}{N}\sum_{i=1}^{n}\left[\cancel{E_{y^0}(Y_i^{02})} - 2E_{y^0}E_y(Y_i^0\hat{y}_i) - \cancel{E_y(y_i^2)} + 2E_y(y_i\hat{y}_i)\right]$$

$\Rightarrow E_{y^0}(Y_i^{02}) = E_y(y_i^2)$

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

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

$\Rightarrow E_{y^0}(Y_i^0) = E_y(Y_i)$

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

$$= \boxed{\frac{2}{N}\cdot\sum_{i=1}^{n}\mathrm{cov}(y_i, \hat{y}_i)}$$

The Y^0 notation indicates that we observe N new response values at each of the training points x_i, $i = 1, 2, \ldots, N$.

$\Rightarrow$ 원래 Training Set $T = \{(x_1, y_1), (x_2, y_2) \cdots (x_N, y_N)\}$