6.1 **보스턴 주택가격 예측** BonstonHousing.csv 파일은 미국 인구통계조사국에 의해 수집된 메사추세츠주 보스턴 지역의 주택정보를 담고 있다. 이 데이터 세트는 보스턴 지역의 506개의 인구통계조사 주택단지 정보를 포함한다. 데이터 수집의 목표는 범죄율, 공해, 방의 개수에 대한 정보에 근거하여 새로운 주택단지의 주택가격의 중앙값(median)을 예측하는 것이다. 데이터 세트는 13개의 예측변수와 반응(목표)변수인 주택가격의 중앙값(MEDV)을 포함한다.

a. 데이터를 학습 세트와 검증 세트로 나누는 이유는 무엇인가? 학습 세트와 검증 세트는 어떠한 용도로 사용되는가? CRIM, CHAS 및 RM를 이용하는 함수로서 주택가격 중앙값에 대한 다중선형회귀모델을 만드시오.

>>> 데이터를 학습 세트와 검증 세트로 나누는 이유는 학습 세트를 통해 모델을 구축하고 검증 세트를 통해 구축된 모델들을 비교하기 위해서입니다.

>>> 다중선형회귀모델 만드는 과정

```
setwd("c:/rdata")
BHouse.df <- read.csv("BostonHousing.csv")

#select variables for regression
selected.var <- c(1, 4, 6, 13)

#partition data
set.seed(1)
train.index <- sample(c(1:506),300)
train.df <- BHouse.df[train.index, selected.var]
valid.df <- BHouse.df[-train.index, selected.var]

#linear regression
str(train.df)
house.lm <- lm(MEDV~., data=train.df)
```

b. 모델의 예측변수들로부터 주택가격 중앙값을 예측하기 위한 식을 쓰시오.

$\widehat{MEDV} = -28.65705 - 0.24463\,CRIM + 5.38191\,CHAS + 8.25004\,R.M$

c. 찰스강 경계에 위치하지 않고, 범죄율이 0.1이고, 평균 방의 개수가 6개인 보스턴 주택단지의 주택가격 중앙값은 얼마로 예측되는가? 예측오차($\hat{y} - y$)는 얼마인가?

>>> $-28.65705 - 0.24463*0.1 + 5.38191*0 + 8.25004*6 = 20.81873$

주택가격의 중앙값은 20.81873($1,000 단위)로 예측된다.

```
library(forecast)
house.lm.pred <- predict(house.lm, valid.df)
accuracy(house.lm.pred, valid.df$MEDV)
accuracy(house.lm$fitted.values, train.df$MEDV)
> accuracy(house.lm.pred, valid.df$MEDV)
                ME    RMSE     MAE      MPE    MAPE
Test set -0.36088 6.2714 4.2333 -9.8834 23.842
> accuracy(house.lm$fitted.values, train.df$MEDV)
                                ME    RMSE     MAE      MPE    MAPE
Test set -0.000000000000000094949 6.0854 4.2279 -7.1468 22.868
```

>>> 이 문제에서는 예측오차는 ($\hat{y} - y$)를 구할 수 없다.