6.1 **Predicting Boston Housing Prices.** The file *BostonHousing.csv* contains information collected by the US Bureau of the Census concerning housing in the area of Boston, Massachusetts. The dataset includes information on 506 census housing tracts in the Boston area. The goal is to predict the median house price in new tracts based on information such as crime rate, pollution, and number of rooms. The dataset contains 13 predictors, and the response is the median house price (MEDV). Table 6.9 describes each of the predictors and the response.

**TABLE 6.9 DESCRIPTION OF VARIABLES FOR BOSTON HOUSING EXAMPLE**

| | |
|---|---|
| CRIM | Per capita crime rate by town |
| ZN | Proportion of residential land zoned for lots over 25,000 $ft^2$ |
| INDUS | Proportion of nonretail business acres per town |
| CHAS | Charles River dummy variable (= 1 if tract bounds river; = 0 otherwise) |
| NOX | Nitric oxide concentration (parts per 10 million) |
| RM | Average number of rooms per dwelling |
| AGE | Proportion of owner-occupied units built prior to 1940 |
| DIS | Weighted distances to five Boston employment centers |
| RAD | Index of accessibility to radial highways |
| TAX | Full-value property-tax rate per $10,000 |
| PTRATIO | Pupil/teacher ratio by town |
| LSTAT | Percentage lower status of the population |
| MEDV | Median value of owner-occupied homes in $1000s |

**a.** Why should the data be partitioned into training and validation sets? What will the training set be used for? What will the validation set be used for?

a. Training set will be used for building the model, and the validation set is used for assessing the model. The data is partitioned into training and validation sets because testing on training set is not accurate since the model can be overfitted. The goal of predictive modeling is to build a robust model for unseen data.

**b.** Fit a multiple linear regression model to the median house price (MEDV) as a function of CRIM, CHAS, and RM. Write the equation for predicting the median house price from the predictors in the model.

```r
#6.1.b
#Fit a multiple linear regression model to the median house price (MEDV) as a
#function of CRIM, CHAS, and RM. Write the equation for predicting the median
#house price from the predictors in the model.
b.df <- read.csv("BostonHousing.csv")
head(b.df)
dim(b.df)
train.rows <- sample(c(1:500),300)#300 rows
length(train.rows)

train.df <- b.df[train.rows,]

reg <- lm(MEDV ~ CRIM+CHAS+RM, data = train.df)
summary(reg)

head(data.frame(reg$fitted.values, train.df$MEDV))
median(reg$fitted.values)#predicted median
median(train.df$MEDV)#real median
```

**c.** Using the estimated regression model, what median house price is predicted for a tract in the Boston area that does not bound the Charles River, has a crime rate of 0.1, and where the average number of rooms per house is 6? What is the prediction error?