Problem Set 2

YILIN LI

Chapter 3 exercises

Exercise 1

The null hypothesis for p-value of "TV" is that given the advertisement of radio and newspaper, TV has no effect on sales. The null hypothesis for p-value of "radio" is that given the advertisement of TV and newspaper, radio has no effect on sales.The null hypothesis for p-value of "newspaper" is that given the advertisement of radio and TV, newspaper has no effect on sales. We can draw a conclusion that given the advertising bugets of TV, radio and newspaper, it suggests that TV and radio have some relation with sales, but newspaper has no relation with sales.

Exercise 4

a. In fact, the cubic regression will have a lower training RSS rather than linear regression, because cubic regression is a more flexible model than linear regression. Therefore, with the same dataset, the cubic regression will fit with a lower RSS.
b. Estimate the cubic regression model with a test set will generate a higher RSS rather than the linear model because the cubic model will overfit the data and thus unable to reflect the true relationship.
c. In fact, the cubic regression will have a lower training RSS no matter what the true relationship is, because the cubic regression is more flexible than the linear regression. Thus, the cubic regression will have a fitting line closer to the data points, resulting a lower training RSS.
d. There is not enough information to tell which one has the lower testing RSS. If the true relationship is closer to the linear regression than the cubic regression, then the linear regression will have the lower testing RSS. If the true relationship is closer to the cubic regression, then the cubic regression will have the lower testing RSS.

Exercise 5

$$\hat{y_i} = x_i\hat{\beta} = x_i\frac{\sum_{j=1}^{n} x_j y_j}{\sum_{k=1}^{n} x_k^2} = \sum_{j=1}^{n}\frac{x_i x_j}{\sum_{k=1}^{n} x_k^2} y_j = \sum_{j=1}^{n} a_j y_j$$

Thus we know $a_j = \frac{x_i x_j}{\sum_{k=1}^{n} x_k^2}$ where x_i is the corresponding input value for a given predicted value.

Additional exercises

The k-nearest neighbor regression was defined as:

$$\hat{f}(x) = \frac{1}{k}\sum_{x_i \in \mathcal{N}(x)} y_i$$

We have derived from class that:

$$MSE_{test} = Var(f(\hat{x_0})) + [E(f(x_0) - f(\hat{x_0}))]^2 + Var(\epsilon)$$

Thus, the variance of KNN regression model is:

$$\begin{aligned} Var(f(\hat{x_0})) &= Var(\frac{1}{k}\sum_{x_i \in \mathcal{N}(x_0)} y_i) \\ &= \frac{1}{k^2} Var(\sum_{x_i \in \mathcal{N}(x_0)} y_i) \\ &= \frac{1}{k^2} Var(\sum_{x_i \in \mathcal{N}(x_0)} (f(x_i) + \epsilon_i)) \\ &= \frac{1}{k^2} Var(\sum_{x_i \in \mathcal{N}(x_0)} \epsilon_i) \qquad\qquad x_0 \text{ has been fixed.} \\ &= \frac{1}{k^2} |\mathcal{N}(x_0)| \sigma_\epsilon^2 \\ &= \frac{\sigma_\epsilon^2}{k} \end{aligned}$$

Then, the $Bias^2$ of KNN regression model is:

$$\begin{aligned} Bias^2(\hat{f}(x_0)) &= [E(f(x_0) - f(\hat{x_0}))]^2 \\ &= [E(f(x_0) - \frac{1}{k}\sum_{x_i \in \mathcal{N}(x_0)} y_i)]^2 \\ &= [E(y_0 - \epsilon_0 - \frac{1}{k}\sum_{x_i \in \mathcal{N}(x_0)} y_i)]^2 \\ &= (y_0 - \frac{1}{k}\sum_{x_i \in \mathcal{N}(x_0)} y_i)^2 \qquad\qquad \text{the expected value of } \epsilon_0 \text{ is 0.} \end{aligned}$$

Therefore, we derived the MSE_{test} as follows:

$$\begin{aligned} MSE_{test} &= Var(f(\hat{x_0})) + Bias^2(\hat{f}(x_0)) + Var(\epsilon) \\ &= \frac{\sigma_\epsilon^2}{k} + (y_0 - \frac{1}{k}\sum_{x_i \in \mathcal{N}(x_0)} y_i)^2 + \sigma_\epsilon^2 \end{aligned}$$

```r
library(ggplot2)
x <- c(1:3, 5:12)
y <- c(-7.1, -7.1, .5, -3.6, -2, -1.7, -4, -.2, -1.2, -1.2, -3.5)
#bias_sq calculate the bias^2 based on test data(x0,y0) and number of nearest neighbors k
bias_sq <- function(k, x0, y0, x, y) {
  (y0 - sum(y[sort(abs(x - x0),index.return=TRUE)$ix[1:k]])/k)^2
}

bias = rep(NA,11)
for (k in 1:11){
  bias[k] = bias_sq(k,4,1,x,y) # Assume the test data is (4.0, 1.0)
}

k = 1:11
sigma2 = 1  # Assume variance of the irreducible error term is 1
variance = sigma2/k
```

```r
ggplot()+
  geom_line(aes(x = k, y = bias, color = "bias"))+
  geom_line(aes(x = k, y = variance, color = "variance"))+
  geom_line(aes(x = k, y = sigma2, color = "irreducible" ))+
  geom_line(aes(x = k, y = bias + variance + sigma2, color = "MSE"))+
  theme(legend.title = element_blank())+
  ylab("variability")
```