Data Mining HW 1

Due: 2022.04.07 24:00

Exercises for Linear Regression

1. 설명변수가 1개(X)인고, 반응변수가 1개(Y)인데이터를 가지고 있다고 하자. ($n = 100$) 그리고 다음의 두 모형(linear regression, cubic regression)을 적합시키려고 한다.

$$Y = \beta_0 + \beta_1 X + \epsilon \quad (1)$$
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon \quad (2)$$

   (a) 실제 X, Y가 선형(linear)관계가 있다고 가정 하자. 모델 (1),(2)의 SSE(잔차제곱합)의 크기를 비교할 수 있는지 설명하여라.

   (b) 실제 X, Y가 비선형(non-linear)관계가 있다고 가정 하자. 대신 실제 모형에 대한 정보는 없다. 모델 (1),(2)의 SSE(잔차제곱합)의 크기를 비교할 수 있는지 설명하여라.

2. 'Auto.csv' 데이터를 이용하여 단순선형 회귀 모형을 적합한다.

   (a) 반응변수 mpg, 설명변수는 horsepower로 하는 단순선형회귀모형을 적합 시킨 후 summary() 함수의 결과 확인하고 다음의 물음에 답하여라.

      i. 두 변수 사이에 관계가 있는가?
      ii. 두 변수 사이의 관계는 얼마나 강한가?
      iii. 두 변수는 음의 관계가 있는가? 양의 관계가 있는가?
      iv. horsepower의 값이 98일 때, mpg의 예측값은 무엇인가 95% 신뢰구간은 무엇인가?

   (b) 설명변수와 반응변수의 산점도를 그리고, 회귀직선을 추가하여라. (abline() 사용)

3. 이 문제는 다중공선성(collinearity)에 관련한 것이다.

   (a) R에 다음의 명령문을 실행하여라.

```
> set.seed (1)
> x1=runif (100)
> x2 =0.5* x1+rnorm (100) /10
> y=2+2* x1 +0.3* x2+rnorm (100)
```

   마지막 줄이 두개의 설명변수를 이용한 중회귀모형이다. 회귀모형을 쓰시오. (β 등을 이용하여)

   (b) 두 설명변수 $x1$과 $x2$ 사이에 상관관계(correlation)이 있는가? 산점도를 그려서 확인하여라.

   (c) 생성된 데이터를 이용하여 (a) 모형의 회귀계수를 추정하여라. 실제 회귀계수와 추정된 회귀계수와 비교하여라. $H_0 : \beta_1 = 0$을 기각할 수 있는가? $H_0 : \beta_2 = 0$을 기각할 수 있는가?

   (d) 이번에는 $x1$만을 이용한 단순선형회귀 모형을 적합하여라. 결과를 분석하여라. $H_0 : \beta_1 = 0$을 기각할 수 있는가?

   (e) 이번에는 $x2$만을 이용한 단순선형회귀 모형을 적합하여라. 결과를 분석하여라. $H_0 : \beta_2 = 0$을 기각할 수 있는가?

   (f) (c)-(e)의 결과가 서로 모순되는가? 설명하여라.

   (g) 새로운 데이터가 관측되었다고 하자.(이 데이터는 잘못 측정된 것이다.)

```
> x1=c(x1 , 0.1)
> x2=c(x2 , 0.8)
> y=c(y,6)
```

   추가된 데이터를 이용하여 (c)-(e)를 다시 적합하여라. 결과가 어떻게 달라졌는가? 각 모형에서 새로운 데이터는 이상점인가?(잔차가 기존에 있는 데이터에 비해 많이 큰가?) 아니면 영향점인가?(추가된 데이터로 인해 회귀계수의 값이 많이 바뀌었는가?) 설명하여라.

Exercises for Logistic Regression

1. 두개의 설명변수 (X_1 = 공부시간, X_2 = 학부평점)를 이용하여 A학점을 받을 확률을 예측하기 위해 로지스틱 회귀모형을 적합하였다. 추정된 회귀계수는 $\hat{\beta}_0 = -6, \hat{\beta}_1 = 0.05, \hat{\beta}_2 = 1$이다.
   (a) 40시간 공부하고, 평점이 3.5인 학생이 A를 받았을 확률을 예측하여라.
   (b) 평점이 3.5인 학생은 얼마나 공부를 해야 A를 받을 확률이 50%를 넘을 것인가?

2. 다음은 odds에 관한 문제이다.
   (a) 신용카드결재 문제에서 결재를 하지 못하는 경우(default)에 대한 odds가 0.37인 사람들이 실제로 defalut할 확률은 평균적으로 얼마인가?
   (b) 어떤 개인이 default할 확률이 16% 라고 하자. 그 사람이 default할 odds는 얼마인가?