고급회귀분석론

Ch3. Multiple Linear Regression

양성준

중선형회귀모형

- 둘 이상의 예측변수와 반응변수 하나의 관계를 선형관계(linear relationship)로 모형화

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

  $E(\epsilon) = 0,\ var(\epsilon) = \sigma^2.$

- $E(y|x_1, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ and $var(y|x_1, \ldots, x_p) = \sigma^2.$
- 각 β_j는 x_j를 제외한 다른 예측변수들의 값이 정해졌을 때(혹은 변하지 않을 때) x_j의 1단위 변화로 나타나는 반응변수 y에서의 변화량으로 해석할 수 있다.
- 예측변수들과 반응변수 사이의 함수관계를 모형화 하는 가장 간단한 방법 중 하나이다.

중선형회귀모형

- 다항회귀모형 또한 중선형 회귀모형의 일종으로 간주할 수 있다.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 + \epsilon$$

- 교호작용(interaction) 효과를 포함한 모형 또한 중선형 회귀모형으로 간주할 수 있다.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon$$

- 다항함수와 교호작용 효과를 동시에 포함한 모형도 중선형 회귀모형의 일종이다.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon$$

중선형회귀모형의 추정

- 먼저 얻게 된 관측치 쌍이 $(x_{1i}, \ldots, x_{ki}, y_i),\ i = 1, 2, \ldots, n$이라 하자.

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \epsilon_i$$

- 회귀계수 : $\beta = (\beta_0, \ldots, \beta_k)^\top$
- ϵ_i들은 평균이 0이고 분산이 σ^2인 분포로부터의 iid random sample
- 추정대상은 β 혹은 오차항의 분산 σ^2.

최소제곱추정(least-squares estimation)

- 최소제곱추정법은 모형에 의한 반응변수의 추정치와 실제 반응변수의 관측치 사이의 거리의 제곱합을 최소화하는 직선을 추정모형으로 선택하는 것이다.

$$S(\beta_0, \beta_1, \ldots, \beta_k) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_k x_{ki})^2$$

- 위 식이 어떤 $\beta_0, \beta_1, \ldots, \beta_k$에서 최소가 되는지를 푸는 문제로 귀결된다.
- $\frac{\partial}{\partial \beta_j} S(\beta_0, \beta_1, \ldots, \beta_k) = 0,\ j = 0, 1, \ldots, k$을 연립해서 풀어 얻어지는 해가 최소제곱추정량이다.
- 즉, $p = k + 1$원 일차 연립방정식을 푸는 문제로 볼 수 있다.

최소제곱추정량

- 행렬형식으로 최소제곱 추정 문제를 다루면 매우 편리하다.
- $x_j = (x_{1j}, \ldots, x_{nj})^\top$를 j번째 예측변수의 관측치 벡터,
  $y = (y_1, \ldots, y_n)^\top$을 반응변수의 관측치 벡터로 정의하자.
  예측변수들의 관측치를 모아놓은 행렬을 $X = (1_n, x_1, \ldots, x_k)$라 하면
  X는 $n \times (k+1)$행렬이 된다. 여기서 $1_n = (1, \ldots, 1)^\top$을 나타낸다.
- X를 전통적으로는 design matrix라 부른다.
- 행렬 형식으로 오차제곱합을 재표현하면 다음과 같다.

$$S(\beta) = \sum_{i=1}^{n} (y - X\beta)^\top (y - X\beta)$$

최소제곱추정량

- $S(\beta)$를 전개하면

$$S(\beta) = y^\top y - 2\beta^\top X^\top y + \beta^\top X^\top X\beta$$

- 최소제곱추정량은 다음 식의 해로 표현된다. 이를 정규방정식이라 한다.

$$\frac{\partial S(\beta)}{\partial \beta} = -2X^\top y + 2X^\top X\beta = 0$$

- 따라서 최소제곱추정량은

$$\hat{\beta} = (X^\top X)^{-1} X^\top y$$

- 위 추정량은 $(X^\top X)^{-1}$이 존재한다는 전제 하에 유일하게 정의된다.

적합치 및 잔차

- 주어진 x_i에서 최소제곱직선에 의해 결정되는 y_i의 값을 적합치(fitted value)라 한다.

$$(\hat{y}_1, \ldots, \hat{y}_n)^\top = \hat{y} = X\hat{\beta} = X(X^\top X)^{-1}X^\top y = Hy$$

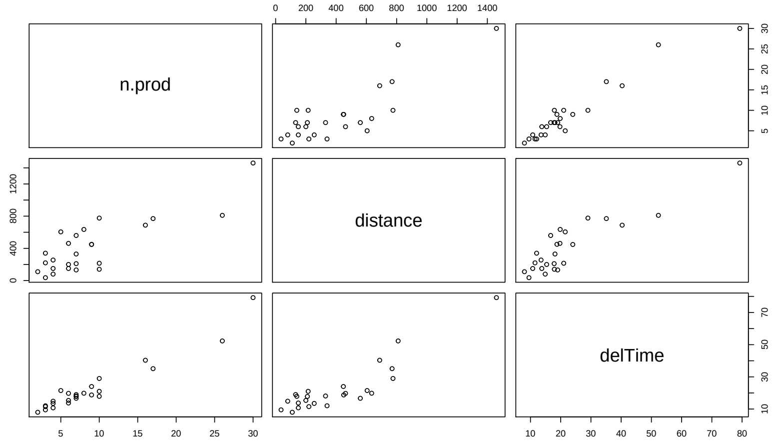
- $n \times n$ 행렬 $H = X(X^\top X)^{-1}X^\top$를 hat matrix라 한다. 이 행렬은 반응변수 벡터 y를 적합치벡터 $\hat{y}$로 연결해 주는 역할을 하게 된다.
- H와 그 성질은 중회귀분석에서 매우 핵심적인 역할을 한다.
- 잔차벡터는 다음과 같이 정의된다.

$$(e_1, \ldots, e_n)^\top = e = y - \hat{y} = y - Hy = (I - H)y$$

Example : Delivery time data

- x_1 : number of products, x_2 : distance, y : delivery time

```r
library(robustbase);plot(delivery)
```



Example : Delivery time data

```r
n = nrow(delivery)    # sample size
# design matrix
X = cbind(rep(1,n),as.matrix(delivery[,-3]))
y = delivery$delTime   # response vector
head(cbind(y,X))
```

```
##          y   n.prod distance
## [1,] 16.68 1      7      560
## [2,] 11.50 1      3      220
## [3,] 12.03 1      3      340
## [4,] 14.88 1      4       80
## [5,] 13.75 1      6      150
## [6,] 18.11 1      7      330
```

Example : Delivery time data

```
# least squares estimator
hbeta = solve(t(X)%*%X)%*%t(X)%*%y
hy = X%*%hbeta # fitted value
# fitting by built-in function
fit1 = lm(delTime~ . ,data=delivery)
cbind(fit1$coefficients, hbeta) # comparison of estimates
```

```
##                    [,1]       [,2]
## (Intercept) 2.34123115 2.34123115
## n.prod      1.61590721 1.61590721
## distance    0.01438483 0.01438483
```

```
# comparison of fitted values
sum(abs((fit1$fitted.values - hy)))
```

```
## [1] 1.056044e-12
```

Example : Delivery time data

```r
head(cbind(y, hy , y-hy),10)
```

```
##             y
##  [1,] 16.68 21.708084 -5.0280843
##  [2,] 11.50 10.353615  1.1463854
##  [3,] 12.03 12.079794 -0.0497937
##  [4,] 14.88  9.955646  4.9243539
##  [5,] 13.75 14.194398 -0.4443983
##  [6,] 18.11 18.399574 -0.2895743
##  [7,]  8.00  7.155376  0.8446235
##  [8,] 17.83 16.673395  1.1566049
##  [9,] 79.24 71.820294  7.4197062
## [10,] 21.50 19.123587  2.3764129
```

최소제곱추정량의 기하학적 의미

https://bre.is/rYSSjhvm

- A : 원점으부터 y에 의해 정의되는 n차원 공간상에서의 지점
- B : 원점으로부터 $1_n, x_1, \ldots, x_k$의 선형결합으로 표현되는 벡터로 정의. 선형결합은 가중치 벡터 $\beta \in R^p$에 대하여

$$\beta_0 \cdot 1_n + \beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k = X\beta$$

  으로 표현된다. 이렇게 표현되는 B지점의 모임을 estimation space라 한다.
- A는 실제 관측결과, B는 회귀모형에 의해 표현 가능한 것이다. 즉, 이 둘 사이의 거리가 가까울 수록 좋을 것이다.
- A와 estimation space 상의 한 지점 B 사이의 거리제곱은

$$S(\beta) = (y - X\beta)^\top (y - X\beta)$$

최소제곱추정량의 기하학적 의미

- 위 거리를 최소로 하는 지점을 B0라 하자. 그러면, B0는 A의 estimation space 위로의 정사영이어야 한다.
- 다시 말해 A와 B0를 연결하는 벡터는 estimation space 혹은 임의의 B벡터와 수직이어야 한다.
- B0를 정의하는 가중치 벡터를 $\hat{\beta}$라 하자. 즉, B0는 $X\hat{\beta}$로 표현된다.
- 벡터끼리 수직이려면 내적이 0이면 된다. 즉, $\hat{\beta}$는 임의의 $\beta \in R^p$에 대하여
$$(X\beta)^\top(y - X\hat{\beta}) = 0$$
을 만족해야 한다.
- 이는 $X^\top X\hat{\beta} = X^\top y$로 귀결되고 이는 정규방정식과 같다.

최소제곱추정량의 성질

- 불편성

$$E(\hat{\beta}) = E((X^\top X)^{-1}X^\top y) = E((X^\top X)^{-1}X^\top(X\beta+\epsilon)) = \beta$$

- 공분산행렬

$$var(\hat{\beta}) = (X^\top X)^{-1}X^\top var(y)X(X^\top X)^{-1} = \sigma^2(X^\top X)^{-1}$$

오차분산의 추정

- 잔차제곱합

$$SS_R = \sum_i e_i^2 = e^\top e$$

을 잔차제곱합의 자유도 $n - p = n - k - 1$로 나눈 값으로 추정

$$\hat{\sigma}^2 = MS_R = \frac{SS_R}{n - p}$$

- 자유도가 왜 $n - p$인가? 총 n개의 잔차를 제곱해서 합하지만,

$$e^\top 1_n = 0,\ e^\top x_j = 0,\ j = 1, \ldots, k$$

이 성립하여 총 $p = k + 1$개의 제약식이 존재하기 때문임.

Example：Delivery time data

```r
e=y-as.vector(hy) # residual
(SSR = sum(e^2))
```

```
## [1] 233.7317
```

```r
(MSR = SSR/(n-ncol(X))) # hat sigma^2
```

```
## [1] 10.62417
```

```r
sum(fit1$residuals^2)
```

```
## [1] 233.7317
```

```r
summary(fit1)$sigma # hat sigma
```

```
## [1] 3.259473
```

최대가능도추정량

- 오차항 벡터에 대한 다음의 가정 하에서

$$\epsilon \sim N(0, \sigma^2 I_n)$$

가능도 함수는 다음과 같이 표현된다.

$$L(\epsilon, \beta, \sigma^2) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp(-\frac{1}{2\sigma^2}\epsilon^\top \epsilon)$$

$\epsilon = y - X\beta$이므로, 가능도 함수는

$$L(y, X, \beta, \sigma^2) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp(-\frac{1}{2\sigma^2}(y - X\beta)^\top (y - X\beta))$$

위 가능도 함수를 최대화 하는 β, σ^2이 최대가능도 추정량이다.

$$\hat{\beta} = (X^\top X)^{-1} X^\top y, \ \ \hat{\sigma}^2 = \frac{(y - X\beta)^\top (y - X\beta)}{n}$$