고급회귀분석론

Ch3. Multiple Linear Regression

양성준

중선형회귀모형

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

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$

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

- $E(y|x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ and $var(y|x_1, \ldots, x_k) = \sigma^2.$
- 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.
- $x_j = (x_{j1}, \ldots, x_{jn})^\top$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.

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

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.

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

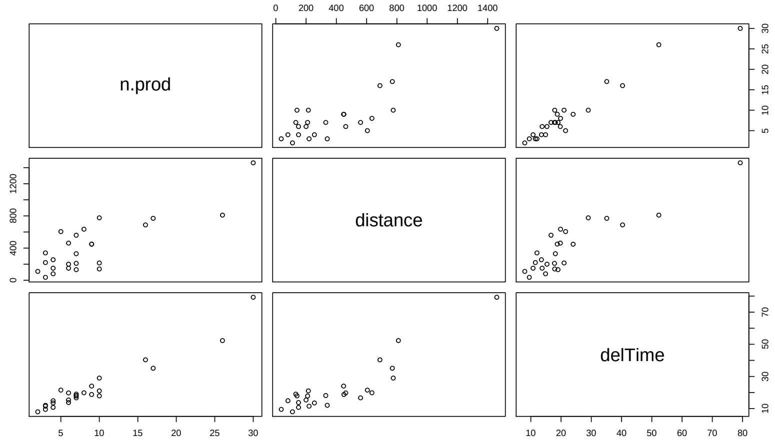
- $n \times n$ 행렬 $H = X(X^\top X)^{-1}X^\top$를 hat matrix라 한다. 이 행렬은 반응변수 벡터 y를 적합치벡터 $\hat{y}$로 연결해 주는 역할을 하게 된다.
- H와 그 성질은 중회귀분석에서 매우 핵심적인 역할을 한다.
- 잔차벡터는 다음과 같이 정의된다.

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

Example : Delivery time data

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

```
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

```r
# 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
```

```r
# 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 : 원점으로부터 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

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

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

$$\hat{\sigma}^2 = MSR = \frac{SSR}{n - p}$$

- 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

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

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

```
(MSE = SSE/(n-ncol(X))) # hat sigma^2
```

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

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

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

```
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.

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

모수(계수)에 대한 검정

- 중회귀분석에서는 크게 다음과 같은 질문에 답하기 위한 검정을 시행할 수 있다.
  - 모형이 전반적으로 적절한가?
  - 개별 예측변수들은 중요한가?
- 기본적인 검정을 위해서는 앞서 가정한 오차항의 독립성, 등분산성 외에도 정규성 가정이 필요한 것이 일반적이다.

회귀모형의 유의성 검정

- Overall or global test
- 다음과 같이 표현할 수 있다.

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k \quad vs \quad H_1 : \beta_j \neq 0 \; for \; some \; j$$

- 즉, 귀무가설이 기각되면 k개의 예측변수들 중 중요한 것이 적어도 하나는 존재한다는 의미로 회귀모형이 완전히 쓸모없는 것은 아니라는 뜻이다.
- 단순선형회귀모형의 경우와 비슷하게 총변동을 분해하여 검정한다.

$$SST = SSR + SSE$$

회귀모형의 유의성 검정

- 다음 사실을 보일 수 있다. (Appendix C.3 참고)
  - $SSE/\sigma^2 \sim \chi^2_{n-k-1}$
  - $SSR/\sigma^2 \sim \chi^2_k$ under $\beta_1 = \cdots = \beta_k = 0.$
- 또한, SSR과 SSE는 서로 독립임을 보일 수 있다. (why?)
- F분포의 정의로부터

$$F_0 = \frac{SSR/k}{SSE/(n-k-1)} \sim F_{k,n-k-1}$$

under $H_0 : \beta_1 = \cdots = \beta_k = 0$

회귀모형의 유의성 검정

- F_0의 관측치가 크면 H_0를 부정하는 증거가 강한 것으로 볼 수 있다.
- 유의수준 α에서 $F_0 > F_{\alpha,k,n-k-1}$이면 귀무가설을 기각한다.
- 제곱합의 행렬 표현
  - $SST = \sum_i (y_i - \bar{y})^2 = y^\top (I_n - 1_n(1_n^\top 1_n)^{-1} 1_n^\top) y$
  - $SSE = y^\top (I_n - H) y$
  - $SSR = SST - SSE = y^\top (H - 1_n(1_n^\top 1_n)^{-1} 1_n^\top) y$
- H와 $1_n(1_n^\top 1_n)^{-1} 1_n^\top$는 멱등행렬(idempotent)임을 이용
- $1_n(1_n^\top 1_n)^{-1} 1_n^\top y = (\bar{y}, \ldots, \bar{y})^\top$

Example : Delivery time data

```
ov = rep(1,n);P1 = ov%*%t(ov)/n
H=X%*%solve(t(X)%*%X)%*%t(X)
SSE = t(y)%*%(diag(1,n) - H)%*%y
SSE1 = t(y)%*%(diag(1,n) - H)%*%(diag(1,n) - H)%*%y
SSE2 = t(y-hy)%*%(y-hy)
c(SSE,SSE1,SSE2)
```

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

```
SST = t(y)%*%(diag(1,n) - P1)%*%y
SST1 = sum((y-mean(y))^2)
c(SST,SST1)
```

```
## [1] 5784.543 5784.543
```

Example：Delivery time data

```r
(F0 = ((SST-SSE)/(ncol(X)-1))/(SSE/(n-ncol(X))))
```

```
##          [,1]
## [1,] 261.2351
```

```r
F0 > qf(0.95,ncol(X)-1,n-ncol(X))
```

```
##      [,1]
## [1,] TRUE
```

```r
1-pf(F0,ncol(X)-1,n-ncol(X))  # p-value
```

```
##              [,1]
## [1,] 4.440892e-16
```

```r
#summary(fit1)
```

결정계수

- 결정계수는 단순선형회귀모형의 경우와 동일하게 정의된다.

$$R^2 = \frac{SSR}{SST}$$

- 결정계수는 예측변수가 추가되면 무조건 증가한다 (why?). 따라서 예측변수의 수를 염두에 둔 결정계수를 정의해서 사용하기도 한다.
- 수정결정계수는 다음과 같이 정의된다.

$$R^2_{adj} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$

Example : Delivery time data

```
(R2 = (SST-SSE)/SST)
```

```
##              [,1]
## [1,] 0.9595937
```

```
(AR2 = 1 - (MSE/var(y)))
```

```
## [1] 0.9559205
```

개별 회귀계수에 대한 검정

- 특정 예측변수가 모형에서 중요한지를 개별 회귀계수에 대한 다음 검정을 통해 살펴본다.

$$H_0 : \beta_j = 0 \quad vs \quad H_1 : \beta_j \neq 0$$

- 만약 귀무가설을 기각할 수 없다면 x_j는 모형에서 제외될 수 있다.
- 모형에 대한 가정 하에서 $\hat{\beta}_j$는 정규분포를 따른다. C_{jj}가 $(X^\top X)^{-1}$의 j번째 대각원소라고 할 때, $var(\hat{\beta}_j) = \sigma^2 C_{jj}$이므로 위 가설 검정을 위한 통계량은

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

- 유의수준 α에서 $|t_0| > t_{\alpha,n-k-1}$이면 귀무가설을 기각한다.
- (marginal test) 이 검정은 다른 예측변수들이 모형에 함께 있는 상황에서 j번째 예측변수의 유의성에 대한 검정이다.

Example : Delivery time data

```
# distance
C = diag(solve(t(X)%*%X))
(t0 = hbeta[3]/sqrt(MSE*C[3]))
```

```
## distance
## 3.981313
```

```
2*(1-pt(t0,n-ncol(X)))  # p-value
```

```
##     distance
## 0.0006312469
```

```
sqrt(MSE*C[3]) # standard error
```

```
##    distance
## 0.003613086
```

Example：Delivery time data

```r
summary(fit1)
```

```
##
## Call:
## lm(formula = delTime ~ ., data = delivery)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7880 -0.6629  0.4364  1.1566  7.4197
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.341231   1.096730   2.135 0.044170 *
## n.prod      1.615907   0.170735   9.464 3.25e-09 ***
## distance    0.014385   0.003613   3.981 0.000631 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.
##
```

예측변수들의 set에 대한 검정 (Partial F test)

- 여러 개의 예측변수들 중 일부의 유의성을 살펴보고자 할 수 있다. 이는, 포함관계에 있는 두 선형모형의 비교를 위한 목적으로 생각할 수도 있다.
- $\beta = (\beta_{01}^\top, \beta_{02}^\top)^\top,\ \beta_{01} = (\beta_0, \beta_1, \ldots, \beta_{k-r})^\top,\ \beta_{02} = (\beta_{k-r+1}, \ldots, \beta_k)^\top$이라 하자.
- 총 k개의 예측변수들 중 β_{02}에 포함된 $r(< k)$개의 예측변수에 대한 유의성을 살펴보자. 즉, 다음과 같은 가설을 검정하는 것이다.

$$H_0 : \beta_{02} = 0 \quad vs \quad H_1 : not\ H_0$$

- β_{01}, β_{02}에 대응되는 예측변수에 대한 design matrix를 각각 X_1, X_2라 하면 모형은 다음과 같이 표현된다.

$$y = X\beta + \epsilon = X_1\beta_{01} + X_2\beta_{02} + \epsilon$$

Full vs Reduced models

- Full model : 전체 예측변수들에 의해 정의되는 모형을 말한다.
  회귀계수에 대한 추정량 등은 앞서 정의되었다.
- Reduced model : 귀무가설 $H_0 : \beta_{02} = 0$ 하에서 정의되는 모형을 말한다. 즉,

$$y = X_1 \beta_{01} + \epsilon$$

  이며, 이때 β_{01}에 대한 최소제곱추정량은

$$\hat{\beta}_{01} = (X_1^\top X_1)^{-1} X_1^\top y$$

  로 주어진다.

Extra sum of squares

- Full model과 reduced model에서의 회귀제곱합의 차이

$$SSR(\beta_{02}|\beta_{01}) = SSR(\beta) - SSR(\beta_{01})$$

  를 Extra sum of squares라 한다. 이는, β_{02}를 모형에 추가함으로써 얻게 되는 추가적인 모형의 설명력을 나타낸다.
- 위 제곱합의 자유도는 $k + 1 - (k - r + 1) = r$이다.
- $SSR(\beta_{02}|\beta_{01})$은 SSE와 독립이다. 귀무가설 하에서

$$F_0 = \frac{SSR(\beta_{02}|\beta_{01})/r}{SSE/(n-k-1)} \sim F_{r,n-k-1}$$

  임을 보일 수 있다.

Example : Delivery time data

- x_2 : distance 에 대한 유의성 검정을 Partial F test를 이용하여 해 보자.

```
X1 = cbind(rep(1,n),delivery$n.prod)
hbeta1 = solve(t(X1)%*%X1)%*%t(X1)%*%y
SSE_reduced = sum((y-X1%*%hbeta1)^2)
diff_SSR = SSE_reduced - SSE  # why?
(F_p = (diff_SSR/1) / MSE)
```

```
##          [,1]
## [1,] 15.85085
```

```
1-pf(F_p,1,n-ncol(X))      # p-value
```

```
##              [,1]
## [1,] 0.0006312469
```

Example : Delivery time data

```r
fit2 = lm(delTime~n.prod, data=delivery) # reduced model
anova(fit1,fit2)    # parial F test
```

```
## Analysis of Variance Table
##
## Model 1: delTime ~ n.prod + distance
## Model 2: delTime ~ n.prod
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     22 233.73
## 2     23 402.13 -1    -168.4 15.851 0.0006312 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

```r
c(t0^2,F_p)      # t test vs partial F test?
```

```
## distance
## 15.85085 15.85085
```

General linear hypotheses

- 계수벡터의 일반적인 선형변환에 대한 검정을 살펴보자. 즉, β에 대하여 적당한 행렬 T를 곱한 결과에 대한 검정

$$H_0 : T\beta = 0 \quad vs \quad H_1 : not\ H_0$$

- $T\hat{\beta}$의 분포로부터 검정통계량의 형태를 예측할 수 있다.

$$T\hat{\beta} \sim N(T\beta, \sigma^2 T(X^\top X)^{-1} T^\top)$$

- 위 검정을 위한 통계량은

$$F_0 = \frac{\hat{\beta}^\top T^\top [T(X^\top X)^{-1} T^\top]^{-1} T\hat{\beta}/r}{SSE/(n-k-1)}$$

  로 주어지고 귀무가설 하에서 $F_{r,n-k-1}$ 분포를 따른다. 여기서 r은 $T\beta = 0$에 의해서 정의되는 constraint의 개수이다.
- 위 통계량의 분자는 full model과 $T\beta = 0$ 조건 하에서 생성되는 reduced model사이의 회귀제곱합의 차이, 즉 extra sum of squares이다.

General linear hypotheses

- 다음과 같은 모형을 고려하자.

  $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

  만약 $\beta_1 = \beta_3$인지를 검정하고자 한다면, 이는 $T = (0, 1, 0, -1)$에 대하여

  $$H_0 : T\beta = 0 \;\; vs \;\; H_1 : not \; H_0$$

  로 표현 가능하다.
- 귀무가설 하에서 reduced model은

  $$y = \beta_0 + \beta_1(x_1 + x_3) + \beta_2 x_2 + \epsilon$$

  가 된다.

Example : Boston housing data

```
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 3.6.2
```

```
y1 = Boston$medv
X1 = model.matrix(medv ~ crim + nox + lstat, data=Boston)
T = t(c(0,1,0,-1))    # beta1 = beta3
m_f = lm(medv ~ crim + nox + lstat, data=Boston)
(beta_f = m_f$coefficients)
```

```
## (Intercept)        crim         nox       lstat
## 33.27286960 -0.07655712  2.36745067 -0.93073395
```

```
SST = sum((y1-mean(y1))^2)
SSE_f = sum((y1-m_f$fitted.values)^2)
```

Example : Boston housing data

```
m_r = lm(medv ~ I(crim  + lstat) + nox, data=Boston)
(beta_r = m_r$coefficients)
```

```
##     (Intercept) I(crim + lstat)             nox
##      32.7401536      -0.4046863      -6.5341890
```

```
SSE_r = sum((y1-m_r$fitted.values)^2)
F0 = ((SSE_r - SSE_f)/1) / (SSE_f/(nrow(X1)-ncol(X1)))
1-pf(F0,1,nrow(X1)-ncol(X1))
```

```
## [1] 0
```

```
diff_SSR = t(T%*%beta_f)%*%
  solve(T%*%solve(t(X1)%*%X1)%*%t(T))%*%T%*%beta_f
c(SSE_r - SSE_f, diff_SSR)
```

```
## [1] 5732.839 5732.839
```

Example : Boston housing data

```r
anova(m_f,m_r)
```

```
## Analysis of Variance Table
##
## Model 1: medv ~ crim + nox + lstat
## Model 2: medv ~ I(crim + lstat) + nox
##   Res.Df   RSS Df Sum of Sq     F    Pr(>F)
## 1    502 19302
## 2    503 25035 -1   -5732.8 149.1 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

회귀계수에 대한 신뢰구간

- 최소제곱추정량은 linear estimator이다 즉,

$$\hat{\beta} = Ay, \ A = (X^\top X)^{-1} X^\top$$

이고, 따라서

$$\hat{\beta}_j = \sum_{i=1}^{n} a_{ji} y_i$$

즉, 각 회귀계수의 추정량은 반응변수의 선형결합으로 표현된다.
- 모형에 대한 기본 가정에서 반응변수는 정규분포를 따르므로, 각 회귀계수 추정량도 정규분포를 따르게 된다.
- $var(\hat{\beta}) = \sigma^2 (X^\top X)^{-1}$로 부터 $(X^\top X)^{-1}$의 j번째 대각원소를 C_{jj}라 하면

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \sim t_{n-k-1}$$

회귀계수에 대한 신뢰구간

- β_j에 대한 $100(1-\alpha)\%$ 신뢰구간은

$$(\hat{\beta}_j - t_{\alpha/2,n-k-1}\sqrt{\hat{\sigma}^2 C_{jj}}\ ,\ \hat{\beta}_j + t_{\alpha/2,n-k-1}\sqrt{\hat{\sigma}^2 C_{jj}})$$

- $se(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 C_{jj}}$

Example : Delivery time data

```
est <- summary(fit1)$coefficients
C = diag(solve(t(X)%*%X))
cbind(sqrt(C*MSE), est[,2])   # standard error
```

```
##                  [,1]        [,2]
##           1.096730168 1.096730168
## n.prod    0.170734918 0.170734918
## distance  0.003613086 0.003613086
```

```
cbind(est[,1] - qt(0.975,22)*est[,2] # confidence int.
      , est[,1] + qt(0.975,22)*est[,2])
```

```
##                    [,1]       [,2]
## (Intercept) 0.066751987 4.61571030
## n.prod      1.261824662 1.96998976
## distance    0.006891745 0.02187791
```

Example : Delivery time data

```r
#cbind(est[,1] - qt(0.975,22)*est[,2]
#      , est[,1] + qt(0.975,22)*est[,2])
confint(fit1 , level = 0.95)
```

```
##                   2.5 %     97.5 %
## (Intercept) 0.066751987 4.61571030
## n.prod      1.261824662 1.96998976
## distance    0.006891745 0.02187791
```

평균 반응치에 대한 신뢰구간

- 특정 예측변수의 값 $x_0 = (1, x_{01}, x_{02}, \ldots, x_{0k})^\top$에서 평균 반응변수의 예측치는 다음과 같다.

$$\hat{y}_0 = x_0^\top \hat{\beta}$$

- $var(\hat{y}_0) = x_0^\top var(\hat{\beta}) x_0 = \sigma^2 x_0^\top (X^\top X)^{-1} x_0$ 로부터 신뢰구간 구성 가능

simultaneous confideince intervals

- 여러 회귀계수들이 포함되는 영역을 제시하는 것이다. 개별 회귀계수들이 신뢰구간의 조합으로 구성하면 전체적인 신뢰도가 하락한다. (예]내일 비올 확률 90%, 내일 안개 낄 확률 90%, 내일 비가 오고 안개가 낄 확률은?)
- 개선을 위해 크게 두 가지 접근법을 생각할 수 있다.
- 첫째는 관심있는 회귀계수들의 추정량 벡터의 분포를 이용하는 것이다. 이 경우 신뢰구간은 타원(체)의 형태로 주어지게 된다.
  - 장점 : 정확한 신뢰도를 보장하는 신뢰영역을 제시할 수 있다.
  - 단점 : 신뢰영역의 형태가 각 회귀계수에 대해서 따로 주어지기 보다는 어떤 수식에 의해 정의되고, 2차원 이상의 공간에서는 표현이 어렵다.
- 둘째는 각 회귀계수에 대한 신뢰구간이 원하는 신뢰도 이상을 만족하도록 적절히 수정하여 주는 것이다. 이 경우 신뢰구간은 각 회귀계수에 대하여 구간으로 주어진다.
  - 장점 : 적용과 신뢰구간 표현이 간단하다.
  - 단점 : 원하는 신뢰도 이상의 결합 신뢰도를 가지게 되며, 예측변수의 차원이 크면 매우 보수적이 될 수 있다.

Bonferroni 신뢰구간

- 결합 신뢰도가 최소한 원하는 수준 이상이 되도록 각 회귀계수에 대한 신뢰구간을 수정하는 방법 중 하나
- 만약 모든 회귀계수에 대한 신뢰구간이 실제 회귀계수를 포함할 확률이 최소한 $1-\alpha$가 되기를 원한다면 다음과 같이 신뢰구간을 수정한다.

$$\hat{\beta}_j \pm t_{\alpha/(k+1),n-k-1} se(\hat{\beta}_j)$$

  즉, t분포의 분위수를 $t_{\alpha,n-k-1}$에서 $t_{\alpha/(k+1),n-k-1}$로 수정하는 것이다.
- 이 방법은 다음 식으로부터 정당화될 수 있다.

$$P(A_1 \cap A_2) = 1 - P(A_1^c \cup A_2^c) \geq 1 - [P(A_1^c) + P(A_2^c)]$$

  여기서, $P(A_1^c) = P(A_2^c) = \alpha/2$로 두면, $P(A_1 \cap A_2) \geq 1-\alpha$가 보장된다.

Why do regression coefficients have the wrong sign?

- Delivery time data

```r
y <- c(1,5,3,8,5,3,10,7)
x1 <- c(2,4,5,6,8,10,11,13)
x2 <- c(1,2,2,4,4,4,6,6)
lm(y~x1)$coefficients
```

```
## (Intercept)          x1
##   1.8347935   0.4630788
```

```r
lm(y~x1+x2)$coefficients
```

```
## (Intercept)          x1          x2
##    1.035506   -1.222276    3.649319
```

Why do regression coefficients have the wrong sign?

- Delivery time data

```
plot(x1,y,col=x2,cex=3,pch=x2)
```