통계학의 이해 II

다중회귀모형의 표현

다중회귀모형의 표현

- 설명변수가 여러 개인 경우 회귀모형을 벡터와 행렬로 표시하는 방법을 알아본다.
- 최소제곱추정량의 형태와 성질을 알아본다.

다중선형회귀모형

관계식 가정

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad \varepsilon_i \sim iid\ N(0, \sigma^2)$$

- 설명변수가 2개 이상인 선형회귀모형
- β_j: 다른 변수를 고정시키고 j번째 변수를 1증가시킬 때 Y의 평균 증가량

◇ 표시: 행렬 & 벡터

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad \varepsilon_i \sim iid\ N\left(0, \sigma^2\right)$$

- 회귀계수: $\beta = \left(\beta_0, \beta_1, \dots, \beta_p\right)^T$
- 설명변수: $x_i = \left(1, x_{i1}, \dots, x_{ip}\right)^T$

$$\Rightarrow Y_i = x_i^T \beta + \varepsilon_i = \beta^T x_i + \varepsilon_i$$

- 반응변수: $Y = (Y_1, Y_2, \dots, Y_n)^T$
- 오차: $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)^T$
- 설명변수: $X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & & x_{2p} \\ & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}$

$$\Rightarrow Y = X\beta + \varepsilon$$

◇ 회귀계수 추정: 최소제곱법

$$D(\boldsymbol{\beta}) = \sum \boldsymbol{\varepsilon}_i^2 = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = (\boldsymbol{Y} - \boldsymbol{X\beta})^T(\boldsymbol{Y} - \boldsymbol{X\beta})$$

- 최소제곱추정량: $\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$
- $E\left(\widehat{\boldsymbol{\beta}}\right) = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T E(\boldsymbol{Y}) = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{X}\,\boldsymbol{\beta} = \boldsymbol{\beta}$
- $Cov\left(\widehat{\boldsymbol{\beta}}\right) = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\, Cov(\boldsymbol{Y})\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} = \boldsymbol{\sigma}^2\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$

$$\Rightarrow Var\left(\widehat{\boldsymbol{\beta}}_j\right) = \boldsymbol{\sigma}^2\, c_{jj}$$

  - c_{jj}: $\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$의 j번째 대각원소

다중회귀모형의 표현

- 설명변수가 여러 개인 경우 회귀모형을 벡터와 행렬로 표시하는 방법을 알아본다.
  - $Y_i = x_i^T \beta + \varepsilon_i = \beta^T x_i + \varepsilon_i$
  - $Y = X\beta + \varepsilon$
- 최소제곱추정량의 형태와 성질을 알아본다.
  - $\hat{\beta} = \left(X^T X\right)^{-1} X^T Y \sim N_{p+1}\left(0, \sigma^2 \left(X^T X\right)^{-1}\right)$

통계학의 이해 II

분산분석과 t-검정

분산분석과 t-검정

- 회귀모형의 유효한 모형인지를 검정하는 방법을 알아본다.
- 유효한 모형이라고 했을 때 어떤 회귀계수가 유의한가를 확인하는 방법을 알아본다.

모형의 유의성

복습】 분산분석

- $Y_{ij} - \mu_i = \varepsilon_{ij},\ \varepsilon_{ij} \sim iid\ N(0, \sigma^2)$
  - $Y_{ij} = \mu_i + \varepsilon_{ij} = \mu + (\mu_i - \mu) + \varepsilon_{ij} = \mu + \alpha_i + \varepsilon_{ij}$
- $H_0 : \mu_1 = \cdots = \mu_p \Rightarrow H_0 : \alpha_1 = \cdots = \alpha_p = 0$
- 검정통계량:

$$F_0 = \frac{\sum_{i=1}^{p} n_i(\bar{Y}_i - \bar{Y})^2/(p-1)}{\sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_i\right)^2/\sum_{i=1}^{p}(n_i - 1)} \sim F_{p-1, N-p}$$

- 분산분석에서는 μ_i를 i번째 수준의 표본평균으로 추정

◇ 회귀분석에서 모든 설명변수가 설명력이 없다는 것은 모든 반응변수의 평균이 같음

- $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$

  $\Rightarrow E(Y_i) = \mu_i$라고 하면 $H_0: \mu_1 = \mu_2 = \cdots = \mu_n = \beta_0$

- 회귀분석에서는 μ_i를 $\hat{Y}_i = x_i^T \hat{\beta}$으로 추정

- 검정통계량:

$$F_0 = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 / p}{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 / (n-p-1)} \sim F_{p,n-p-1}$$

☑ 예제】 단순선형모형

- H_0: $\beta_1 = 0$
- $T_0 = \dfrac{\hat{\beta}_1}{\sqrt{MSE/S_{xx}}} \sim t_{n-2} \quad \Rightarrow T_o^2 = \dfrac{S_{xx}\hat{\beta}_1^2}{MSE} \sim F_{1,n-2}$
- $\left(\hat{Y}_i - \bar{Y}\right)^2 = \left(\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{Y}\right)^2 = \hat{\beta}_1^2(x_i - \bar{x})^2$

$$\Rightarrow T_o^2 = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2/(n-2)} \sim F_{1,n-2}$$

◇ 각 회귀계수에 대한 추론

- $\hat{\beta}_j \sim N(\beta_j, \sigma^2 c_{jj})$, $\quad c_{jj}$: $(X^T X)^{-1}$의 j번째 대각원소
- 중심축량:

$$T = \frac{\hat{\beta}_j - \beta_j}{\sqrt{MSE}\sqrt{c_{jj}}} \sim t_{n-p-1}$$

- $100(1-\alpha)\%$ 신뢰구간: $\hat{\beta}_j \pm t_{\frac{\alpha}{2},n-2}\sqrt{MSE}\sqrt{c_{jj}}$
- 검정통계량: H_0: $\beta_j = \beta_j^*$

$$T_0 = \frac{\hat{\beta}_j - \beta_j^*}{\sqrt{MSE}\sqrt{c_{jj}}} \sim t_{n-p-1}$$

  ○ 유의하지 않는 회귀계수는 모형에서 제외하는 것이 parsimony (모수절약) 차원에서 권장

◇ 예제】 시멘트 성분과 발생 열량(Woods, Steinour & Starke, 1932)

- 4개 성분비율과 180일 후의 g당 열량, 16개 자료
- 분산분석표(R 기준)

```
Analysis of Variance Table

Response: Y
          Df  Sum Sq Mean Sq F value    Pr(>F)
X2         1 1900.14 1900.14  282.26 3.384e-10 ***
X4         1 1317.89 1317.89  195.77 3.247e-09 ***
Residuals 13   87.52    6.73

Residual standard error: 2.595 on 13 degrees of freedom
Multiple R-squared:  0.9735,    Adjusted R-squared:  0.9695
F-statistic:   239 on 2 and 13 DF,  p-value: 5.604e-11
```

☑ T-검정

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 52.18516    2.45670   21.24 1.77e-11 ***
X2           1.48017    0.10839   13.66 4.37e-09 ***
X4           0.67754    0.04842   13.99 3.25e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

분산분석과 t-검정

- 회귀모형의 유효한 모형인지를 검정하는 방법을 알아본다.
  - $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$
  - 검정통계량: $F_0 = \dfrac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 / p}{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 / (n-p-1)} \sim F_{p,n-p-1}$
- 유효한 모형이라고 했을 때 어떤 회귀계수가 유의한가를 확인하는 방법을 알아본다.
  - β_j에 대한 추론
  - 중심축량: $T = \dfrac{\hat{\beta}_j - \beta_j}{\sqrt{MSE}\sqrt{c_{jj}}} \sim t_{n-p-1}$

통계학의 이해 II

다중회귀모형에서의 주요 문제

다중회귀모형에서의 주요 문제

- 여러 개 설명변수가 있는 모형에서 발생할 수 있는 문제와 이를 해결하는 방법에 대해 알아본다.

다중공선성(Multicollinearity)

- 설명변수들 간 선형관계가 존재
  - $(X^T X)^{-1}$이 존재하지 않거나 일부 대각원소가 상당히 커짐
- 다중공선성이 있는 경우 현상
  - 추정된 회귀계수의 값이나 부호가 상식적이지 않음
  - 중요하다고 생각되는 변수가 유의하지 않게 나옴
  - 설명변수가 약간만 변해도 회귀계수가 크게 변함
  - 관측치가 하나만 추가되거나 제거되어도 회귀계수가 크게 변함

확인하는 방법

- VIF(분산팽창계수): $VIF_j = \frac{1}{1-R_j^2}$
    - ○ R_j^2: j번째 변수를 반응, 나머지를 설명변수로 설정한 모형의 변동계수
    - ○ VIF가 10 이상이면 다른 변수와 선형관계가 있는 것으로 의심
- 상태수(condition number, 조건수), 공차한계(tolerance), …

해결방법

- 변수선택: 설명변수들 중 불필요한 변수를 모형에서 제거
- 주성분회귀분석, …

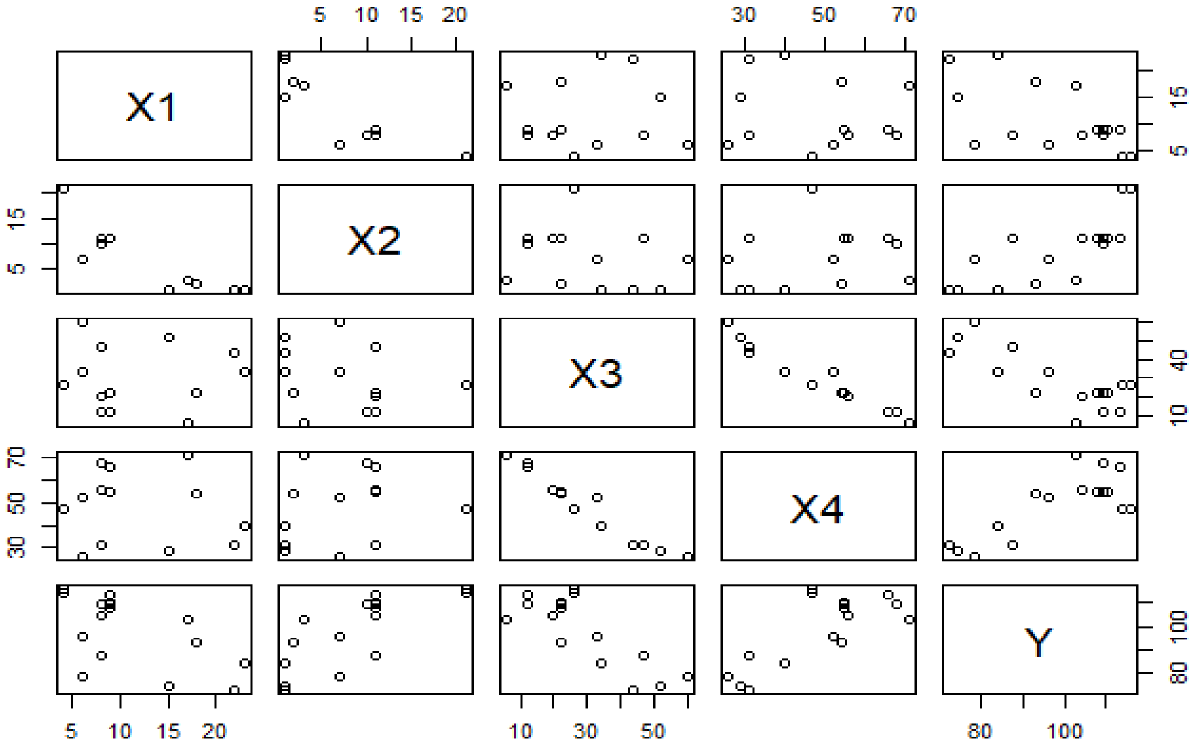
◇ 예제】 시멘트 성분과 발생 열량(Woods, Steinour & Starke, 1932)

☑ 4개 성분비율과 180일 후의 g당 열량, 16개 자료

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  85.8969    75.1575   1.143    0.277
X1           -0.1751     0.8028  -0.218    0.831
X2            1.2757     0.7801   1.635    0.130
X3           -0.3798     0.7615  -0.499    0.628
X4            0.2875     0.7791   0.369    0.719

Residual standard error: 2.664 on 11 degrees of freedom
Multiple R-squared:  0.9764,	Adjusted R-squared:  0.9678 
F-statistic: 113.7 on 4 and 11 DF,  p-value: 7.2e-09
```

- **VIF: X1(50.57), X2(51.18), X3(284.34), X4(255.73)**


X1
X2
X3
X4
Y

상대적 영향력

- $\hat{\beta}$는 해당변수의 관측척도(scale)에 영향을 받음
  - j번째 변수의 척도를 10배 크게 하면 $\hat{\beta}_j$는 10배 작아짐
  - $\hat{\beta}$ 값이 크다(작다)고 해당변수가 반응변수에 영향을 더(덜) 준다고 할 수 없음
- 반응변수와 각 설명변수를 표준화하여 회귀분석 실시
  - 변수의 척도에 영향을 받지 않음
  - 모든 변수를 표준화하면 절편은 모형에서 제외 됨

$$Y_i^* = \beta_1 x_{i1}^* + \cdots + \beta_p x_{ip}^* + \varepsilon_i$$

- $x_{ij}^* = \dfrac{x_{ij} - \bar{x}_j}{s_j}$

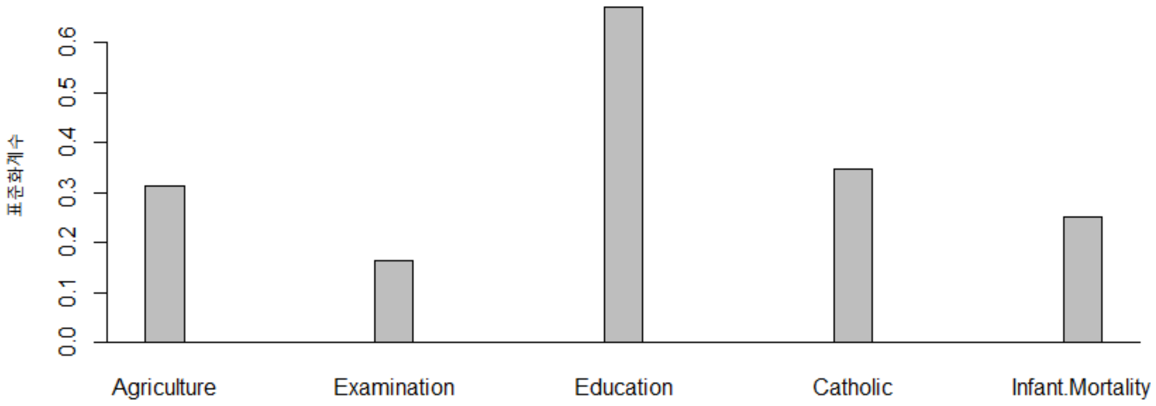
◇ 예제】 Swiss Fertility and Socioeconomic Indicators (1888) Data

- Fertility와 5개 변수 간의 관계 유도, 47개 관측값

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       66.91518   10.70604   6.250 1.91e-07 ***
Agriculture       -0.17211    0.07030  -2.448  0.01873 *
Examination       -0.25801    0.25388  -1.016  0.31546
Education         -0.87094    0.18303  -4.758 2.43e-05 ***
Catholic           0.10412    0.03526   2.953  0.00519 **
Infant.Mortality   1.07705    0.38172   2.822  0.00734 **
```

☑ 표준화 계수

- Agriculure: -0.313, Examination: -0.165, Education: -0.670
  Catholic: 0.348, Infant.Mortality: 0.251



설명변수에 범주형 변수가 포함

- 공분산분석(ANCOVA) 모형: 설명변수에 공변량과 요인이 모두 포함
- 가변수(dummy variable) 사용
  - 두 범주: A, B ⇒ A이면 $x_{iA} = 1$, B이면 $x_{iA} = 0$
  - 세 범주: A, B, C

    ⇒ A: $(x_{iA}, x_{iB}) = (1, 0)$, B: $(x_{iA}, x_{iB}) = (0, 1)$, C: $(x_{iA}, x_{iB}) = (0, 0)$
  - 예제】 두 범주

$$Y_i = \beta_0 + \beta_1 x_i + \beta_A x_{iA} + \beta_I x_i x_{iA} + \varepsilon_i$$

    ○ β_A: A범주의 절편이 B범주보다 β_A만큼 큼

    ○ β_I: A범주의 기울기가 B범주보다 β_I만큼 큼($x_i x_{iA}$: 상호작용)

◇ 예제】 올림픽 육상 100m 우승기록

- 1990년~2016년 자료 (1986년 제외)
- 남녀간 절편 또는 기울기에 차이가 있는가?

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  40.543225   2.785297  14.556  < 2e-16 ***
genderM      -9.003392   3.317439  -2.714  0.00946 ** 
year         -0.014878   0.001410 -10.550 1.25e-13 ***
genderM:year  0.004020   0.001683   2.388  0.02128 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1681 on 44 degrees of freedom
Multiple R-squared:  0.9314,    Adjusted R-squared:  0.9267 
F-statistic: 199.1 on 3 and 44 DF,  p-value: < 2.2e-16
```

○ 남자 기록의 기울기: -0.0149-0.0040 = -0.0109

다중회귀모형에서의 주요 문제

- 여러 개 설명변수가 있는 모형에서 발생할 수 있는 문제와 이를 해결하는 방법에 대해 알아본다.
  - 다중공선성
    - 확인방법: VIF, condition number, tolerance, …
    - 해결방법: 변수선택, …
  - 상대적 영향력
    - 표준화 회귀모형
  - 설명변수에 범주형 변수가 포함
    - 가변수로 변환 후 분석

통계학의 이해 II

변수선택

변수선택

- 모수절약의 원칙에서 필요한 변수를 선택하거나 불필요한 변수를 제거하는 방법을 알아본다.

변수선택(Variable Selection)

- 불필요한 변수가 추가되는 경우 다중공선성의 가능성이 높아지거나 추정의 정밀도가 낮아짐
  - 복잡한 모형은 해석이 어려움: 모수절약의 원칙(parsimony)
- p개의 설명변수가 있는 경우 비교 가능한 모형의 수: $2^p - 1$
  - 고전적인 변수선택법
    - 전진선택법(forward selection), 후진제거법(backward elimination), 단계적방법(stepwise method), …
    - 선택제거의 기준: t-검정(F-검정), AIC
  - LASSO, …

전진선택법

- β_0만 있는 모형에서 시작
- 설명변수를 추가해 가며 AIC가 가장 작은 모형을 선택
- 추가만 하기 때문에 모형에 한번 포함된 변수는 제외되지 않음

후진제거법

- 모든 변수가 포함되어 있는 모형에서 시작
- 설명변수를 제거하며 AIC가 가장 작은 모형을 선택
- 한번 제거된 변수는 모형에 다시 포함되지 않음

단계적 방법

- β_0만 있는 모형에서 시작
- 설명변수를 추가, 제거를 해 가며 AIC가 가장 작은 모형을 선택
- 모형에 포함되었던 변수가 제거되기도하고 제거된 변수가 포함되기도 함

◇ 예제】 시멘트 성분과 발생 열량(Woods, Steinour & Starke, 1932)

☑ 전진선택법

```
Start:  AIC=87.29   (Y ~ 1)
Step:  AIC=72.87   (Y ~ X3)
Step:  AIC=37.21   (Y ~ X3 + X2)
Step:  AIC=33.43   (Y ~ X3 + X2 + X4)

       Df Sum of Sq    RSS    AIC
<none>                78.419 33.432
+ X1    1   0.33759 78.082 35.363
```

✓ 후진제거법

```
Start:  AIC=35.36   (Y ~ X1 + X2 + X3 + X4)
Step:  AIC=33.43    (Y ~ X2 + X3 + X4)
Step:  AIC=33.19    (Y ~ X2 + X4)
       Df Sum of Sq      RSS    AIC
<none>                 87.52 33.188
- X2    1    1255.5 1342.98 74.881
- X4    1    1317.9 1405.41 75.608
```

✓ 단계적 방법

```
Start:  AIC=87.29   (Y ~ 1)
Step:  AIC=72.87    (Y ~ X3)
Step:  AIC=37.21    (Y ~ X3 + X2)
Step:  AIC=33.43    (Y ~ X3 + X2 + X4)
Step:  AIC=33.19    (Y ~ X2 + X4)
       Df Sum of Sq      RSS    AIC
<none>                 87.52 33.188
+ X3    1      9.10   78.42 33.432
+ X1    1      7.67   79.85 33.721
- X2    1   1255.46 1342.98 74.881
- X4    1   1317.89 1405.41 75.608
```

변수선택

- 모수절약의 원칙에서 필요한 변수를 선택하거나 불필요한 변수를 제거하는 방법을 알아본다.

  - 고전적인 변수선택법
    - 전진선택법(forward selection),
      후진제거법(backward elimination),
      단계적방법(stepwise method), …
    - 선택제거의 기준: t-검정(F-검정), AIC
  - LASSO, …

통계학의 이해 II

강의정리 및 실습

회귀모형의 형태

- 설명변수가 여러 개인 경우 회귀모형을 벡터와 행렬로 표시하는 방법을 알아본다.
  - $Y_i = x_i^T \beta + \varepsilon_i = \beta^T x_i + \varepsilon_i$
  - $Y = X\beta + \varepsilon$
- 최소제곱추정량의 형태와 성질을 알아본다.
  - $\hat{\beta} = (X^T X)^{-1} X^T Y \sim N_{p+1}(0, \sigma^2 (X^T X)^{-1})$

분산분석과 t-검정

- 회귀모형의 유효한 모형인지를 검정하는 방법을 알아본다.
  - $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$
  - 검정통계량: $F_0 = \dfrac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 / p}{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 / (n-p-1)} \sim F_{p,n-p-1}$

- 유효한 모형이라고 했을 때 어떤 회귀계수가 유의한가를 확인하는 방법을 알아본다.
  - β_j에 대한 추론
  - 중심축량: $T = \dfrac{\hat{\beta}_j - \beta_j}{\sqrt{MSE}\sqrt{c_{jj}}} \sim t_{n-p-1}$

다중회귀모형에서의 주요 문제

- 여러 개 설명변수가 있는 모형에서 발생할 수 있는 문제와 이를 해결하는 방법에 대해 알아본다.
  - 다중공선성
    - 확인방법: VIF, condition number, tolerance, …
    - 해결방법: 변수선택, …
  - 상대적 영향력
    - 표준화 회귀모형
  - 설명변수에 범주형 변수가 포함
    - 가변수로 변환 후 분석

변수선택

- 모수절약의 원칙에서 필요한 변수를 선택하거나 불필요한 변수를 제거하는 방법을 알아본다.
  - 고전적인 변수선택법
    - 전진선택법(forward selection), 후진제거법(backward elimination), 단계적방법(stepwise method), …
    - 선택제거의 기준: t-검정(F-검정), AIC
  - LASSO, …

R 실습

- 다중공선성: vif(car)
- 변수선택: step
- 표준화회귀분석: lm.beta(QuantPsyc)

과제

- 올림픽 100m 우승기록(1900년부터 2004년까지 자료)
  - 남녀간 기울기에 차이가 있는지 확인하고 해당 모형을 근거로 기록의 역전이 언제 일어나는지를 구하여라.

- "cement28.csv"자료를 이용하여
  - 다중공선성을 확인하여라.
  - 세가지 변수선택방법을 적용하여 최종모형을 비교하여라.
  - 단계적 방법에 의한 모형을 근거로 어떤 변수가 더 많이 영향을 주는지 표준화 회귀계수로 비교하여라.