김은영

제3장 중선형회귀모형

3.8 기타논제

3.8.1 회귀계수의 표준화

회귀모형에서 회귀계수 β_j가 갖는 단위는 $\dfrac{\text{종속변수 } y\text{의 단위}}{\text{독립변수 } X_j\text{의 단위}}$로 주어지므로

회귀계수의 의미는 항상 두 변수의 단위와 함께 해석되어야 한다.

회귀계수 간의 비교를 쉽게 하기 위해서 단위가 없는 회귀계수를 사용할 수 있다.
이를 위해 단위에 따라 변하지 않는 변환(scale-invariant transformation)이 필요하다.

$$y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i$$

$$y_i^* = \frac{y_i - \bar{y}}{\sqrt{s_{yy}}} \qquad \text{여기서, } s_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$w_{ij} = \frac{X_{ij} - \bar{X}_j}{\sqrt{s_{jj}}} \qquad \text{여기서, } s_{jj} = \sum_{i=1}^{n} (X_{ij} - \bar{X}_j)^2 \qquad (j = 1, \cdots, p-1)$$

$$y^* = \begin{bmatrix} y_1^* \\ \vdots \\ y_n^* \end{bmatrix}, \quad W = \begin{bmatrix} w_{11} \cdots w_{1,p-1} \\ \vdots \qquad \vdots \\ w_{n1} \cdots w_{n,p-1} \end{bmatrix}, \quad \beta^* = \begin{bmatrix} \beta_1^* \\ \vdots \\ \beta_{p-1}^* \end{bmatrix}$$

$$y^* = W\beta^* + \epsilon$$

$$\hat{\beta}^* = \underset{\beta^*}{\arg\min} (y^* - W\beta^*)^t (y^* - W\beta^*) = (W^t W)^{-1} W^t y^*$$

$$W = (w_{ij}) \qquad (i = 1, \cdots, n;\ j = 1, \cdots, p-1)$$

$W^t W$ 의 jl번째 원소는 $(W^t W)_{jl} = \sum_{i=1}^{n} W_{ij} W_{il} = \sum_{i=1}^{n} \frac{(X_{ij} - \bar{X}_j)}{\sqrt{s_{jj}}} \frac{(X_{il} - \bar{X}_l)}{\sqrt{s_{ll}}}$

$$= \frac{\sum_{i=1}^{n} (X_{ij} - \bar{X}_j)(X_{il} - \bar{X}_l)}{\sqrt{\sum_{i=1}^{n} (X_{ij} - \bar{X}_j)^2 \sum_{i=1}^{n} (X_{il} - \bar{X}_l)^2}} \equiv \gamma_{jl} \quad \rightarrow X_j\text{와 } X_l \text{ 의 표본상관계수}$$

$$\Rightarrow \; W^t W = \begin{bmatrix} 1 & \gamma_{12} & \cdots & \gamma_{1,p-1} \\ & 1 & \cdots & \gamma_{2,p-1} \\ & & \ddots & \vdots \\ & & & 1 \end{bmatrix}$$

$$W^t y^* = \begin{bmatrix} \gamma_{y1} \\ \vdots \\ \gamma_{y,p-1} \end{bmatrix} \quad \rightarrow \gamma_{y,j}\text{: } y^*\text{와 } X_j \text{ 의 표본상관계수}$$

※ 정규방정식의 각 원소가 상관계수로 −1과 1 사이의 값을 가지게 되어 이들을 사용하면 계산오차는 상대적으로 줄어든다.

3.8.2 다중공선성

중회귀모형에서 설명변수들 간에 선형종속 관계가 존재하지 않더라도 변수들 간의 상관관계가 높은 경우에는 회귀분석의 추론과정에서 몇 가지 문제가 발생한다.

[예] 두 개의 설명변수가 있는 모형에서 각 변수가 표준화되고
두 설명변수 간의 상관계수를 γ_{12}라고 하면

$$\hat{\beta}^* = \underset{\beta^*}{\arg\min} (y^* - W\beta^*)^t (y^* - W\beta^*) = (W^t W)^{-1} W^t y^*$$

$$= \begin{bmatrix} 1 & \gamma_{12} \\ \gamma_{12} & 1 \end{bmatrix}^{-1} \begin{bmatrix} \gamma_{y1} \\ \gamma_{y2} \end{bmatrix} = \frac{1}{1-\gamma_{12}^2} \begin{bmatrix} 1 & -\gamma_{12} \\ -\gamma_{12} & 1 \end{bmatrix} \begin{bmatrix} \gamma_{y1} \\ \gamma_{y2} \end{bmatrix}$$

■ 만약 X_1과 X_2 간의 상관관계가 높으면 γ_{12}는 ±1에 가까운 값을 가지게 되고, 행렬식의 값은 0에 가깝게 되어, 행렬 $W^t W$ 는 비정칙행렬이 될 수 있다.

■ 표준화 회귀계수 $\hat{\beta}^*$의 분산도 $(1-\gamma_{12}^2)^{-1}$의 함수로 주어지므로 만약 γ_{12}는 ±1에 가까운 값을 가지면 이들 분산이 매우 커지게 되어 회귀계수에 대한 추정의 정확도가 낮아지게 된다.

■ 설명변수들 간의 상관관계가 높으면 최소제곱추정량의 계산이 불가능할 수 있고, 추정량의 분산이 커지는 문제가 발생할 수 있는데, 이때 설명변수들 간에 '다중공선성(multicollinearity)'이 존재한다고 한다.

■ 다중공선성은 설명변수들 간의 상관관계가 높거나, 설명변수들 간에 선형종속 관계가 있는 경우에 발생한다고 할 수 있다.