Cov matrix는 왜 안하나!

* Scale의 영향을 받아서 선형관계정도를 판단하는게 어려워서?

* Multicollinearity in multiple regression data

- Correlation matrix

$X = (\underset{\sim}{1}, \underset{\sim}{x_1}, \underset{\sim}{x_2}, \cdots, \underset{\sim}{x_k})$ $n \times p$ matrix

$\downarrow$ centering & Scaling $\quad x_{ij} \Rightarrow x_{ij}^* = \dfrac{x_{ij} - \bar{x}_j}{\sqrt{\sum_{i=1}^{n} (x_{ij} - \bar{x})^2}}$

*centering → mean이 '0'이되게
scaling → 길이가 '1'이되게
∴ 한 변수에대한 값들의 제곱합 = 1
$\sum_{i=1}^{n} x_{i1}^{*2} = 1$

$$X = \begin{pmatrix} \frac{1}{\sqrt{n}} & x_{11}^* & x_{12}^* & \cdots & x_{1k}^* \\ \frac{1}{\sqrt{n}} & x_{21}^* & x_{22}^* & \cdots & x_{2k}^* \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{n}} & x_{n1}^* & x_{n2}^* & \cdots & x_{nk}^* \end{pmatrix} = \left(\frac{1}{\sqrt{n}}\underset{\sim}{1}, \underset{\sim}{x_1^*}, \underset{\sim}{x_2^*}, \cdots, \underset{\sim}{x_k^*}\right)$$

$\uparrow$

y는 centering 하지않았기 때문에 상수항의 보정이 필요

$y = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \cdots + \beta_k x_k^* + \varepsilon$

$\bar{y} \neq 0$ But $\bar{x}_1^* = \bar{x}_2^* = \cdots = \bar{x}_k^* = 0$

$\therefore \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1^* - \cdots - \hat{\beta}_k \bar{x}_k^*$

$= \bar{y}$

symmetric

$$X'X = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & r_{21} & \cdots & r_{k1} \\ \vdots & r_{12} & 1 & \cdots & r_{k2} \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & r_{1k} & r_{2k} & \cdots & 1 \end{pmatrix}$$

$X^{*\prime}X^*$

: x들의 correlation matrix

(eg) $(X'X)_{23} = \underset{\sim}{x_1^{*\prime}}\underset{\sim}{x_2^*} = \sum_{i=1}^{n} \dfrac{x_{i1} - \bar{x}_1}{\sqrt{\sum (x_{i1} - \bar{x}_1)^2}} \cdot \dfrac{x_{i2} - \bar{x}_2}{\sqrt{\sum (x_{i2} - \bar{x}_2)^2}}$

$= \widehat{\mathrm{corr}}(x_1, x_2)$

$= r_{12}$

$\therefore (X'X)_{ij} = \underset{\sim}{x_{i-1}^{*\prime}}\underset{\sim}{x_{j-1}^*} = r_{i-1, j-1}$

- Problem of Multicollinearity

(그러니까 회귀계수의 추정이 잘되지 않음.)

: x들의 correlation이 높으면, coefficient 들의 variance에 inflation이 일어난다.

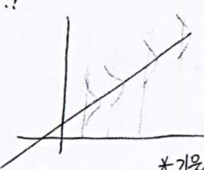
if $\begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.99215 \\ 0.99215 & 1 \end{bmatrix} \xrightarrow{\text{Inverse}} \underbrace{\begin{bmatrix} 63.94 & -63.44 \\ -63.44 & 63.94 \end{bmatrix}}_{(X^{*\prime}X^*)^{-1}}$

$\hat{\underset{\sim}{\beta}} \sim (\underset{\sim}{\beta}, (X'X)^{-1}\sigma^2)$ 이니까!

$(X^{*\prime}X^* = I)$ 일때 대비 약 64배!!

⇒ 따라서 회귀계수 자체에 관심이 있다면, (y와 x사이의 관계). 치명적인 문제

But. ŷ에 관심이 있다면, 크게 문제가 되지 않음.



⇒ 이런 Multicollinearity 의 정도를 나타내는 것이 VIF (Variance Inflation Factor)

*기울기가 불안정해도 꼭지점의 ŷ는 변화가 거의 X

— VIF

: 한 변수(x_j)가 나머지 다른변수들과 선형관계를 가지는 것때문에 $\hat{\beta}_j$의 variance가 얼마나 증가한 것인가를 보여줌.

↔ j-th 변수가 나머지 변수들과 linearly independent 또는 orthogonal 할때 $Var(\hat{\beta}_j)=1$ 이라하면, VIF_j는 그것들의 선형관계에의해 몇배나 증가하는가.

$$VIF_j = \frac{1}{1-R_j^2}$$

R_j^2: '$x_j^* = \beta_1 x_1^* + \cdots + \beta_{j-1} x_{j-1}^* + \beta_{j+1} x_{j+1}^* + \cdots + \beta_k x_k^* + \varepsilon$' 에서의 R^2 (모든 x^*들은 centering 했기때문에 상수항이 없다)

* VIF 구하기 - 1

i) $X^* = (\underset{\sim}{x_1}^*, \underset{\sim}{x_2}^*, \cdots, \underset{\sim}{x_k}^*)$: $n \times k$ matrix

$= (\underset{\sim}{x_1}^*, X_1^*)$ — X_1은 제외한 나머지 $k-1$개의 독립변수

$$X^{*\prime}X^* = \begin{pmatrix} \underset{\sim}{x_1}^{*\prime} \\ X_1^{*\prime} \end{pmatrix} (\underset{\sim}{x_1}^*, X_1^*) = \begin{pmatrix} \underset{\sim}{x_1}^{*\prime}\underset{\sim}{x_1}^* & \underset{\sim}{x_1}^{*\prime}X_1^* \\ X_1^{*\prime}\underset{\sim}{x_1}^* & X_1^{*\prime}X_1^* \end{pmatrix}$$

(cf) $C = \begin{pmatrix} A & B \\ B & D \end{pmatrix}$ symmetric 이니까

$C^{-1} = \frac{1}{AD-BC}\begin{pmatrix} D & -B \\ -C & A \end{pmatrix}$

$\therefore C^{-1}_{11} = (A - BD^{-1}B')^{-1}$

ii) $$(X^{*\prime}X^*)^{-1}_{11} = (\underset{\sim}{x_1}^{*\prime}\underset{\sim}{x_1}^* - \underset{\sim}{x_1}^{*\prime}\underbrace{X_1^*(X_1^{*\prime}X_1^*)^{-1}X_1^{*\prime}}_{H}\underset{\sim}{x_1}^*)^{-1}$$

$\because H = X^*(X^{*\prime}X^*)^{-1}X^{*\prime}$

$$= (\underset{\sim}{x_1}^{*\prime}(I-H)\underset{\sim}{x_1}^*)^{-1}$$

$$= \left(\frac{\underset{\sim}{x_1}^{*\prime}(I-H)\underset{\sim}{x_1}^*}{\underset{\sim}{x_1}^{*\prime}\underset{\sim}{x_1}^*}\right)^{-1}$$

$\because \underset{\sim}{x_1}^{*\prime}\underset{\sim}{x_1}^* = 1$ (scaled! 제곱합 1)

$$= \left(\frac{SSE}{SST}\right)^{-1}$$

$\underset{\sim}{x_1}^* = X_1^*\underset{\sim}{\gamma} + \varepsilon$ 이 모델의 SSE

$\because SST = \sum(x_1^* - \bar{x}_1^*)^2 = \sum x_1^{*2} = \underset{\sim}{x_1}^{*\prime}\underset{\sim}{x_1}^*$

$\bar{x}_1^* = 0$: centered, $= 1$: scaled

$$= \frac{1}{1-R_1^2}$$

iii) $$Var(\hat{\beta}_1) = (X'X)^{-1}_{11}\sigma^2$$
$$= (X^{*\prime}X^*)^{-1}_{11}\sigma^2$$
$$= VIF_1 \cdot \sigma^2$$

$\therefore$ VIF의 뜻이 $Var(\hat{\beta}_j) = \sigma^2$ 일때 대비 몇배나 증가하느냐!

$s.e(\hat{\beta}_1)^2 / MSE = VIF_1$?

<Remark>

- 모든 변수에 대해서 VIF를 계산

  만약 변수 x_j^*가 다른 변수들에 의해 설명이 많이된다면, $Var(\hat{\beta}_j)$의 값이 커짐.

- VIF > 10 이면 Multicollinearity 의 문제가 존재한다고 한다.