Multiple Linear Regression

- Target (Dependent variable) : Y
- Feature (Independent variable) : X
- 목표 : $Y = F(X)$ Find (F)
  - linear relationship Find (Y, X)
  - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$ (noise)

Find $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_n x_n$

- OLS (Ordinary least square)
  - ⇒ Minimize the squared difference between $Y, \hat{Y}$.

- OLS solve.

X : $n \times (d+1)$ matrix ⇒ data.

y : $n \times 1$ vector ⇒ actual value

$\hat{\beta}$: $(d+1) \times 1$ vector ⇒ parameter

$$\min \frac{1}{2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min \frac{1}{2} (y - X\hat{\beta})^T (y - X\hat{\beta})$$

$$\Rightarrow (y - X\hat{\beta})' (y - X\hat{\beta}) = y'y - \hat{\beta}'X'y - y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta}$$

$$= y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta}$$

$$\Rightarrow \frac{\partial E(\hat{\beta})}{\partial \hat{\beta}} = -2X'y + X'X\hat{\beta} + \hat{\beta}'X'X$$

$$= -2X'y + 2X'X\hat{\beta}$$

$$\Rightarrow -2X'y + 2X'X\hat{\beta} = 0$$

$$\Rightarrow \quad 2X'X\hat{\beta} = 2X'y$$

$$\Rightarrow \quad (X'X)\hat{\beta} = X'y$$

$$\Rightarrow \quad \hat{\beta} = (X'X)^{-1}X'y.$$

↳ Unique, explicit solution

✱ 4가지 조건 (OLS 의 해가 좋게)

① ε 는 normal Distribution

② Homoskedasticity.

③ linear relationship is correct.

④ independent.

Check 1

→ QQ-plot. Residual

Check 2.

→ Residual plot.

- 평가방법!

R^2

Sum-of-Squares Decomposition

$$\underbrace{\sum_{j=1}^{n}(y_j-\bar{y})^2}_{\text{total sum of squares about mean}}=\underbrace{\sum_{j=1}^{n}(\hat{y}_j-\bar{y})^2}_{\text{regression sum of squares}}+\underbrace{\sum_{j=1}^{n}\hat{\varepsilon}_j^2}_{\text{residual (error) sum of squares}}.$$

SST SSR SSE ← SSE가 OLS의 목

Total variance Explained variance

$$R^2 = 1 - \frac{SSE}{SST}$$

SST SSE

⇒ 1 - 전체 데이터에서의 변동성과 회귀식에서 설명되는 변동성의 비율

Adjusted R^2

$$R^2_{Adj} = 1 - \left[\frac{n-1}{n-(p+1)}\right]\frac{SSE}{SST} \leq R^2$$

(p ↳ 변수의 수.)

~> 변수가 늘어나면 무조건 커지는 R^2의 한계를 극복하기 위해 (쓸모 없는 변수 줄이기 위해)

* 다중 공산성.

$$C = 1 - A - B$$

* p-value

$H_0 : x = 0$ (귀무가설)

$H_1 : x \neq 0$ (대립가설)