회귀분석팀

6팀

심은주
진수정
문병철
이수정
임주은

INDEX

- **회귀분석의 정의**

둘 또는 그 이상의 변수들 간의 인과관계를 파악하고, 이를 통해 특정 변수의 값을 다른 변수들을 이용하여 설명하고 예측하는 분석

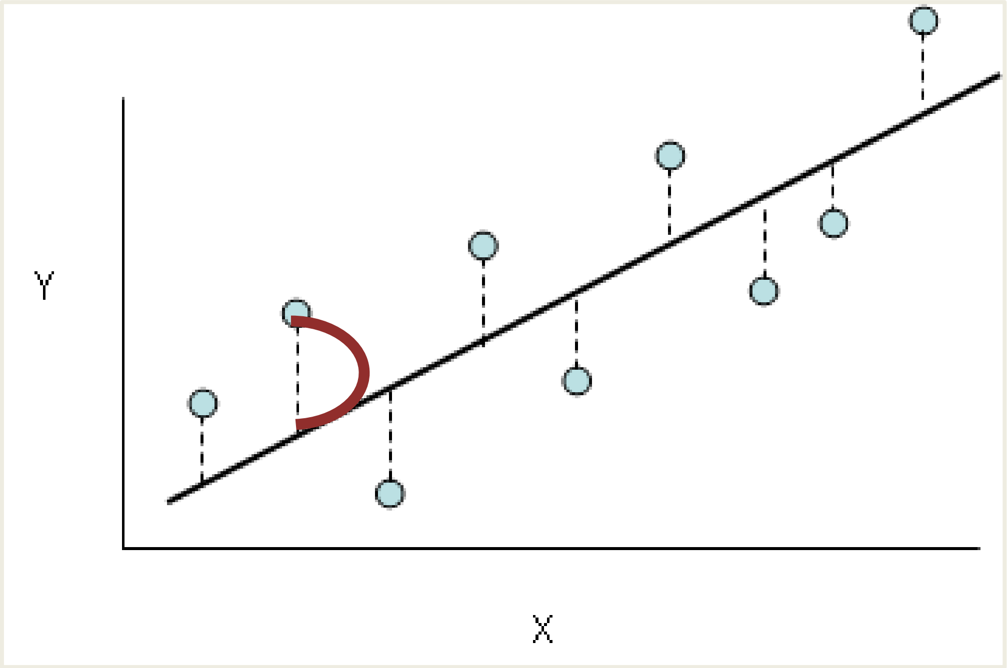
- **회귀식**

$$Y = f(X_1, X_2, ..., X_p) + \varepsilon$$

- Y : 반응변수(Response Variable), 종속변수(Dependent Variable)
- X : 설명변수(Explanatory Variable), 예측변수(Predictor Variable)
- ε : 오차항(random error), 모형이 데이터를 정확하게 적합하지 못하는 정도
- f : 독립변수들 간의 관계

- **모수 추정 - 최소제곱법(Least Square Estimation Method)**

각 점으로부터 구하고자 하는 최적 직선까지의 수직거리의 **제곱합을 최소**로 하는 방법



실제 데이터와 우리가 추정한 값의 오차가 작을 수록 좋은 추정

- 모수 추정 - BLUE

  - BLUE(Best Linear Unbiased Estimator)

    : 선형의 불편추정량 중 분산이 가장 작은 추정량

- 오차들의 평균은 0
- 오차들의 분산은 σ^2 으로 동일
- 오차 간에는 자기상관이 없음

* 정규성 조건은 필요하지 않음

세 가지 조건이 충족될 때, 최소제곱추정량은 BLUE!

- 적합성(Goodness of fit) 검정
  - 잔차를 통한 적합성 검정

결정 계수: 총 변동에서 회귀식이 설명하는 부분

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

- 범위: $0 \leq R^2 \leq 1$

- 값이 클수록 회귀식으로 설명되는 변동의 비율이 크므로,
  1에 가까울수록 좋음

- 유의성 검정

$\varepsilon_i \sim N(0, \sigma^2)$ 라는 정규분포 가정하에 개별 베타 계수에 대한 통계적 검정 가능

귀무가설 H_0: $\beta = 0$

대립가설 H_1: $\beta \neq 0$

귀무가설을 기각하지 못하여도,

X와 Y 사이에 선형적 관계가 없을 뿐

아무 의미가 없는 것은 아님!

- **다중선형회귀란?**

여러 개의 설명변수 X와 종속변수 Y의 관계를 표현한 식을 찾는 것

단순선형회귀

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

설명변수를 p개로 확장

다중선형회귀

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i$$

나머지 X 변수들이 고정되었을 때,
x_1이 1단위 증가하면 y는 평균적으로 β_1만큼 증가함을 의미

- **모수의 추정: 최소제곱법(LSE)**

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & \cdots & x_{p1} \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_{pn} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

회귀식

$$\mathrm{Y} = X\beta + \varepsilon$$

목적함수

$$\mathrm{S}(\beta) = \sum_i \varepsilon_i^2 = (\mathrm{Y} - X\beta)'(\mathrm{Y} - X\beta)$$

Normal Equation

$$\frac{\partial S}{\partial \beta} = -2X'\left(\mathrm{Y} - X\hat{\beta}\right) = 0$$

$$\hat{\beta}^{LSE} = argmin\, S(\beta) = (X'X)^{-1}X'Y \quad \text{when } (X'X)^{-1} \text{ exists}$$

- 유의성 검정: 회귀식의 독립변수가 통계적으로 유의미한가?

1. F-test: 모델 전체에 대한 검정

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

$H_1:\ \beta_1, \beta_2, \dots, \beta_\mathrm{p}$ 중 적어도 하나는 0 이 아니다.

if 기각되지 않는다면?

$$y = \beta_0 + \varepsilon \quad (\because \beta_1 = \beta_2 = \cdots = \beta_p = 0)$$

회귀식이 아무런 의미가 없음을 의미!

전부 $\bar{y}$로 예측하게 됨…

- **유의성 검정: 회귀식의 독립변수가 통계적으로 유의미한가?**

3. t-test

: 개별 회귀계수의 유의성을 검정함

$H_0\colon\ \beta_j = 0$	다른 변수들이 적합된 상태에서 x_j는 통계적으로 유의하지 않다
$H_1\colon\ \beta_j \neq 0$	다른 변수들이 적합된 상태에서 x_j는 통계적으로 유의하다

검정통계량

$$t_j = \frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)} \sim t_{n-p-1}$$

- **적합성 검정: 모형이 주어진 데이터를 잘 설명하는가?**

수정결정계수

$$R_a^2 = \frac{SSP/p}{SST/(n-1)} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

- SSE와 SST를 각각의 자유도로 나누어 계산한 형태

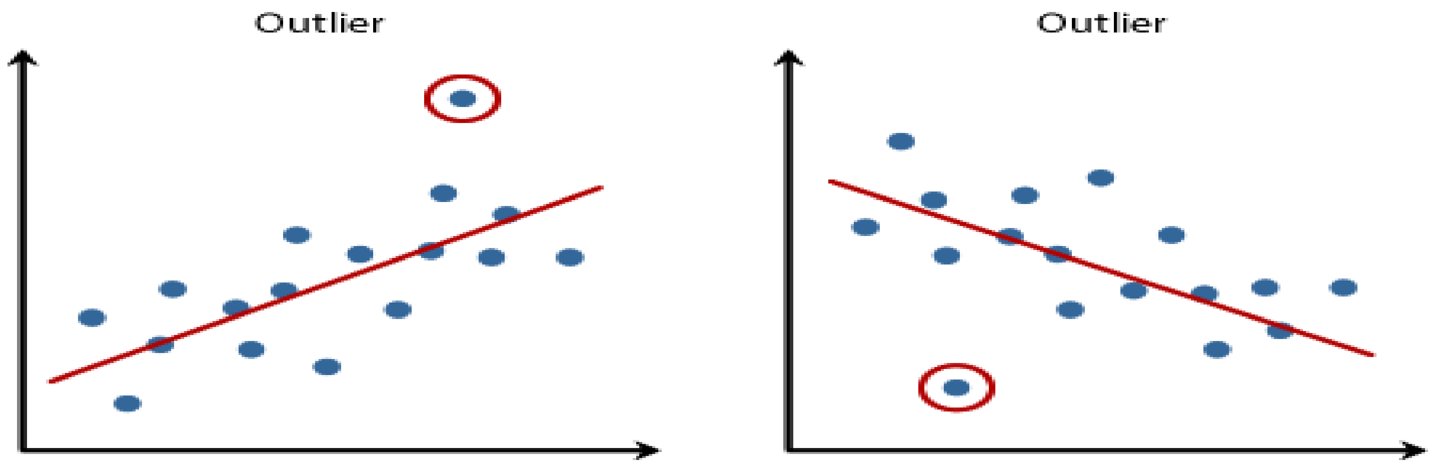
- R_a^2 값이 더 높은 회귀식이 더 좋은 회귀식

- 변수의 개수가 다른 두 회귀식을 비교할 때 사용 가능

- **이상치(Outlier)**

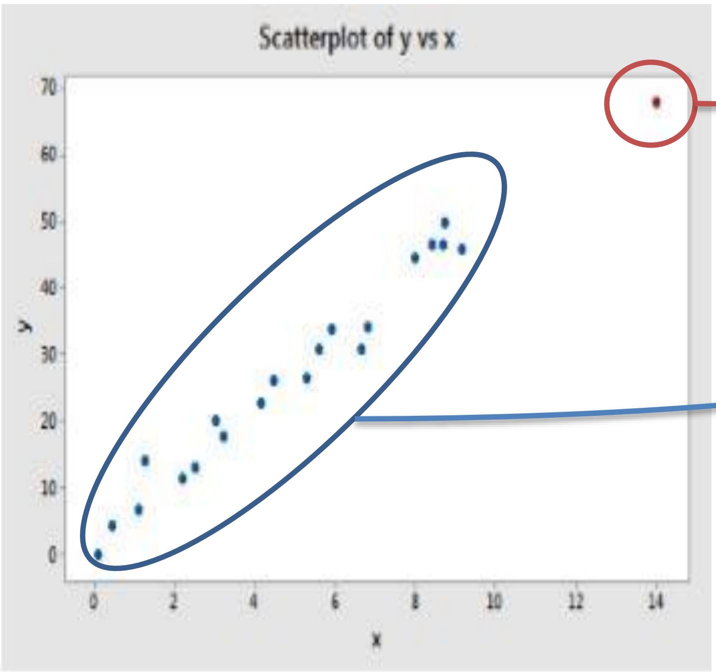
표준화 잔차가 매우 큰 값!

$|r_i| > 3$이면 이상치로 판단!



- **지렛값(Leverage Point)**

표준화했을 때, x 기준에서 절대값이 큰 값!



$h_{ii} \geq \frac{2(p+1)}{n}$ 이면 지렛값!

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$$

- 로버스트(Robust) 회귀란?

  → 건장한, 탄탄한

  이상치의 영향력을 크게 받지 않는 회귀모형

- 로버스트(Robust) 회귀 종류

Median Regression

Huber's M-estimation

Least Trimmed Square