

INDEX

1. 회귀분석이란?
2. 단순선형회귀
3. 다중선형회귀
4. 데이터 진단
5. 로버스트 회귀

- 회귀모델링 과정

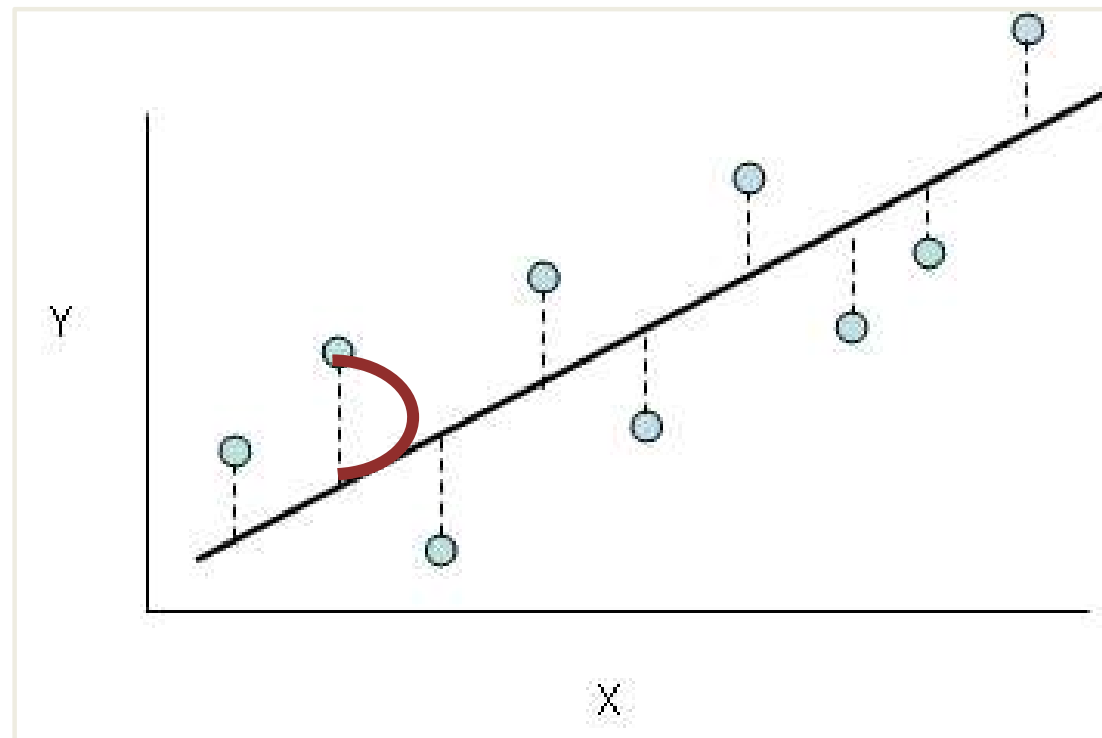
예시

학점과 통학거리는 관련이 있을까??



- 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!

- 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

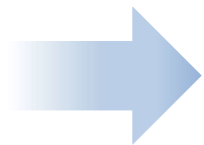
$$Y = X\beta + \varepsilon$$

목적함수

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

Normal
Equation

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



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

- 유의성 검정: 회귀식의 독립변수가 통계적으로 유의미한가?

2. Partial F-test

$$H_0: \beta_{q+1} = \beta_{q+2} = \dots = \beta_p = 0 \quad \text{RM이 적절}$$

$$H_1: \beta_{q+1}, \beta_{q+2}, \dots, \beta_p \text{ 중 적어도 하나는 0이 아니다.} \quad \text{FM이 적절}$$

- 회귀식 전체에 대한 F-test는 Partial F-test의 한 케이스임

$$F = \frac{\frac{SSR(FM) - SSR(RM)}{p}}{\frac{SSE(FM)}{n - p - 1}} = \frac{\frac{SSR}{p}}{\frac{SSE}{n - p - 1}} = \frac{MSR}{MSE} \quad \rightarrow \quad \sum_i (\bar{Y} - \bar{Y})^2 = 0$$

- 데이터진단, 왜 필요할까?

일반적인 경향에서 벗어나는 데이터

ex) 이상치, 지렛값, 영향점 등



회귀 모형에 큰 영향을 미침



어떻게 해결할까?

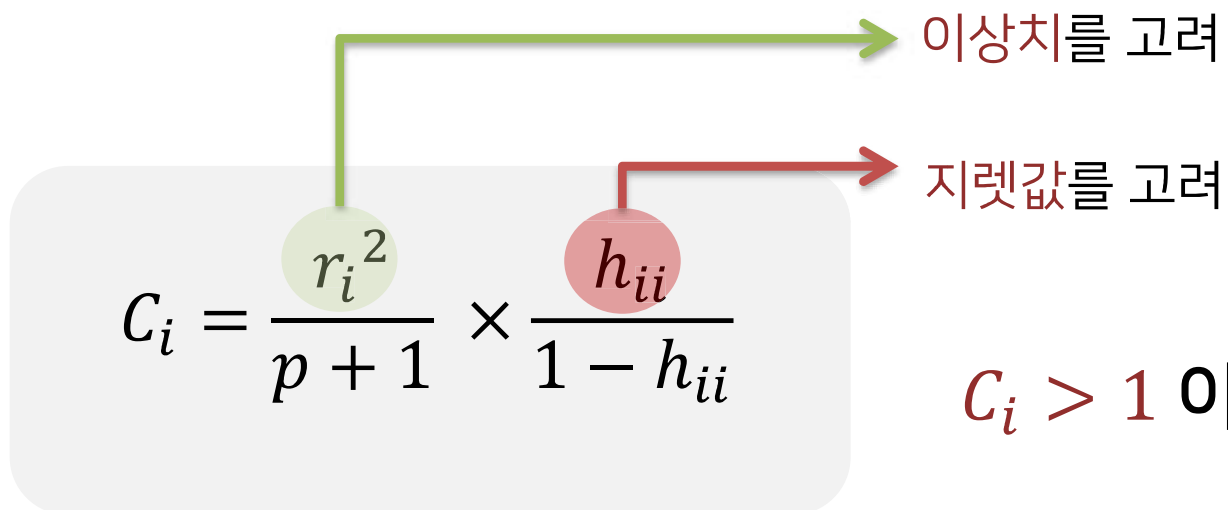
표준화 잔차

(Standardized residual)를
이용!

표준화 잔차값 -> 관측치가
경향성에서 벗어나는지 판단

- Cook's distance

이상치와 지렛값을 동시에 고려하여, 특정 데이터를 지웠을 때 회귀선이 변하는 정도를 나타내는 지표


$$C_i = \frac{r_i^2}{p + 1} \times \frac{h_{ii}}{1 - h_{ii}}$$

$C_i > 1$ 이면 영향점이라 판단!

- 로버스트(Robust) 회귀란?

↘ 건장한, 탄탄한

이상치의 영향력을 크게 받지 않는 회귀모형

- 로버스트(Robust) 회귀 종류

Median Regression

Huber's
M-estimation

Least Trimmed
Square