회귀분석팀

6팀

권남택
윤주희
진효주
한유진
황유나

INDEX

1 회귀분석이란?

상관분석과 차이

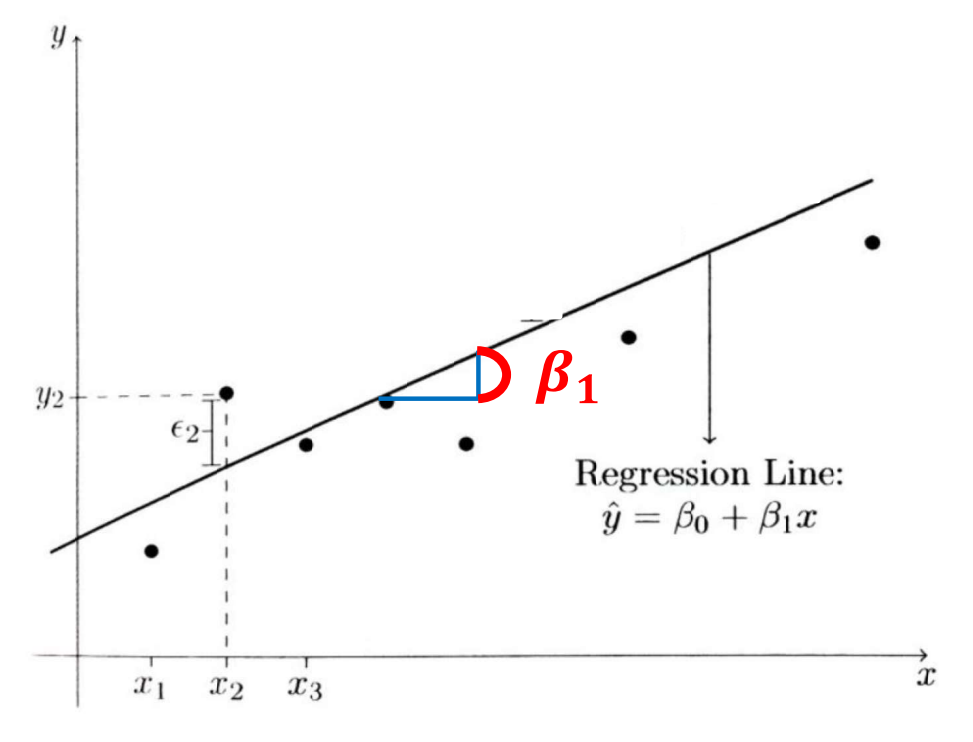
1. 변수 간의 수치적 관계를 표현함으로써 예측력과 설명력 get!

2. 두 개 이상의 변수에 대한 상관관계 설명 가능!

3. 꼭 선형관계가 아니어도 OK! 비선형관계에 대해서도 표현 가능

모델의 해석

: X가 한 단위 증가할 때, Y는 β_1 만큼 증가한다.



평균적으로 β_1 만큼 증가한다는 것!

LSE의 가정과 특징

1. 오차들의 평균은 0

2. 오차들의 분산은 σ^2으로 동일 (등분산)

3. 오차간에는 자기상관이 없다 (uncorrelated)

전부 만족하면 BLUE(Best Linear Unbiased Estimator)

$\varepsilon_i \sim N(0, \sigma^2)$ 라는 정규분포 가정하에서의

개별 베타 계수에 대한 통계적 검정

귀무가설은 β=0 ➡

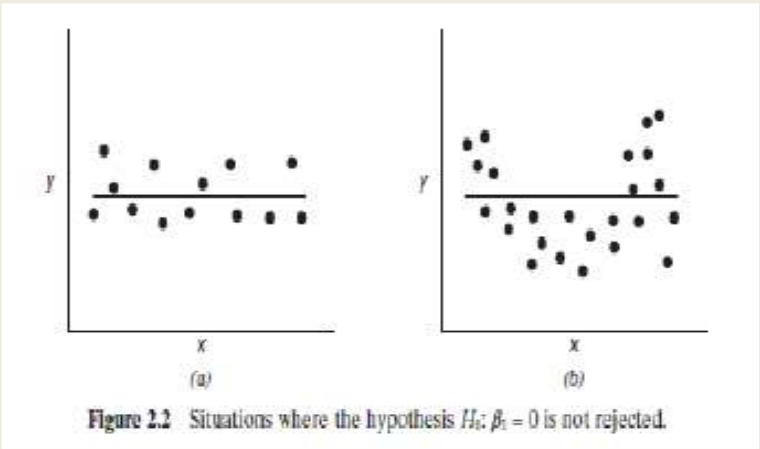


Figure 2.2 Situations where the hypothesis H_0: $\beta_1 = 0$ is not rejected.

기각하지 못하면, 개별 회귀계수는 0이라는 것

↓

X와 Y사이에 아무 의미가 없다는 것이 아니라

단지 **선형적 관계**가 없는 것!

"RM에는 없고 FM에는 있는 변수들, 즉 FM에서 변수를 추가할 때 설명력이 유의미하게 증가하는가?"

<Partial F-test>--→ 몰라도 ok

일반적으로 회귀식 전체에 대한 검정을 한다!

$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$ $H_1: at\ least\ one\ \beta_j \neq 0$

$y = \beta_0 + \varepsilon,$ VS $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$

<T-test: 가 의성 검정>

$H_0: \beta_j = 0,$ $H_1: \beta_j \neq 0$

H0 : 다른 변수들이 다 적합 통계적으로 유의하지 않다.

H1 : 다른 변수들이 다 적 는 통계적으로 유의하다.

!T-test로 변수를 선택하는 것은 매우 위험!

"다른 변수들이 고정되어 있는 상황에서"
변수의 유의성 판단

다른 회귀식을 가정했을 때는
해당 변수가 유의할 수도 있음

다른 을 때

est
일

Cook's distance

: 영향점을 확인하는 표준적인 지표

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

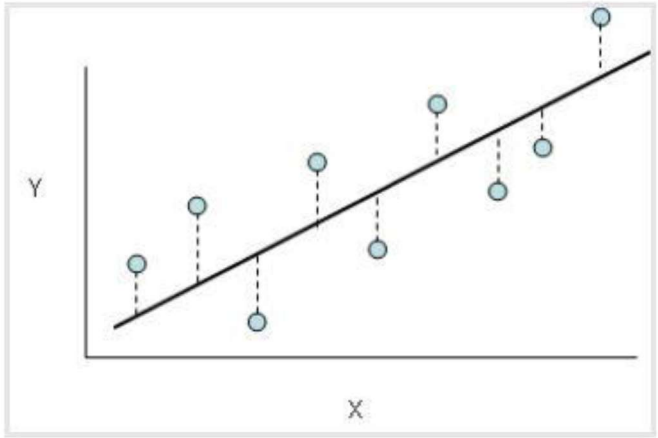
outlier leverage

- C_i가 1보다 크면 영향점으로 간주
- 보통 이러한 영향점은 제거 할 수 있지만 데이터 삭제는 늘 조심해야 함

Median Regression

: 평균, 중앙값, 최빈값 중: 중앙값이 가장 이상치의 영향을 덜 받는다는 생각에 기초

→ 회귀계수를 추정할 때 x에 따른 평균적인 y를 반환하는 것이 아닌 **x에 따른 y의 중앙값**을 반환

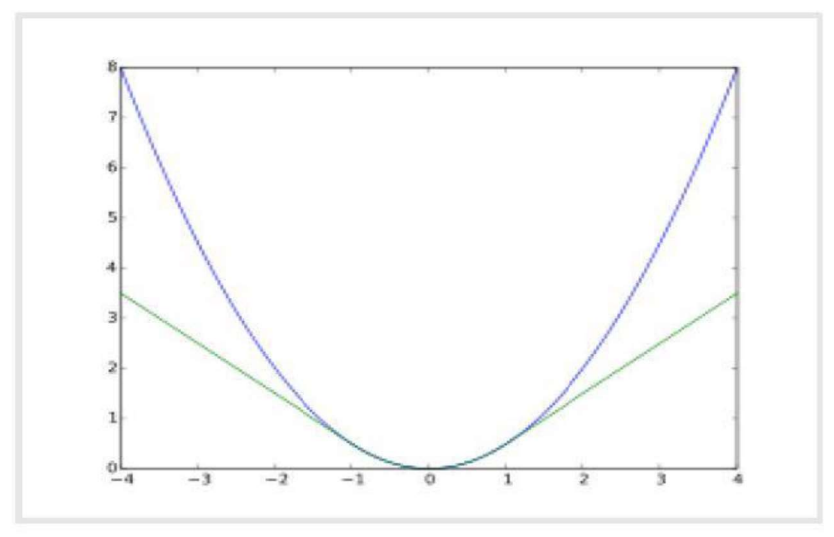


Classical linear regression: $\underset{\beta}{\operatorname{argmin}} \sum(\varepsilon_i)^2)$

Robust regression: $\underset{\beta}{\operatorname{argmin}} \sum |\varepsilon_i|$

Huber's M-estimation

: 잔차에 특정 상수값보다 크면 잔차의 제곱이 아닌 1차식으로 바꿔 회귀계수를 추정



$$if\ |e| \le c, \rho(e) = \frac{1}{2}e^2,$$

$$otherwise\ \rho(e) = c|e| - \frac{1}{2}c^2$$

- 이때의 목적함수(최적화할 함수)는 $\sum \rho(e)$

- R에서는 MASS 패키지의 rlm함수를 사용

최소제곱 형태는 유지하면서 이상치에 대한 가중치를 완화!