회귀분석팀

6팀

신성민
신유정
김찬영
윤주희
이혜인

INDEX

회귀분석의 기본 정의 예시 단계 회귀모형

회귀분석

변수들 사이의 함수적 관계를 탐색하는 방법

회귀모형

$$Y = f(X_1, X_2 \cdots X_p) + \epsilon$$

Y : 반응변수 $f(X_1, X_2 \cdots X_p)$: 예측변수들의 집합 ϵ: 확률오차

공분산

공식 해석 및 설명 한계

공식

$$\mathrm{cov}(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = E(XY) - E(X)E(Y)$$

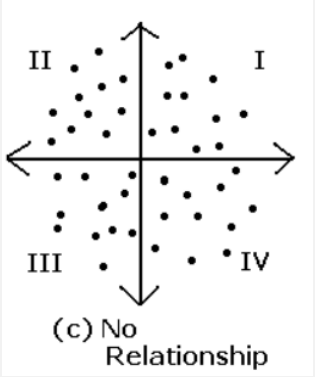
- 두 개의 확률 변수의 상관정도를 혹은 or 두 변수의 관계를 나타내는 값
- 부호에 따라 Y와 X사이의 선형관계에 대한 방향을 나타냄
  (범위 : $-\infty < \mathrm{cov}(x,y) < \infty$)

공분산

공식 **해석 및 설명** 한계

해석 및 설명

$\mathrm{cov}(x,y)>0$	변수 X와 Y가 양의 선형관계
$\mathrm{cov}(x,y)<0$	변수 X와 Y가 음의 선형관계
$\mathrm{cov}(x,y)=0$	변수 X와 Y가 선형적으로 관련X (아무런 관계가 없다는 뜻X)



$\mathrm{cov}(x,y)=0$ 일때 변수 X와 Y 사이에 선형적인 관계는 없지만 두 변수는 서로 독립적인 관계에 있다

상관계수 공식

공식

$$r_{xy} = \frac{Cov(X,Y)}{S_x S_y} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

$$= \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

- 표준화된 변수들 사이의 공분산
- 상관계수의 크기 비교 가능
- $-1 \leq \mathrm{cov}(x,y) \leq 1$; 1과 -1에 가까울수록 더 강한 상관관계

단순회귀분석 정의 최소제곱법 잔차 검정

단순회귀식이란?

X와 Y변수 간의 산점도에 둘의 **관계**를 잘 설명하는 가장 **이상적**인 선

반응변수 Y에 대해 예측변수 X가 **하나**인 회귀식

단순회귀식

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

단순회귀분석 정의 최소제곱법 잔차 검정

설명

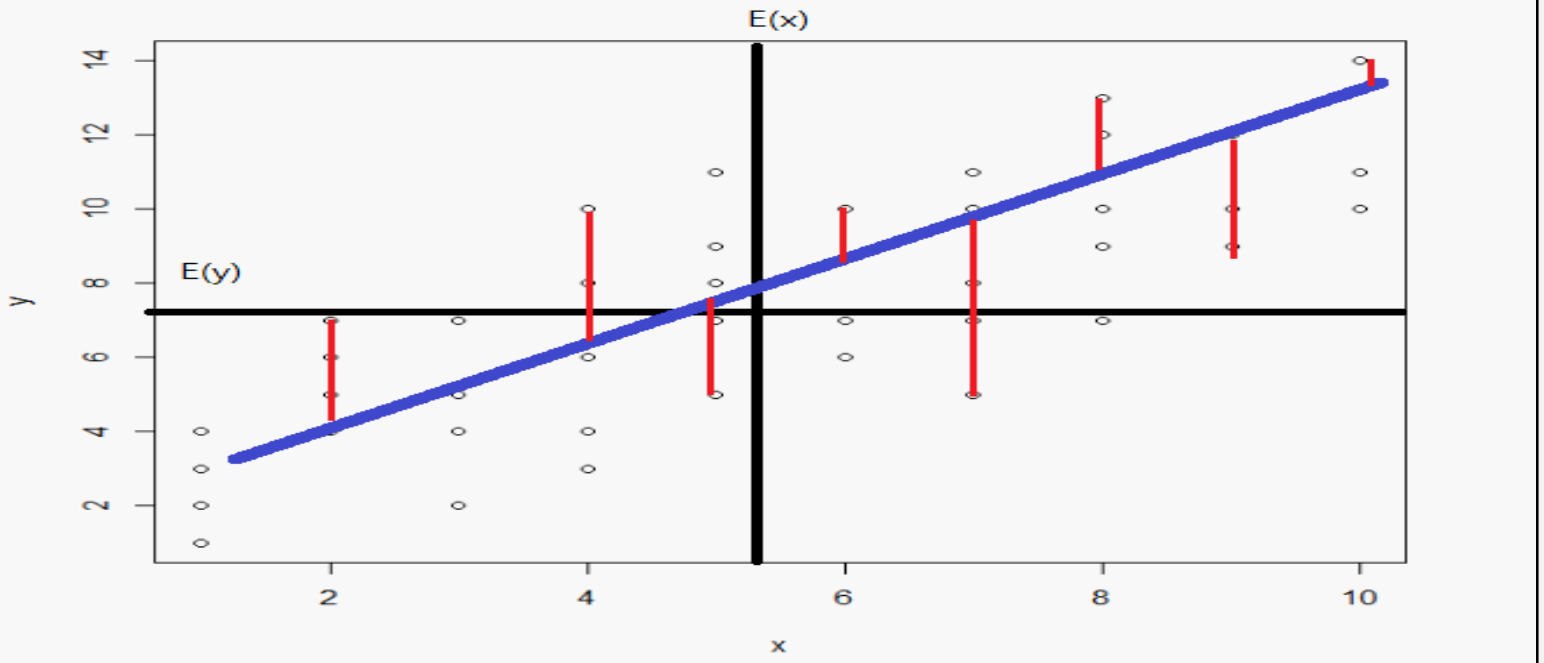
β_0 ⇨ 회귀선의 절편(intercept)
'X=0일 때의 Y의 기댓값'

β_1 ⇨ 회귀선의 기울기(slope)
'X의 한 단위 변화에 대한 Y에 대한 변화'

ε_i ⇨ $N(0,\sigma^2)$을 따른다고 가정
동일한 X값임에도 Y값이 달라질 수 있음을 표현

단순회귀분석 정의 **최소제곱법** 잔차 검정

LSE



즉, 빨간선(오차)들의 제곱합을 최소로 하는 파란선(회귀선)을 구하는 것!

단순회귀분석 정의 최소제곱법 **잔차** 검정

잔차란?

i번째 관측값과 추정한 i번째 값의 차이

잔차식

잔차 $e_i = y_i - \hat{y}_i$

- y_i: i번째 관측값
- $\hat{y}_i$: 추정한 i번째 값

단순회귀분석 정의 최소제곱법 잔차 검정

회귀계수가 유의한가?

결론

P-value < 0.05

↓

귀무가설 기각

↓

개별 회귀 계수가 유의하다!

개별 t-test

H0 : $\beta_1 = 0$

H1 : $\beta_1 \neq 0$

다중선형회귀 정의 모수추정 검정 적합성

다중선형회귀

X와 Y변수 간의 관계를 설명해주는 이상적인 선을 그린다는

단순선형회귀와 개념은 같음. 다만, 다중선형회귀에서는 예측변수인 X가 많음.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_p X_p + \epsilon$$

- β_0: 회귀모수 (P+1개)
- X_2: 예측변수
- ϵ: 오차 (평균이 0이고 분산이 σ^2인 확률변수)

→ β_q: 나머지 예측변수를 고정시켰을 때, x_{jq} 의 한단계 증가에 따른 y_j 의 변화량

다중선형회귀 정의 모수추정 **검정** 적합성

개별 회귀 계수 검정

β_j에 대한 검정

- t검정 이용
- $H_0 : \beta_j = 0,\ H_1 : \beta_j \neq 0$
- 검정통계량

$$t_j = \frac{\hat{\beta}_j - 0}{s.e.(\hat{\beta}_j)}$$

모형에 대한 가설 검정

회귀식에 대한 검정

- F검정 이용
- H_0 : RM이 적절하다

  H_1 : FM이 적절하다

다중선형회귀 정의 모수추정 검정 **적합성**

회귀모델 적합성 측정법

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

회귀식이 데이터를 잘 설명하고 있는지 평가할 수 있는 척도

SSR? SST? SSE? 이게 다 뭐죠..?

다중선형회귀 정의 모수추정 잔차 검정 적합성

수정결정계수 R_a^2

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

- SSE와 SST를 각각의 자유도로 나누어 준 후 계산한 것
- R_a^2 은 예측변수 X의 개수가 다른 모델간 비교가 가능함

  (R^2는 X의 개수가 많아질수록 값이 커짐 -> 개수가 다른 모델 비교 X)
- 다만, R_a^2은 R^2처럼 Y의 변이가 X에 의해 설명되는 비율로 표현 불가함