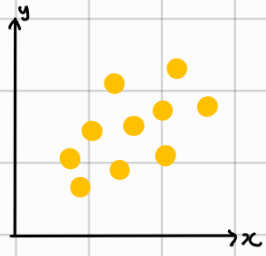
• 05강 선형회귀 →

공분산과 상관계수는 두 데이터 간의 관계를 밝히기 위함이다.
단순 숫자가 아니라 전과 방정식을 구하고 싶은 것.
두 데이터 간에 어떤 관계가 있는지 정확히 수식으로 나타내고 싶다.
그래서 한 변수가 변할 때,
다른 변수가 정확히 어떤 값이 나오는지 알고싶은 거임
실제로 선형회귀, 두 데이터 간의 정확한 방정식을 구하는 것 자체로도
큰 연구 결과일 수 있다. 그리고 2차, 3차가 결코 좋은게 아니다.
공학적으로 어떤 효과를 냈을 때, 최대한 단순하게 좋은거임

원래대로 돌아가다, 본질로 돌아가다 라는 의미

- 회귀: 에러를 제거한 두 데이터 간의 관계도출

• 선형회귀: $y = ax + b$ (1차식)
(= 단순회귀)



$y = ax + b + \varepsilon$
입실론, 에러

• 선형회귀 유도 (기울기)

기울기 a, 상수항 b

$$y = ax + b + \varepsilon$$

$$Cov(x,y) = Cov(x, ax+b+\varepsilon)$$

$$= Cov(x, ax) + Cov(x, b) + Cov(x, \varepsilon)$$

$Cov(x, b)$: b는 상수기 때문에 어떻게 변하든 같은 값이 나온다.
⇒ 정지된 상태
⇒ 독립 = 0

$Cov(x, \varepsilon)$: 변수와 error는 절대 상관관계가 있어서는 안된다.
있으면 그건 error가 아님.
어떤 요인이 반드시 숨어있음.
만약 내가 움직일 때마다 error가 생기면 그걸 야기하는 다른 요인이 있는거임.
그걸 찾아야지 그걸 ε로 치부하면 안됨

$$= a\,Cov(x,x) = a\,V(x)$$

자기자신과 Cov는 분산과 같은말임

$$Cov(x,y) = aV(x) \text{ 니까 } a = \frac{Cov(x,y)}{V(x)}$$

- 선형회귀 유도 (y절편)

$y = ax + b + \varepsilon$

· y의 평균.

· 기대값 (Expected Value)의 E

· y를 제곱구해서 평균을 내면 회귀를 따라가겠다. 라는 가정.

$E(y) = E(ax + b + \varepsilon)$

$= E(ax) + E(b) + E(\varepsilon)$

b가 상수 있는 것의 평균은 b임.

에러들의 평균은 0임. noise가 0이 안되고 한쪽으로 편향된다면 없애줘야하는 요인임.

$= aE(x) + b$ 이다.

그럼 $b = E(y) - aE(x)$

y의 평균 x의 평균 x a

$b = E(y) - \frac{cov(x,y)}{V(x)} E(x)$

- 최종

$y = ax + b$

$y = \frac{cov(x,y)}{V(x)} x + E(y) - \frac{cov(x,y)}{V(x)} E(x)$

문제) [1,2,3] 과 [1,3,5]의 선형회귀 식은?

[1,2,3] = x, [1,3,5] y 일때

$E(x) = 2$

$E(y) = 3$

$cov(x,y) = \frac{4}{3}$

$V(x) = \frac{2}{3}$

$y = \frac{\frac{4}{3}}{\frac{2}{3}} x + 3 - 2 \cdot 2$

$= 2x - 1$

확인 => 1넣으면 1, 2넣으면 3, 3 넣으면 5나옴

- 선형회귀 기울기와 상관계수와의 관계

$y = ax + b$

선형회귀 기울기 $a = \frac{Cov(x,y)}{V(x)} = \frac{Cov(x,y)}{\sigma(x)^2} = \frac{Cov(x,y)}{\sigma(x)\,\sigma(x)}$

(x,y의 관계)

$V(x)$: 분산 (데이터가 얼마나 퍼져 있는지)

$\sqrt{V} = \sigma$, $V = \sigma^2$

상관계수 $r = \frac{Cov(x,y)}{\sigma(x)\,\sigma(y)}$ ← Cov를 ← $\sigma(x)\sigma(y)$로 나눈 값.

여기까지 비슷하기 때문에, 아래식 유도 가능

$$r \cdot \frac{\sigma(y)}{\sigma(x)} = a$$

r에 x의 표준편차를 나누고
y의 표준편차를 곱하면 → 선형회귀의 기울기가 나온다.