상관 분석

상관성의 정의 :
- 방향, 크기 기준
- 동시에 같은 방향으로 움직이는지
- 두 변수의 선형 관계만을 판단 (비선형일 경우 상관분석 고려대상 X)

기본가정 : 두 변수가 이변량 정규분포 (두 변수가 정규분포를 따를 때의 결합확률밀도함수)
→ 만족 X → 스피어만 상관계수를 사용.

*Cov, 상관계수 차이

Cov는 X,Y 단위에 따라 공분산에 영향을 미치는 정도가 다름.
→ 단위에 영향 많이 받음

$$\rho_{xy} = \frac{E[(X-\mu_x)(Y-\mu_y)]}{\sigma_x \sigma_y}$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

피어슨 상관계수

유의성 검정 — Why? 자체 분포 알수 없기 때문에 t분포 검정에 사용 불가능.

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 1$$

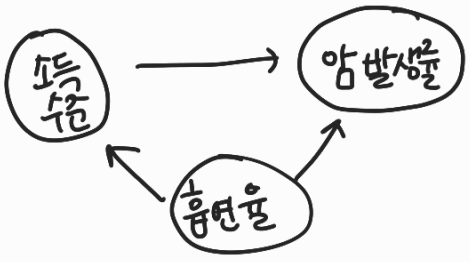
$$t = r\sqrt{\frac{(n-2)}{(1-r^2)}} \sim t_{n-2}$$

(bh) 문제점
검정 통계량의 값이 표본 크기 n과 직관적으로 연결
→ n이 커지면 기각 가능성 ↑

한계점
① 인과성 보여주지 X
② 비선형 관계 고려 X
③ 데이터 구조에 민감
- 이질적 소표본으로 구성된 경우 소표본에 각각에 대한 상관분석하는 것이 바람직

상관계수 ρ < 실제 두 변수 사이 인과관계
인과관계를 존재 X → 허위 상관 → 영향을 주는 제 3의 변수 (교란 변수 or 공통반응 변수)

* 교란변수
: 두 변수와 떨어져 위치에 있으면서 두 변수와 모두 연관성이 변수로 인해 X→Y 처럼 보이는 관계 발생



부분상관계수 (편상관계수)
새로 상관성을 보이는 변수들 즉 2개 선택 후 다른 변수들의 영향을 제거하고 오직 두 변수만의 상관계수를 측정한 값

$r_{12\cdot3} = \dfrac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}$

1, 2 변수간 편 상관계수

스피어만 상관계수 - 서열척도 data

$$\rho = \frac{\sum (R(x_i) - \overline{R(x)}) \sum (R(y_i) - \overline{R(y)})}{\sqrt{\sum (R(x_i) - \overline{R(x)})^2} \sqrt{\sum (R(y_i) - \overline{R(y)})^2}}$$

$$\overline{R(x)} = \overline{R(y)} = \frac{n+1}{2}, \quad \sum (R(x_i) - \overline{R(x)})^2 = \frac{n(n^2-1)}{12}$$

$$\rho = 1 - \frac{6\sum (R(x_i) - R(y_i))^2}{n(n^2-1)}$$

독립성

서로 독립이다 ⇄ 상관관계 존재 X (uncorrelated)