- 데이터 탐색 과정에서 평균, 분산, 왜도, 첨도 그리고 결측치 등 각 변수들의 특성을 파악
- 이때 변수들간의 관계도 파악한다
- Y와 X와 관계뿐만 아니라 특성들간의 관계도 살펴보아야 한다
- 이를 통해 독립 션수의 변화에 따른 종속 변수의 변화량을 크게하여 통계적 정확도를 감소시키는 다중공선성을 방지할 수 있으며, 데이터에 대한 이해도를 높일 수 있다.
- 대표적인 방법인 공분산과 상관계수를 알아보자

상관분석

- 상관 분석을 위해서는 데이터가 등간이나 비율 척도이며, 두 변수가 선형적 관계라는 기본 과정을 함

- 공분산(Covariance)과 상관계수(Correlation coefficient)는 각 변수의 변동이 얼마나 닮았는지를 표현하는 지표
- 두 방식은 계산 방식에서 차이가 있음

공분산

- Covariance
- 서로 공유하는 분산
- 한 변수의 각각의 데이터가 퍼진 정도를 나타내지만, 공분산은 두 분산의 관계를 뜻함
- 두 변수 사이의 상호연관성을 측정하는 지수가 공분산(covariance)

공분산

$$\text{모집단} : \mathrm{Cov}(X,\ Y) = \sigma_{XY} = \sum_{i=1}^{N} [X_i - E(X)][Y_j - E(Y)]P(X_iY_j)$$
$$= E[[X - E(X)][Y - E(Y)]]$$
$$= E(XY) - E(X)E(Y)$$

$P(X_iY_j)$: X의 i번째 결과와 Y의 j번째 결과의 발생확률

- 분산하고 다른 점은 음수의 값을 가질 수 있다.
- 공분산이 양수이면 두 변수가 같은 방향으로 움직이고, 음수이면 두 변수가 반대 반향으로 움지기이는 것을 의미

- 공분산은 두 변수의 선형관계를 알려주지만 두 변수의 측정단위에 의존하기 때문에 그의 크기는 두 변수의 선형관계의 강도를 나타내는 지표는 아니다.
- 공분산을 두 변수의 표준편차로 나누는데 이 값이 상관계수이다.

공분산의 특성

- 공분산은 범위에 제한이 없음
- 공분산은 측정단위에 영향을 받음
- 두 변수가 서로 독립이면 공분산은 0
- 공분산이 0이라고 해서 두 변수가 반드시 독립은 아님
- 공분산이 0이면 상관계수도 0
- 공분산은 선형관계의 측도

공분산의 한계

- 공분산의 단위에 영향을 많이 받음
- 그렇기 때문에 선형관계를 비교하는 적당한 통계량은 아님
- 예)
  - A와 B의 신장을 cm로 측정하여 공분산을 구했을 때와
  - m로 변환하여 공분산을 구하면 공분산이 변하게 됨
- 즉, 공분산 값 자체로는 얼마나 강한 연관성이 있는지 알기 어려움

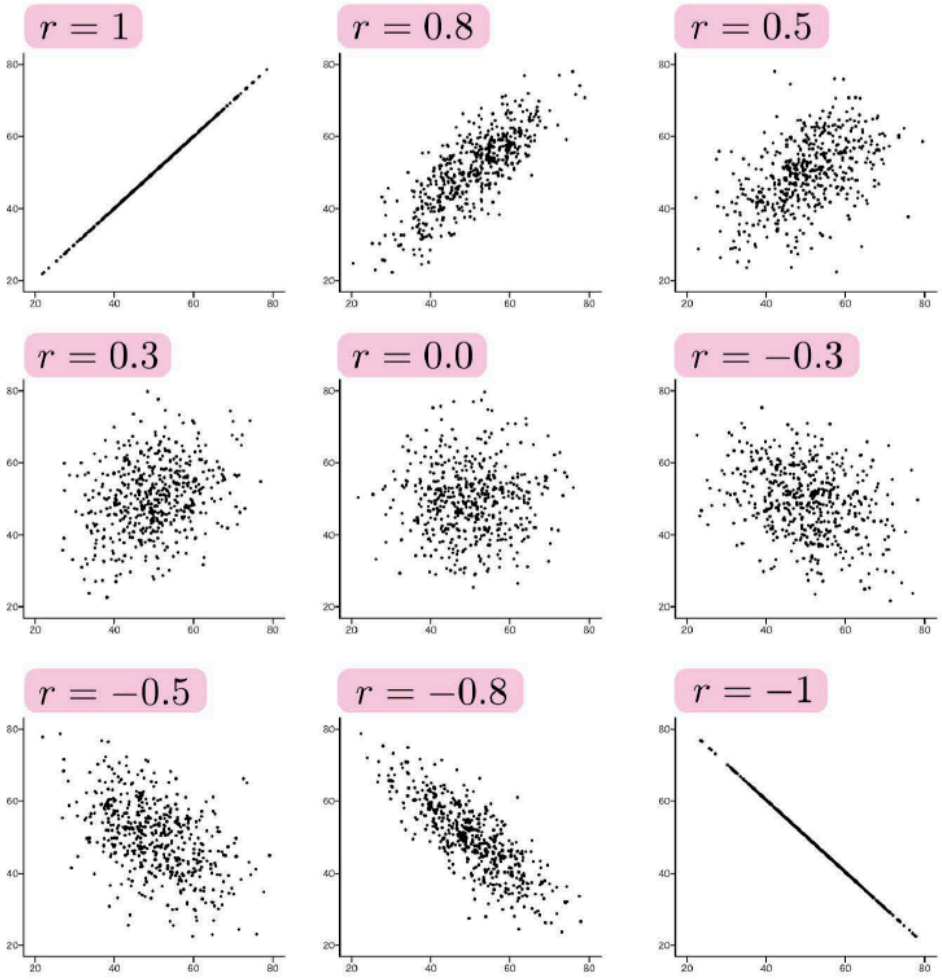
상관계수

- Correlation Coefficient
- 2개의 변수 간의 강도를 정량화한 값이 상관계수
- 사회과학에서 상관계수를 해석하는 기준은 절댓값이 0.7 이상일 때 상관관계가 매우 높다고 판단
- 0.4 이상이면 어느 정도 상관관계가 있다고 해석하지만 분야에 따라서 차이가 있다.

피어슨 상관계수

- 피어슨 상관계수 r(Pearson's correlation r)이라 부르는 값
- 2개 양적 변수 사이의 선형관계가 얼마나 직선 관계에 가까운가를 평가

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2}\sqrt{\frac{1}{n}\sum_{k=1}^{n}(y_k - \bar{y})^2}}$$



- x와 y의 관계 강도를 다양하게 바꾸어 피어슨 상관계수 r을 계산
- 직선에 가까울수록 r의 절댓값은 1에 가까워지고, x와 y 사이에 아무런 관계가 없을 때는 0
- 부호가 양일 때는 x가 커질수록 y도 함께 커지고, x가 작아질수록 y도 함께 작아지는 관계성을 양의 상관(positive correlation)
- 부호가 음일 때는 x가 커질수록 y는 작아지고, x가 작아질수록 y는 커지는 관계성을 음의 상관(negative correlation)
- 아래 그림은 해석의 예이며, 도메인에 따라 정확한 해석이 달라진다

  - $0.7 < |r| \leqq 1$: 강한 상관
  - $0.4 < |r| \leqq 0.7$: 중간 정도 상관
  - $0.2 < |r| \leqq 0.4$: 약한 상관
  - $0.0 < |r| \leqq 0.2$: 거의 상관없음

- 피어슨 상관계수는 모수적인 방법

- 주의할 점은 산점도의 기울기와 상관계수는 관련이 없다는 것 - 분산의 관계성이 같다면, 기울기가 크

  든 작든 상관계수는 같다
- 위의 그림은 기울기가 모드 다르지만 상관계수는 1로 동일하다
- 이처럼 상관계수가 높다는 것은 X_1이 움직일 때 X_2가 많이 움직인다는 뜻이 아니라, X_2를 예상할 수 있는 정확도, 즉 **설명력이 높다는 것**

스피어만 순위상관계수

- Spearman's Rank Correlation Coefficient
- 관측치의 분포가 극단적인 분포를 보이거나 관측치가 순위정도의 정보 밖에 갖고 있지 않을 경우에 이 변수들 간의 상관관계를 빨리 알고자 할 때 사용
- 데이터의 x축, y축 중 적어도 하나 이상에 정규성이 없을 때는 비모수 상관계수인 스피어만 순위상관계수(Spearman's rank correlation coefficient ρ) 사용이 권장
- ρ는 피어슨 상관계수처럼 -1부터 1까지의 실수값
- 데이터 값을 x축, y축 각각에서 크기 순으로 나열했을 때의 1위, 2위... 등의 순위로 변환한 다음 피어슨 상관계수 공식을 사용
- 이렇게 하면 이상값이 있을 때도 사용할 수 있음

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n}$$

여기서, $\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (X_i - Y_i)^2$: 순위에 대한 편차제곱합