7.5 공분산과 상관계수

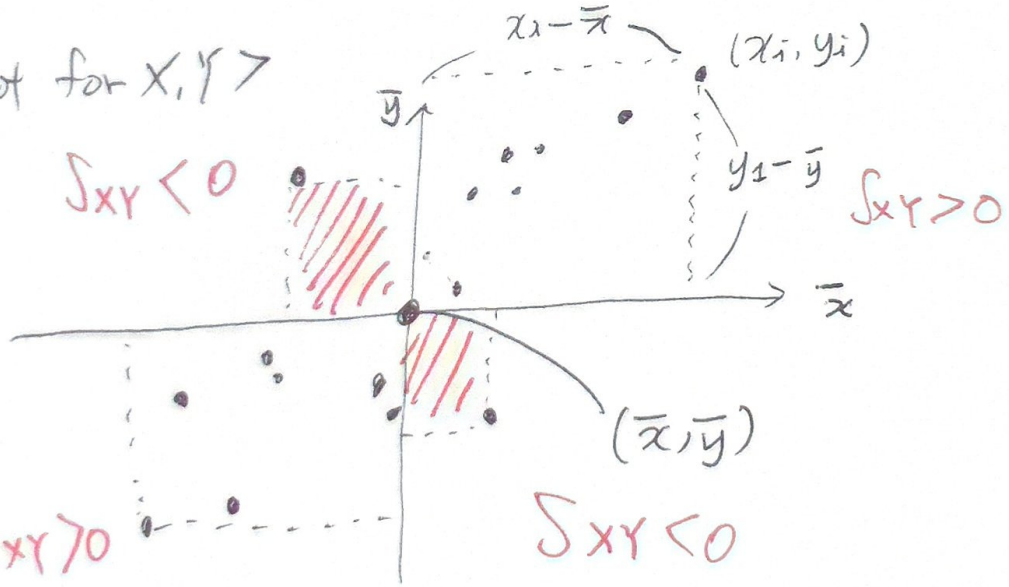
○ 변수 특징변수 간의 상관관계를 숫자로 나타낸 것이
공분산(covariance)과 상관계수(correlation coefficient)다.

표본 공분산 sample covariance

$$S_{XY} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

$X \longleftrightarrow Y$, 얼마나 독립인가? 얼마나 상관성이 존재하는가?

<Scatter plot for X, Y>



(data) (data)

- 1, 3사분면 多, 2·4사분면 小 → $S_{XY} > 0$
- 2·4사분면 多, 1·3사분면 小 → $S_{XY} < 0$
- 2·4사분면 ≅ 1·3사분면 data → $S_{XY} \cong 0$

} 모양 정보

- → $|S_{XY}|$ 작음
- → $|S_{XY}|$ 큼

} 퍼져있는 정도에 대한 정보 (분산 크기)

표본상관계수

표본공분산은 평균을 중심으로 각 자료들이 어떻게 분포되어 있는지 크기와 방향성을 같이 보여준다. 그런데 분포의 크기는 공분산이 아닌 분산만으로도 알 수 있기 때문에 대부분의 경우 자료 분포의 방향성만 분석하여 보는 것이 유용하다.

이때 필요한 것이 표본상관계수(sample correlation coefficient)다.
표본상관계수는 아래와 같이 공분산을 각각의 표본표준편차값으로 나누어 정규화(normalize)하여 정의한다.

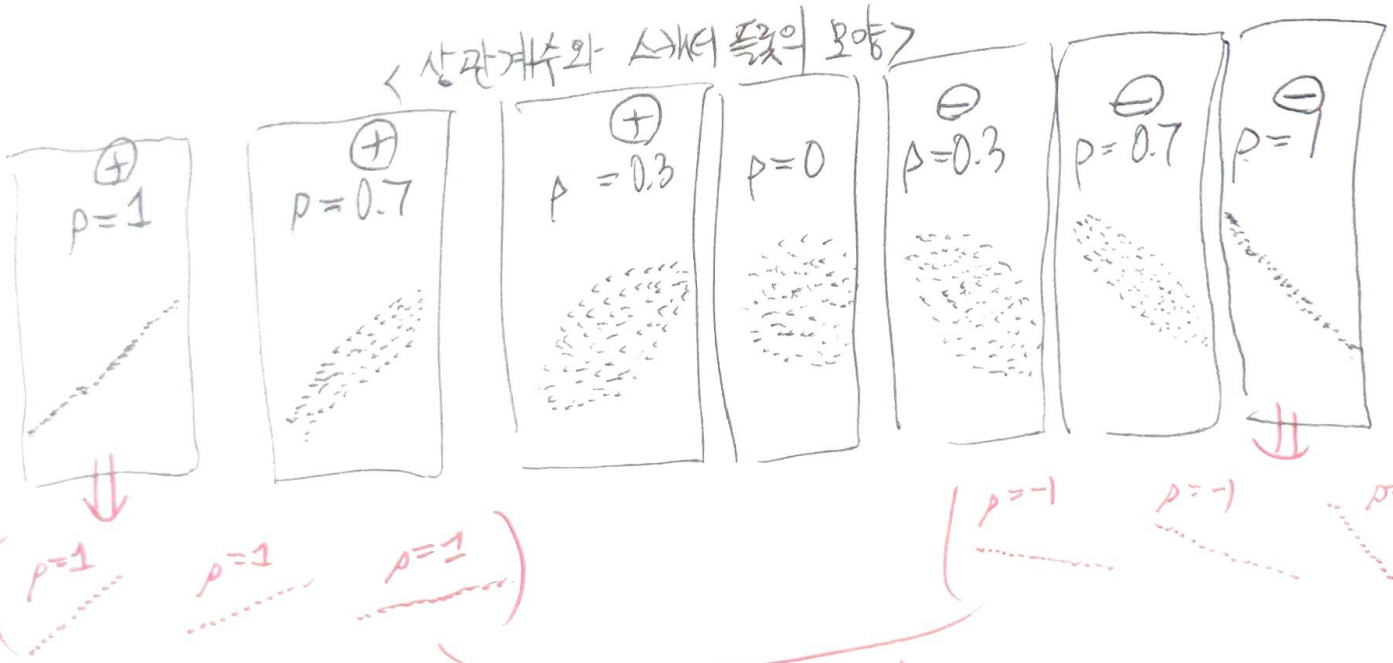
$$\text{Pearson } r = r_{xy} = \frac{S_{xy}}{\sqrt{S_x^2 \cdot S_y^2}} = \frac{\text{공분산}}{\sqrt{\text{표본분산} \times \text{표본분산}}}$$

이와 다르게 정의한 상관계수를 알기 때문에 다른 종류의 상관계수와 비교하여 말할 경우 피어슨(Pearson) 상관계수라고 하기도 한다.

사이파이 stats 서브패키지는 피어슨 상관계수를 계산하는 pearsonr() 함수를 제공한다. pearsonr() 함수는 상관계수와 유의확률을 반환한다. 유의확률에 대해서는 9장에서 공부한다.

scipy.stats.pearsonr(x1, x2)[0] = r
[1] = 유의p

<상관계수와 스캐터 플롯의 모양>



확률변수의 공분산과 상관계수

두 확률변수 X와 Y의 공분산은 기댓값 연산자를 사용하여 다음과 같이 정의된다.

$$Cov[X, Y] = E[(X - E[X])(Y - E[Y])]$$

마찬가지로 두 확률변수 X와 Y의 상관계수를 다음과 같이 정의한다.

$$\rho[X, Y] = \frac{Cov[X, Y]}{\sqrt{Var[X] \cdot Var[Y]}}$$

↑ 이론적 확률, rho

확률변수의 상관계수는 다음과 같은 성질을 가진다.

$$-1 \le \rho \le 1$$

또한 특수한 ρ에 대하여 다음과 같이 부른다.

- rho $\rho = 1$) 완전선형 상관관계
- $\rho = 0$) 무상관 ($\neq$ 독립)
- $\rho = -1$) 완전선형 반상관관계

독립 $\Rightarrow \rho = 0$

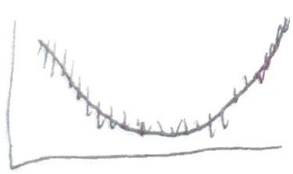
$\rho = 0$ $\not\Rightarrow$ 독립

비선형 상관관계

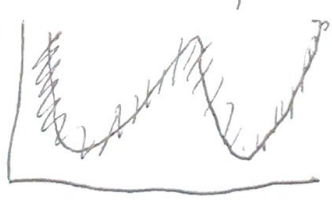
두 확률변수가 상관관계가 있으면, 두 확률변수의 값 중 하나를 알았을 때 다른 확률변수의 값에 대한 정보를 알 수 있다. <반드시 정확한 값을 알 수 있어야 하는 것은 아니다.>
hint를 얻을 수 있다는 의미.

* 피어슨 상관계수는 두 확률변수의 관계가 선형적일 때만 상관관계를 제대로 계산할 수 있다.

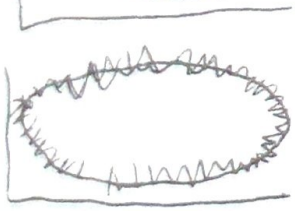
(비선형 상관관계 1) r = 0.0



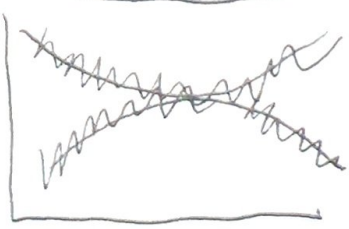
(비선형 상관관계 2) r = 0.0



(비선형 상관관계 3) r = 0.0



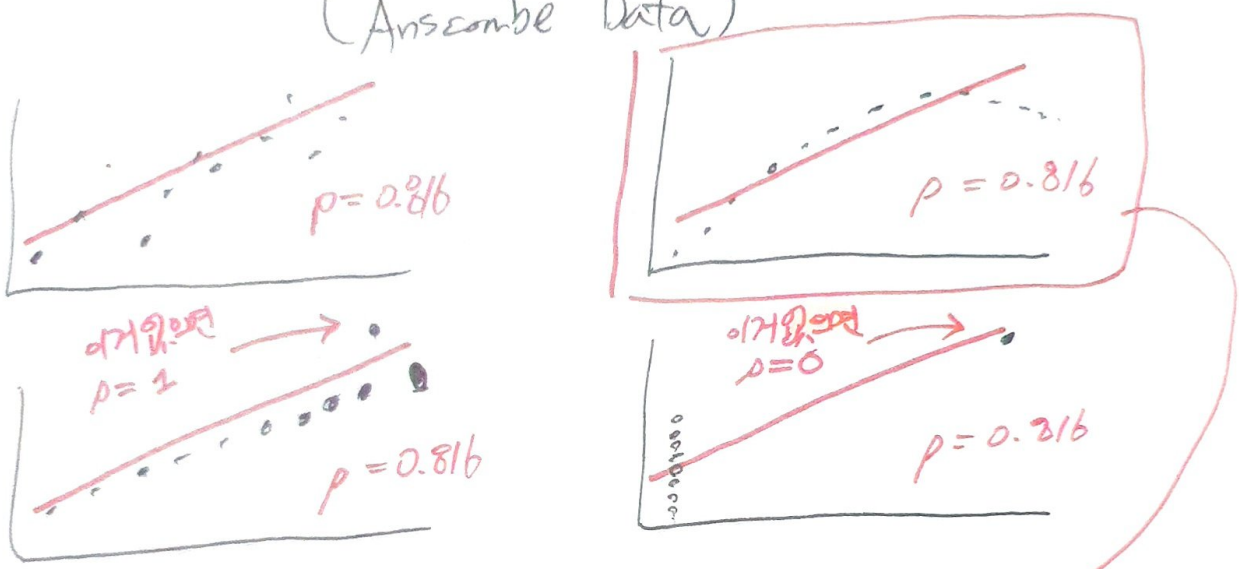
(비선형 상관관계 4) r = 0.0



앤스콤 데이터

상관계수로 분포 모양을 추측할 때, 개별 자료가 상관계수에 미치는 영향력이 각각 다르다는 점에 유의해야 한다. 다음은 프랭크 앤스콤(Frank Anscombe)의 논문에 제시된 데이터다. 이 데이터는 서로 다른 4종류의 2차원 데이터셋을 포함하는데, 4종류 데이터셋 모두 상관계수가 약 0.816으로 동일하다.

(Anscombe Data)



매우 깨끗한 (x→y) 그래프 상관관계로 추측가능함

But ρ = 0.816

Conclusion

pearson 상관계수는
2차식 등 비선형 상관관계를 catch하지 못하며
outlier에 민감/취약하다.

다변수 확률변수의 표본공분산

- 스칼라가 아닌 벡터 표본값을 가지는 다변수 확률변수의 공분산
- $X_1, X_2, \cdots, X_M$ 이라는 서로 다른 M개의 확률변수
- 각 확률변수마다 각각 N개의 표본
  - j $(j = 1, \cdots, M)$ 번째 확률변수의 i $(i = 1, \cdots, N)$ 번째 데이터 $= x_{i,j}$

"이렇게 놓으면 다음과 같은 특징행렬이 됨"

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,M} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \cdots & x_{N,M} \end{bmatrix}$$

M개의 서로 다른 확률변수의 모든 조합에 대한 공분산을 한꺼번에 표기하기 위하여, 다음처럼 **표본 공분산 행렬** (Sample Covariance Matrix)을 정의한다.

- 대각성분 = 각 확률변수의 분산으로 정의
- 비대각성분 = 서로 다른 두 확률변수의 공분산으로 정의 — 교재식 계산한다.

$$\Sigma = \begin{bmatrix} \Sigma x_1^2 & \Sigma x_1 x_2 & \cdots & \Sigma x_1 x_M \\ \Sigma x_1 x_2 & \Sigma x_2^2 & \cdots & \Sigma x_2 x_M \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma x_1 x_M & \Sigma x_2 x_M & \cdots & \Sigma x_M^2 \end{bmatrix}$$

위 행렬의 값은 다음처럼 구한다.

(1) 각 확률변수 x_j $(j=1,\cdots,M)$의 표본평균을 계산한다.

$$\bar{x}_j = \frac{1}{N}\sum_{i=1}^{N} x_{i,j}$$

(2) 각 확률변수 x_j $(j=1,\cdots,M)$의 분산을 계산한다.

$$s_j^2 = \frac{1}{N}\sum_{i=1}^{N}(x_{i,j}-\bar{x}_j)^2$$

(3) 두 확률변수 x_j, x_k의 공분산을 계산한다.

$$s_{j,k} = \frac{1}{N}\sum_{i=1}^{N}(x_{i,j}-\bar{x}_j)(x_{i,k}-\bar{x}_k)$$

만약 x_i $(i=1,\cdots,N)$가 다음과 같은 M차원 표본 벡터로 정의하면

$$x_i = \begin{bmatrix} x_{i,1} \\ \vdots \\ x_{i,M} \end{bmatrix}$$

표본공분산행렬 S는 다음식으로 구할수 있다.

column vector, row vector

$$S = \frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})(x_i-\bar{x})^T$$

$$S = \frac{1}{N} X_0^T X_0$$

고유값 ≥ 0

대칭행렬의 특성을 갖는다

$X_0^T X_0$: 편차값을 뺀 분산행렬 = 공분산행렬

다변수 확률변수의 공분산

M개의 다변수 확률변수 벡터

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_M \end{bmatrix}$$

의 이론적 공분산행렬은 Σ로 표기하며 다음과 같이 정의한다.

$$\Sigma = \mathrm{Cov}[X] = \begin{bmatrix} \sigma_{x_1}^2 & \sigma_{x_1x_2} & \sigma_{x_1x_3} & \cdots & \sigma_{x_1x_M} \\ \sigma_{x_1x_2} & \sigma_{x_2}^2 & \sigma_{x_2x_3} & \cdots & \sigma_{x_2x_M} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{x_1x_M} & \sigma_{x_2x_M} & \sigma_{x_3x_M} & & \sigma_{x_M}^2 \end{bmatrix}$$

$$= \begin{bmatrix} (X_1 - E[X_1])^2 & \cdots & (X_1 - E[X_1])(X_M - E[X_M]) \\ (X_1 - E[X_1])(X_2 - E[X_2]) & \cdots & (X_2 - E[X_2])(X_M - E[X_M]) \\ \vdots & \ddots & \vdots \\ (X_1 - E[X_1])(X_M - E[X_M]) & \cdots & (X_M - E[X_M])^2 \end{bmatrix}$$

다음과 같이 표현할 수도 있다.

$$\Sigma = E\left[(X - E[X])(X - E[X])^T\right]$$

$$= E\left[\begin{bmatrix} X_1 - E[X_1] \\ X_2 - E[X_2] \\ \vdots \\ X_M - E[X_M] \end{bmatrix} \begin{bmatrix} X_1 - E[X_1] & X_2 - E[X_2] & \cdots & X_M - E[X_M] \end{bmatrix}\right]$$