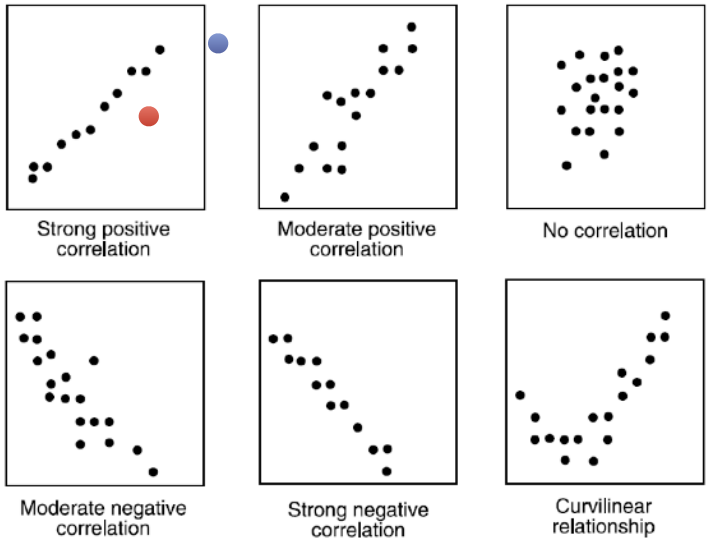
1. 개념

- 두 측정형 (적어도 순서형) 변수의 선형(직선)관계에 대한 척도
- 데이터 (x_i, y_i) 쌍으로 관측치를 활용
- 두 변수간의 관계를 시각적으로 표현하는 산점도는 두 변수 간의 함수 관계를 보여줌

2. 산점도 (SCATTER PLOT)

- 두 측정형 변수의 함수관계를 표현한 2차원 그래프
- 인과관계가 있다면 종속변수에 해당되는 변수 y-축, 설명변수에 해당되는 변수 x-축

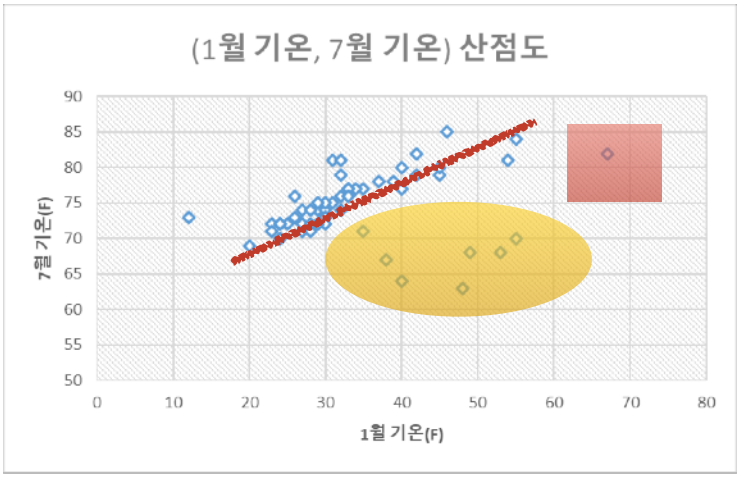


(판단 내용)

✓ 함수형태 (특히 직선, 직선관계가 해석과 활용이 용이함)

✓ 직선의 함수 관계를 벗어난 관측치 시각적 판단 : 이상치(● outlier 직선 관계를 벗어남), 영향치(● influence 데이터의 x-축 범위 내에서 벗어난 관측치)

[사례데이터] 미국도시 기후사회환경 데이터 USA_SMSA.csv



1월 기온과 7월 기온을 2차원 공간에 표현한 산점도이다. 타원의 도시, 사각형 도시를 제외하면 1월 기온이 높아지면 7월 기온도 높아지며, 강한 직선의 관계가 존재하는 것으로 보인다. 타원(이상치)의 도시는 다른 도시에 비해 7월 기온이 낮은 도시, 사각형 도시는 함수관계에 영향을 주는 영향치이다.

이상치, 영향치 진단 후 분석방법

이상치

- 입력 오류 확인 후 오류가 없다면 "제외하고" 상관계수를 재계산한다 - 이상치는 상관계수 값을 낮춘다.
- 함수관계 분석이 주목적이 아니라 함수관계를 이탈하는 관측치(예제 데이터 - 타원의 도시)에 관심이 있다면 이상 관측치에 대한 2차 분석을 한다.

영향치

✓ 입력 오류 확인 후 문제가 없다면 제외하고 상관계수 재계산한다. - 영향치는 상관계수 값을 높인다.

✓ 영향치 주변의 관측값을 추가적으로 수집한 후 분석하거나 영향치(사각형 도시)에 대한 개별 추가 분석을 실시한다.

3. 상관계수 종류

(1) 피어슨 Pearson 상관계수

- 측정형 변수 간의 선형관계 척도
- (계산식) $\rho = \frac{COV(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}}$, $\rho_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E(X)^2}\sqrt{E(Y^2) - E(Y)^2}}$. (모집단)
- (계산식) $r = r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$ (표본 데이터)
- (해석) 데이터 개수가 20개 이상이고 동일 값이 반복되지 않은 경우 $\pm 0.7 \sim \pm 1$ 이상이면 강한 상관계수, $\pm 0.3 \sim \pm 0.7$ 상관관계 존재
- (단점) 데이터 개수가 많이지면 직선 경향이 강한 것처럼 상관계수 값이 증가한다.

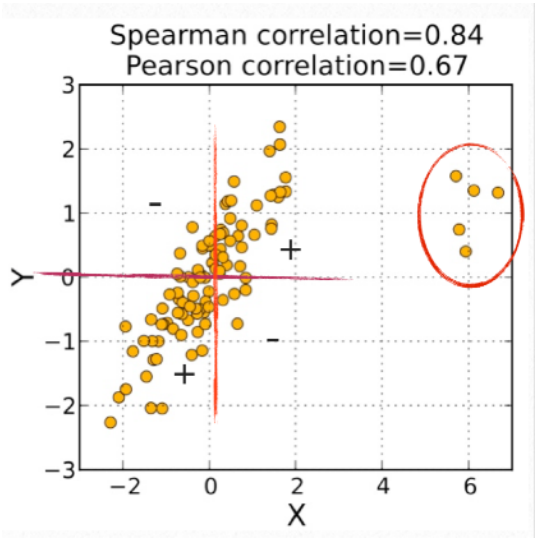
- [사례 데이터] 1월 기온과 7월 기온 피어슨 상관계수 = 0.322 `=CORREL(B3:B61,C3:C61)`

---

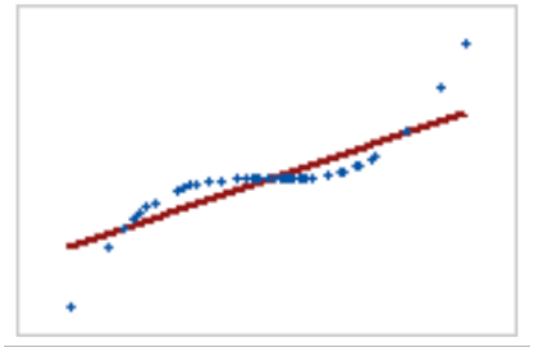
(2) 스피어맨 Spearman 순위 상관계수

- 순서형 변수 간의 선형 관계 척도
- (계산식 방법 1) $R_s = Corr(R_{x_i}, R_{y_i})$, R_{x_i}는 x_i의 순위 이며, R_{y_i}는 y_i의 순위 이다.
- (방법 2) $r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)},\ d_i = R_{X_i} - R_{Y_i}$

상관계수의 분모는 확률변수의 표준편차이므로 상관계수의 부호 를 결정하는 분자항이다. $(x_i - \bar{x})(y_i - \bar{y})$의 부호는 아래 그림(수평선은 Y의 평균, 수직선은 X의 평균, 오른쪽의 관측치 5개를 제외한 경우)에서 시각적으로 확인할 수 있음(출처 : 위키피디아).

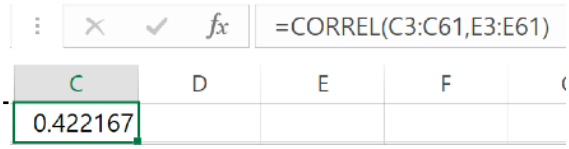


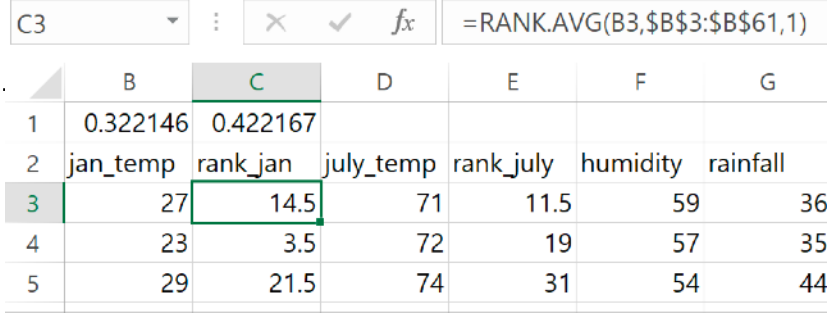
스피어만 상관계수 1 피어슨 상관계수 0.85



스피어만 상관계수는 순위의 상관계수이므로 직선의 관계가 완벽하지 않아도 순서가 동일하면 스피어만 상관계수는 1이다.

- [사례 데이터] 1월 기온과 7월 기온 스피어만 상관계수 = 0.422





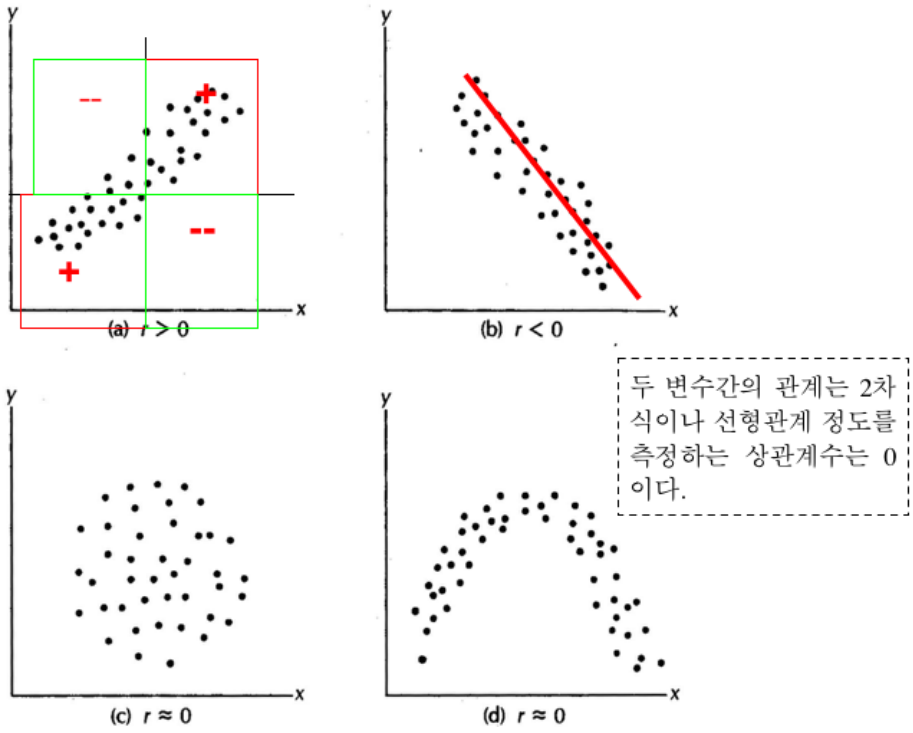
	B	C	D	E	F	G
1	0.322146	0.422167				
2	jan_temp	rank_jan	july_temp	rank_july	humidity	rainfall
3	27	14.5	71	11.5	59	36
4	23	3.5	72	19	57	35
5	29	21.5	74	31	54	44

(3) Kendall Tau 순위 상관계수

- 순서형 변수 간의 선형 관계 척도
- concordant : 쌍의 관측치 값의 크기와 순위의 크기가 일치할 때
- (계산식) $$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\frac{1}{2}n(n-1)}.$$

4. PEARSON 상관계수 추론

(계산식) $$r = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)}\sqrt{\text{var}(Y)}} = \frac{E(X-E(X))E(Y-E(Y))}{\sqrt{\text{var}(X)}\sqrt{\text{var}(Y)}} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}}$$



(1) 데이터 검증

1) 데이터는 이변량 정규분포에 근사해야 한다. 단 n>20 인 대표본에서는 문제 없음

2) x-값의 동일 값에 반복 관측치가 많은지 검증한다. x-축 관측값이 가능하면 모두 달라야 한다.

3) 산점도를 그려 데이터 범위(X-) 밖의 관측치(영향치) 존재 여부를 체크한다. - 존재한다면 제외하거나 활용 시 주의해야 한다.

4) 산점도의 이상 관측치 존재여부를 진단하여 제거 후 상관계수 유의성 검증한다.

(2) 상관관계 유의성 검정

- 귀무가설 : $H_0 : \rho = 0$ (두 측정형 변수의 상관관계는 유의하지 않음)
- 대립가설 : $H_0 : \rho \neq 0$ (두 측정형 변수의 상관관계는 유의하지 않음)
- 검정통계량 : $\frac{r - \rho_0(0)}{\sqrt{(1-r^2)/(n-2)}} \sim t(n-2)$

[사례 데이터] 1월 7월 기온 상관계수 = 0.322

=0.322/SQRT((1-0.322^2)/(59-2))

	E	F	G
	TS=	2.567809	

검정통계량 = 2.56 > 기각역 (2.002)

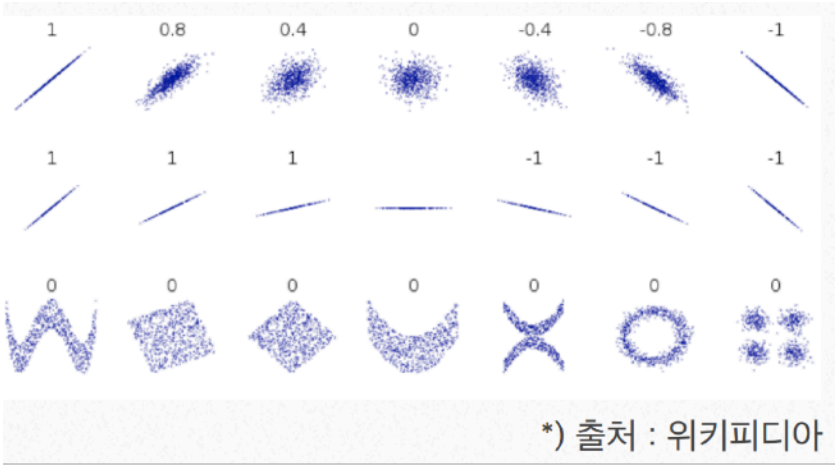
귀무가설 기각, 양의 상관관계사 유의함 - 1월 기온이 오르면 7월 기온도 높아진다.

=T.INV.2T(0.05,59-2)

---

(3) 상관계수 해석

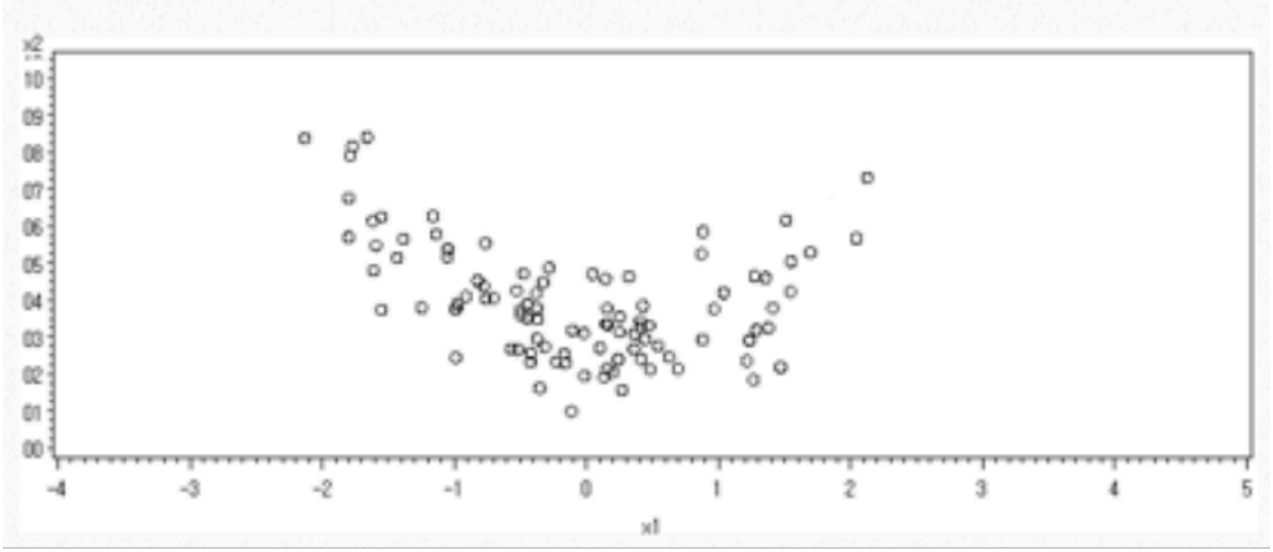
- 상관 계수는 두 변수간의 선형 관계 (linear association)에 대한 척도
- -1과 1사이의 값이다.
- 1에 가까우면 양의 선형 상관 관계가 존재한다. 한 변수의 값이 증가(감소)하면 다른 변수 값도 증가 (감소)한다.



*) 출처 : 위키피디아

- -1에 가까우면 음의 선형 상관 관계가 존재한다. 한 변수의 값이 증가(감소)하면 다른 변수 값은 감 소(증가)한다.
- 두 변수의 상관 관계가 높다는 것은 두 변수가 동 일한(comparable) 개념을 측정한다는 의미도 담 고 있다(두 변수가 유사함). 그러므로 변수를 축약 하거나 개체를 분류하는데 사용되는 다변량 분석 에서는 공분산, 혹은 상관계수 개념을 사용

- 상관 계수가 0에 가깝다는 것은 선형 상관 관계가 없다는 것이지 함수 관계가 없다는 것은 아니다. 두 변수는 이차식에 의한 ($y_i = 100 + x_i^2 - 0.4x_i$) 생성된 데이터이나 상관계수는 0에 가깝다.



(4) $H_0 : \rho = \rho_0$ 검정

- 상관관계 유의성 검정이 아니라 임의의 상관계수와 동일한지 검정
- 활용 : 미국의 경우 부자 키의 상관계수는 0.65이다. 한국의 경우 미국과 부자의 키의 상관계수가 같다고 할 수 있나? 귀무가설 : $H_0 : \rho = 0.65$
- 검정통계량 : $TS = \dfrac{\frac{1}{2}\ln(\frac{1+r}{1-r}) - \frac{1}{2}\ln(\frac{1+\rho_0}{1-\rho_0})}{1/\sqrt{n-3}} \sim N(0,1)$

[사례] 한남대생 30명을 조사한 결과 부자간 상관계수는 0.6이었다. 상관관계가 낮다고 할수 있나? 유의수준 5% 검정 => 한국, 미국 부자 키의 상관계수의 차이가 없다.

=(0.5*LN((1+B2)/(1-B2))-0.5*LN((1+0.65)/(1-0.65)))/(1/SQRT(27))

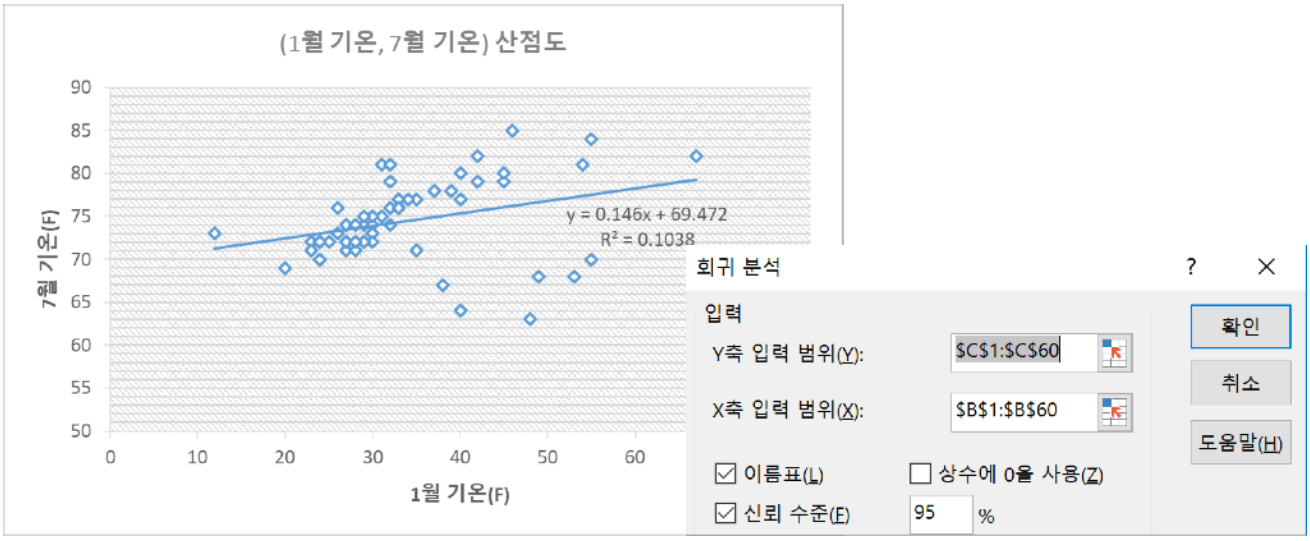
	A	B	C	D	E	F	G	H
1	표본크기	30						
2	표본상관계수	0.6						
3	귀무가설 : 한국 부자 키의 상관계수는 0.65이다.							
4	대립가설 : 한국 부자 키의 상관계수는 0.65보다 작다.							
5	검정통계량	-0.4268718	< 유의수준 5% 기각역 1.96					

(4) 독립인 2집단 상관계수 차이 검정

- 귀무가설 $H_0 : \rho_x = \rho_y$ vs. 대립가설 $H_0 : \rho_x \neq \rho_y$
- 활용 : 한국 부자 키의 상관계수와 미국 부자 키의 상관계수는 동일한가?
- 검정통계량 : $$z(x) = 0.5 \ln \frac{1+r_x}{1-r_x}, z(y) = 0.5 \ln \frac{1+r_y}{1-r_y}$$ $$z = \frac{z(x) - z(y)}{\sqrt{1/(n_x - 3) + 1/(n_y - 3)}} \sim N(0,1)$$

5. 회귀계수와 관계

- 단순 회귀모형 $y_i = \alpha + \beta x_i + e_i$에서 회귀계수 OLS 추정치 $\hat{\beta} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$
- 상관계수와 회귀계수 관계식 $\hat{\beta} = \sqrt{\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}} \times r$: 부호가 동일하며 비례관계
- 상관계수 유의성 검정과 회귀계수 기울기 유의성 검정은 동일하며 $t(n-2)$ 샘플링분포
- 단순 회귀모형에서 결정계수 Determination Coefficient $R^2 = \frac{SSR}{SST}$, $0 < R^2 < 100(\%)$
- 총변동 중 회귀변동이 차지하는 비율 : 모형의 적합 정도를 나타냄
- 상관계수의 제곱 = 결정계수 $r^2 = R^2$



7월 기온=69.5 + 0.146*(1월 기온) : 도시의 1월 기온이 올라가면 7월 기온은 0.146(F) 올라간다. 설명변수 1월 기온의 유의확률은 0.013으로 유의하다.

	계수	표준 오차	t 통계량	P-값
Y 절편	69.47158	2.004366	34.66013	5.37E-40
jan_temp	0.146026	0.056839	2.569104	0.012838

6. 상관계수 해석의 유의사항

- 양의 부호 : 한 변수 값이 커지면(작아지면) 다른 변수 값도 커진다(작아진다)
- 음의 부호 : 한 변수 값이 커지면(작아지면) 다른 변수 값도 작아진다(커진다)
- 상관관계 유의성은 크기로 결정하는 것이 아니라 검정 결과의 “유의확률”의 크기에 의해 판단
- 상관계수의 값의 크기와 상관관계 유의성은 비례하는 것은 아님 - 왜냐하면 측정변수의 관측값이 충분히 연속형이 아닌 경우 (예를 들면 일주일 교통사고 건수처럼 0. 1, 2, .., 7이면 상관계수 값는 낮을 수 있음)

사례연구

(i) ❑USA_CRIME.csv 미국 도시들의 사건 건수에 대한 데이터이다. "폭력사건"과 가장 연관성이 높은 범죄를 분석하시오.

(ii) ❑USA_SMSA.csv 미국 도시들의 사회/인구/환경 지표에 대한 데이터이다. "사망지수 Mortality Index"와 연관성이 높은 3개 지표를 분석하시오.