통계 분석

2019년 2학기

강봉주

상관 분석

상관 분석

[개요]

- 통계학에서 의존성(dependence) 또는 연관성(association)은 2개의 확률변수에 대한 통계적인 관련성을 의미한다. 때로는 인과관계(causal) 일 수도 있다.
- 상관관계(correlation)는 일종의 연관성을 나타내는 측도이며, 2개의 확률변수의 일차 관계 또는 선형 관계(linear relationship)을 나타낸다.
- 상관관계를 표현하는 측도 중의 대표적인 것이 피어슨 상관계수(Pearson correlation coefficient)

상관 분석

[상관 계수]

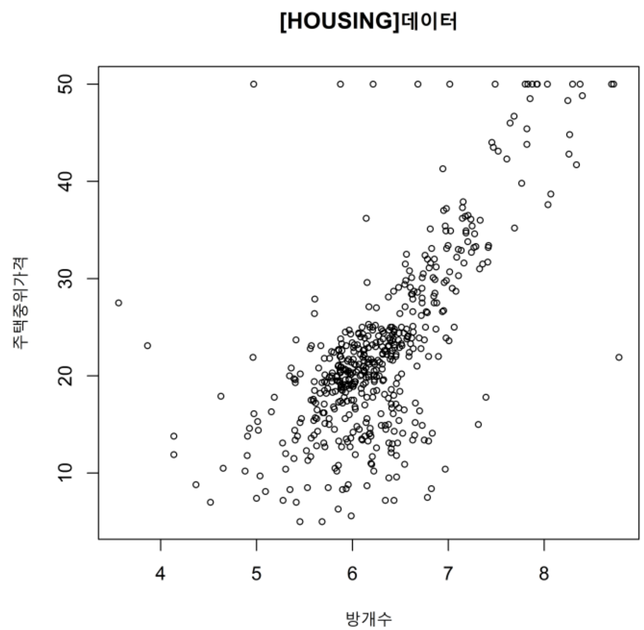
- $\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$
- X와 Y가 독립이면 $\rho_{XY} = 0$
- X가 0을 중심으로 한 대칭 분포이고 $Y = X^2$
- $\sigma_{XY} = E(XX^2) - \mu_X \mu_Y = E(X^3) - E(X)E(X^2) = 0 - 0 = 0$

- 표본 상관계수 : $\hat{\rho} = r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{(n-1) s_x s_y}$
- 표본 상관계수는 항상 -1과 1사이의 값을 갖는다.
- $(x_1 - \bar{x}, \dots, x_n - \bar{x}), (y_1 - \bar{y}, \dots, y_n - \bar{y})$ 의 $\cos(\theta)$

상관 분석

[산점도]

- 두 개의 변수 간의 관련성을 시각적으로 표현



상관 분석

[산점도]

예제)

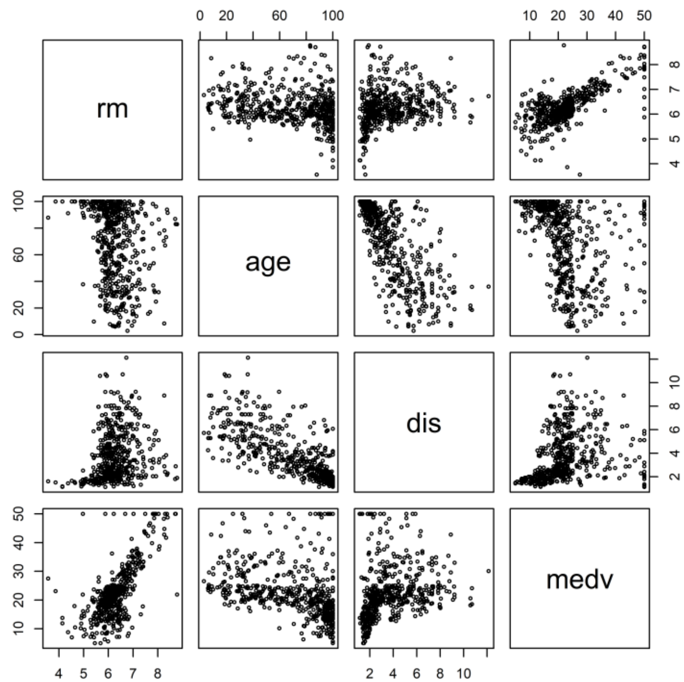
1) [HOUSING] 데이터에서 방의 개수(RM)와 주택 중위 가격(MEDV)의 상관계수를 구하세요.

$$\hat{\rho} = r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

상관 분석

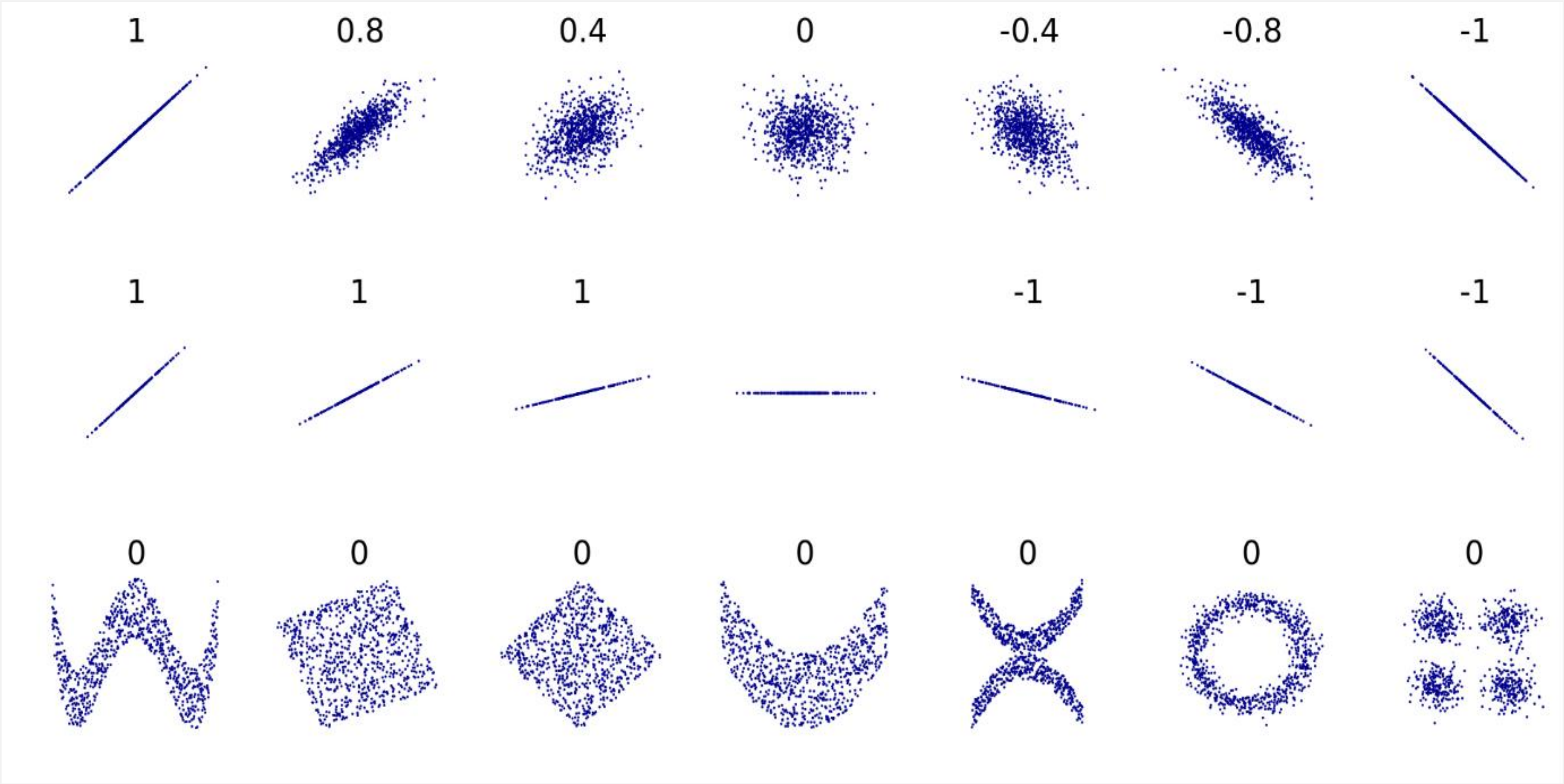
[산점도 행렬]

- pairs() 함수 이용



상관 분석

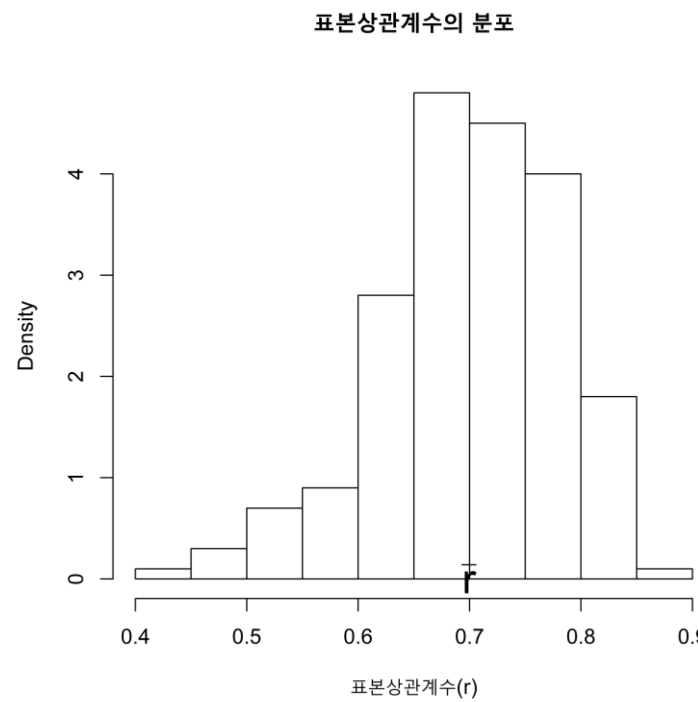
[산점도와 상관계수]



상관 분석

[표본 상관계수의 분포]

```
# 표본분포 추정
out <- c()
set.seed(1234)
num_samples <- c(200)
size <- c(100)
for (i in 1:num_samples) {
  index <- sample(1:nrow(df), size)
  sdf <- df[index, c('rm', 'medv')]
  out[i] <- cor(sdf$rm, sdf$medv)
}
```



- 왼쪽으로 완만한 분포
- 음수의 왜도값

상관 분석

[표본 상관계수의 분포]

- $R = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X - \bar{X})^2 \sum_i (Y_i - \bar{Y})^2}}$
- $T = \frac{\sqrt{n-2}\, R}{\sqrt{1-R^2}} \sim t(n-2)$
- $t_0 = \frac{\sqrt{n-2}\, r_0}{\sqrt{1-r_0^2}}$

대립 가설	P값	영가설 기각
$H_1 : \rho > 0$	$p_0 = \Pr(T \geq t_0)$	$t_0 \geq t_\alpha(n-2)$
$H_1 : \rho < 0$	$p_0 = \Pr(T \geq t_0)$	$t_0 \geq t_{1-\alpha}(n-2)$
$H_1 : \rho \neq 0$	$p_0 = \Pr(\lvert T \rvert \geq \lvert t_0 \rvert)$	$\lvert t_0 \rvert \geq t_{\frac{\alpha}{2}}(n-2)$

상관 분석

[모 상관계수에 대한 추론]

예제)

1) [HOUSING] 데이터에서 방의 개수(RM)와 주택 중위 가격(MEDV)과는 관련이 있다는 가설을 검증하세요. 단, 상관계수의 유의성으로 검증하세요.