AI 기초

2019~2020

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]

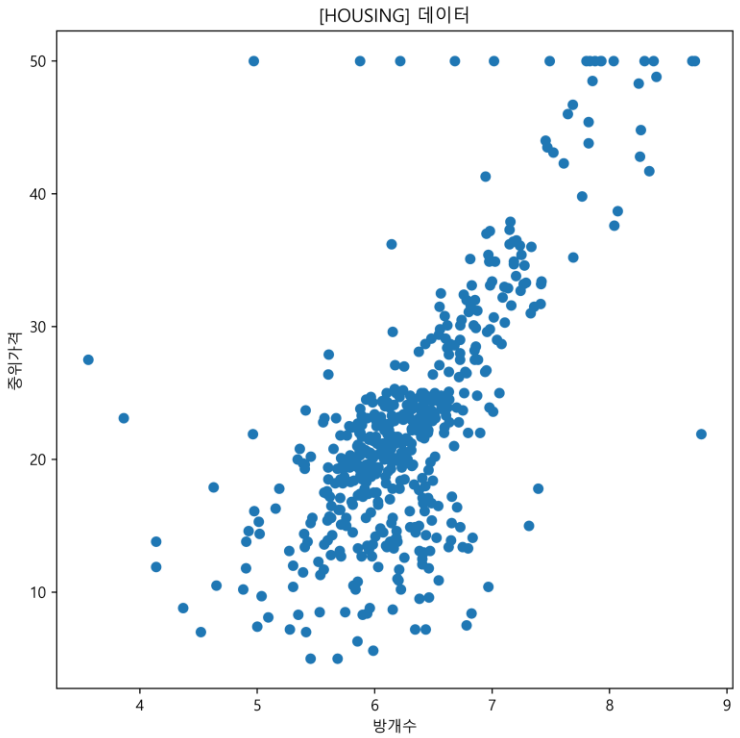
- $\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$
- X와 Y가 독립이면 $\rho_{XY} = 0$
- X가 0을 중심으로 한 대칭 분포이고 $Y = X^2$
- $\sigma_{XY} = E(XX^2) - \mu_X \mu_Y = E(X^3) - E(X)E(X^2) = 0 - 0 = 0$

- 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]

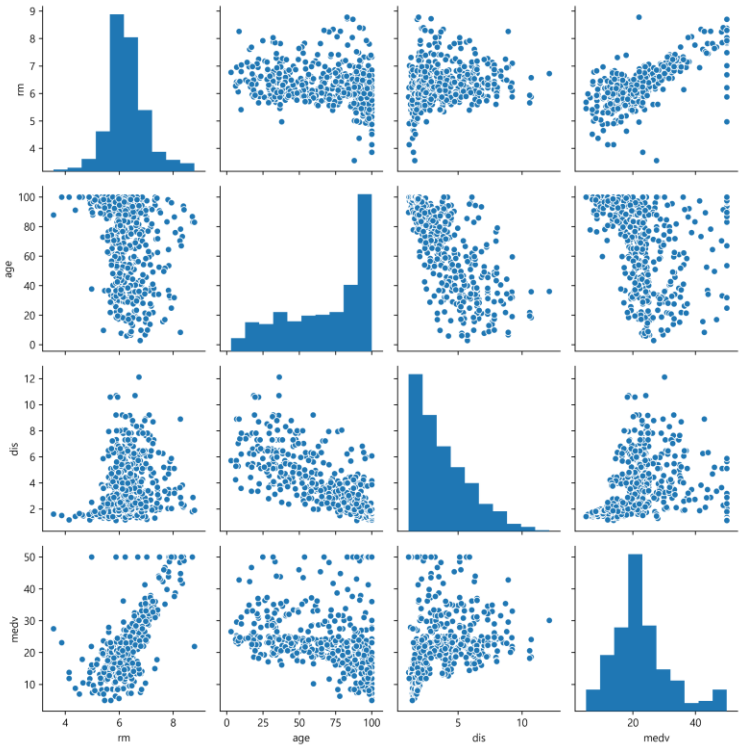
- 두 개의 변수 간의 관련성을 시각적으로 표현



상관 분석

[산점도 행렬]

- 두 개 이상의 변수 간의 관련성을 시각적으로 표현



상관 분석

[상관 계수]

예제 [HOUSING] 자료에서 ['rm', 'age', 'dis', 'medv'] 변수에 대한 상관계수를 구해보자.

In

```
# 데이터 구성: [HOUSING]

# 경로 정의
url = "https://archive.ics.uci.edu/ml/machine-learning-
databases/housing/housing.data"
df = pd.read_csv(url, sep=r'\s+', header=None)

# 컬럼 정보 주기
df.columns = ['CRIM', 'ZN', 'INDUS', 'CHAS', 'NOX',
              'RM', 'AGE', 'DIS', 'RAD', 'TAX', 'PTRATI
O', 'B', 'LSTAT', 'MEDV']
df.columns = df.columns.str.lower()

# 데이터 확인
df.shape
```

Out

```
(506, 14)
```

상관 분석

[상관 계수]

예제 [HOUSING] 자료에서 ['rm', 'age', 'dis', 'medv'] 변수에 대한 상관계수를 구해보자.

In

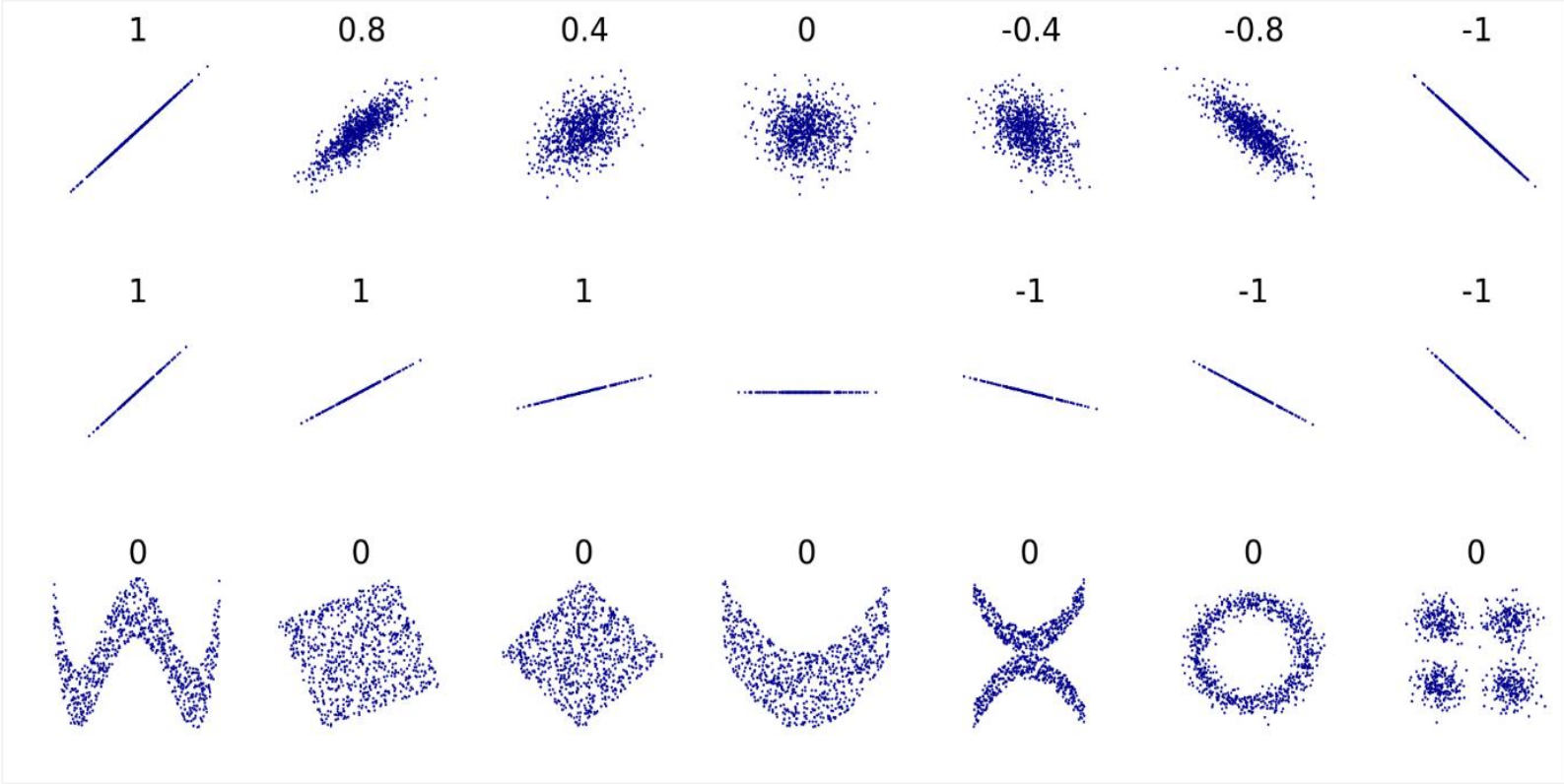
```
# 표본 상관계수의 계산
vars = ['rm', 'age', 'dis', 'medv']
print(df[vars].corr().round(3))
```

Out

```
            rm    age    dis   medv
rm       1.000 -0.240  0.205  0.695
age     -0.240  1.000 -0.748 -0.377
dis      0.205 -0.748  1.000  0.250
medv     0.695 -0.377  0.250  1.000
```

상관 분석

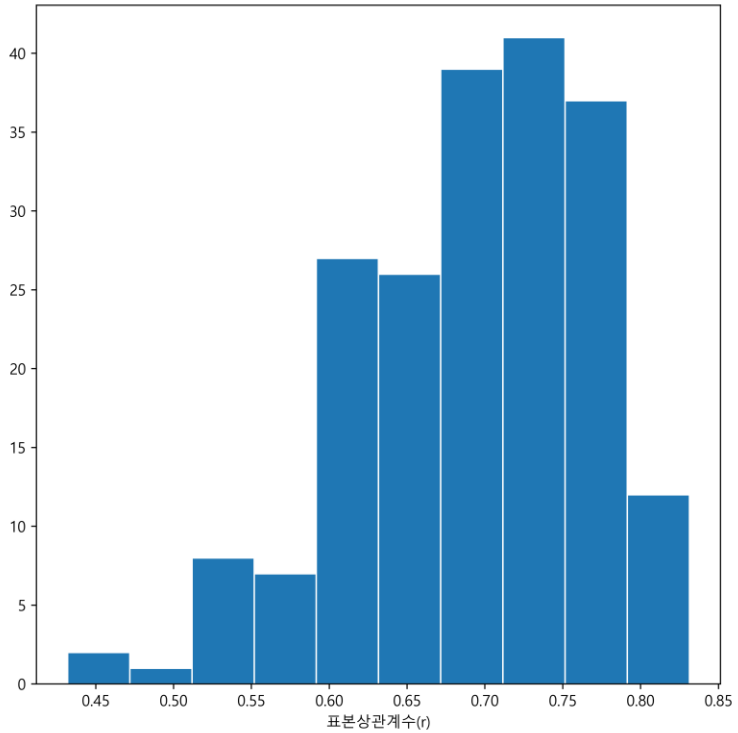
[산점도와 상관계수]



상관 분석

[표본 상관계수의 분포]

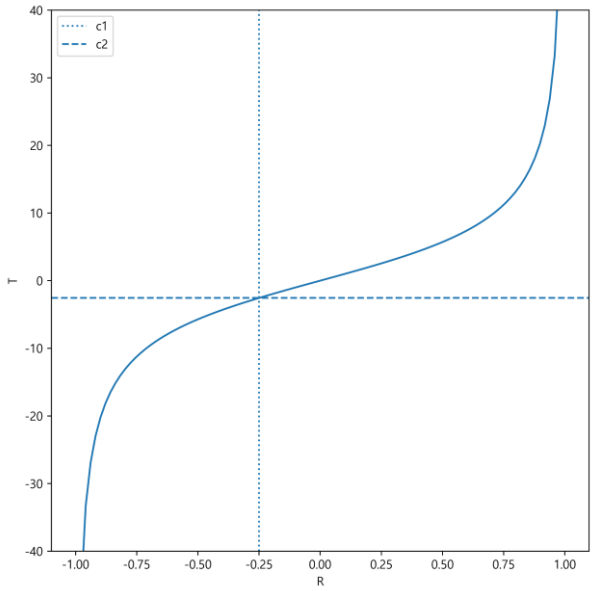
- 중복 허용, 크기가 100인 표본을 200번 추출
- 각 표본마다 방개수와 주택중위가격과의 상관계수 값 계산



상관 분석

[모 상관계수에 대한 추론]

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



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 [HOUSING] 자료에서 ['rm', 'medv'] 변수에 대한 대립가설인 $\boldsymbol{H_1}$: $\boldsymbol{\rho \neq 0}$ 에 대한 검증을 해보자.

상관 분석

[모 상관계수에 대한 추론]

예제 [HOUSING] 자료에서 ['rm', 'medv'] 변수에 대한 대립가설인 H_1: $\rho \neq 0$ 에 대한 검증을 해보자.

```
In    # 표본 상관계수의 추론
      # 대상 변수 정의
      vars = ['rm', 'medv']

      # 표본 상관계수 계산
      r = df[vars].corr().iloc[0, 1]
      r.round(3)
Out   0.695
```

상관 분석

[모 상관계수에 대한 추론]

예제 [HOUSING] 자료에서 ['rm', 'medv'] 변수에 대한 대립가설인 H_1: $\rho \neq 0$ 에 대한 검증을 해보자.

```
In      # t 변환
        n = len(df)
        t = np.sqrt(n-2) * r / np.sqrt(1 - r**2)
        t.round(3)
Out     21.722
```

상관 분석

[모 상관계수에 대한 추론]

예제 [HOUSING] 자료에서 ['rm', 'medv'] 변수에 대한 대립가설인 H_1: $\rho \neq 0$ 에 대한 검증을 해보자.

```
In      # p-value 계산
        pvalue = 2 * (1-ss.t.cdf(t, df= n-2))
        pvalue.round(5)
Out     0.0
```

상관 분석

[모 상관계수의 신뢰 구간]

- 표본 상관계수의 분포는 왼쪽으로 완만(왜도가 음수)하므로 이를 대칭으로 만들어 주어야 함
- 피셔 변환: $z = 0.5 \log\left(\frac{1+r}{1-r}\right)$
- z의 분산은 $\frac{1}{n-3}$
- $100(1-\alpha)\%$ 신뢰 구간은 $\left(z - z_{\frac{\alpha}{2}}\sqrt{\frac{1}{n-3}}\ , z + z_{\frac{\alpha}{2}}\sqrt{\frac{1}{n-3}}\right)$
- $r = \frac{e^{2z}-1}{e^{2z}+1} = \tanh(z)$
- $\left(\tanh\left(z - z_{\frac{\alpha}{2}}\sqrt{\frac{1}{n-3}}\right),\ \tanh\left(z + z_{\frac{\alpha}{2}}\sqrt{\frac{1}{n-3}}\right)\right)$

상관 분석

[모 상관계수의 신뢰 구간]

과제 [HOUSING] 자료에서 ['rm', 'medv'] 변수에 대한 모상관계수의 신뢰구간을 구하세요.