14강

회귀모형 I

통계·데이터과학과 장영재 교수

한국방송통신대학교

목차

01

상관계수

상관계수의 정의

- 산점도 (scatter plot)

  연속형인 두 변수 사이의 관계를 판단하기 위해 한 변수의 값을 X 축으로 하고 다른 변수의 값을 Y축으로 하여 관측값을 나타낸 그래프

- 상관계수(correlation coefficient)

  연속형 두 변수 간 선형관계의 강도를 나타내는 측도

- 표본공분산 C_{XY},표본상관계수 r(표준화): 두 변수 X, Y에 대해 크기 n인 표본의 관측값

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

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}\right)$$

$$r = \frac{s_{XY}}{s_X s_Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\sqrt{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}}$$

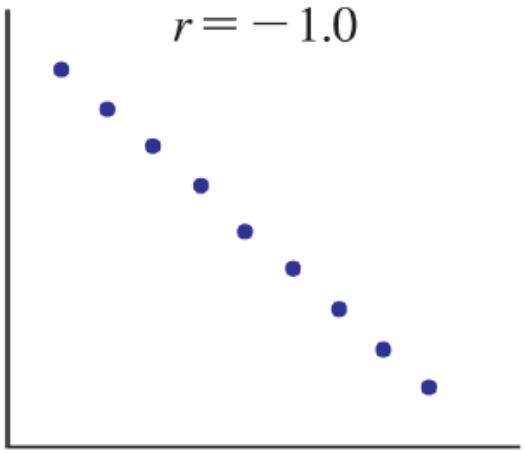
상관계수의 특징

- r은 -1과 +1 사이의 값을 가지며, r의 값이 +1에 가까울수록 강한 양의 선형관계를, -1에 가까울수록 강한 음의 상관관계를 나타내며, r의 값이 0에 가까울수록 선형관계는 약해짐

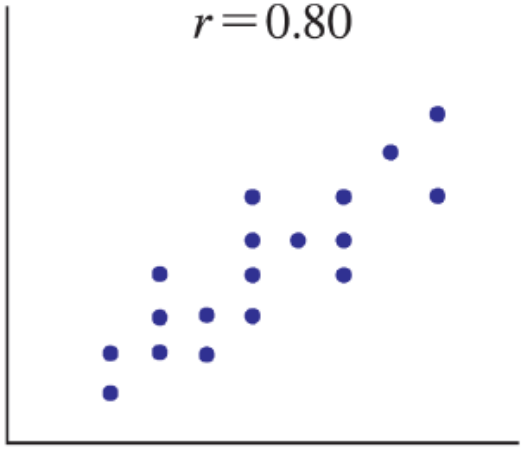
- X와 Y의 대응되는 모든 값이 한 직선상에 위치하면 r의 값은 -1(직선의 기울기가 음인 경우)이나 +1(직선의 기울기가 양인 경우)의 값을 가짐

- 표본상관계수 r은 단지 두 변수의 선형관계만 나타내는 측도이므로 두 변수의 선형상관관계는 없지만 다른 관계를 가질 때에도 r은 '0'에 가까울 수 있음

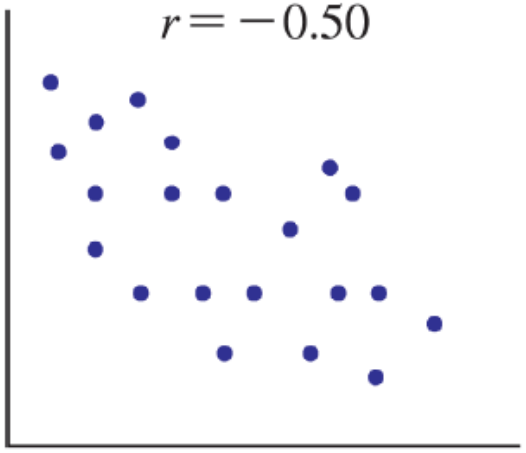
상관계수의 특징



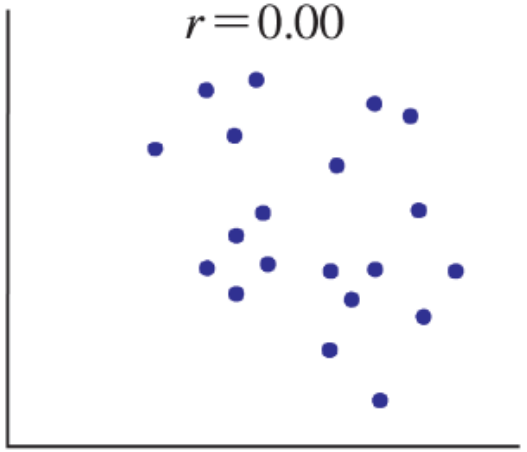
(a)



(b)



(c)



(d)

02

단순선형회귀모형의 적합

회귀분석이란

- 회귀분석(regression analysis)은 변수 간의 함수적 관련성을 구명하기 위하여 어떤 수학적 모형을 가정하고, 이 모형을 측정된 변수의 데이터로부터 추정하는 통계적 분석방법
- 주요 개념

회귀식(regression equation) : 변수 간의 관계를 나타내는 수학적 모형

종속변수(dependent variable) : 설명하고자 하는 대상이 되는 변수,
반응변수(response variable)라고도 함

독립변수(independent variable) : 종속변수에 영향을 주는 변수
설명변수(explanatory variable)라고도 함

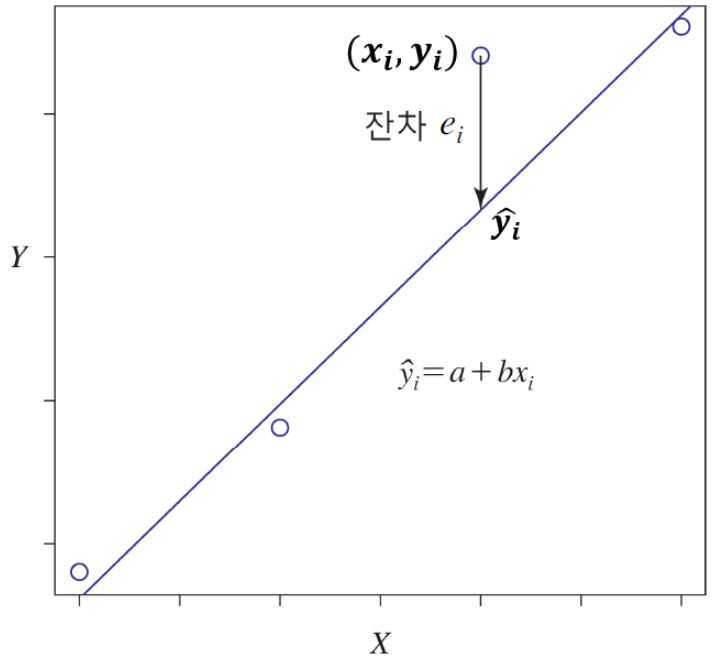
단순선형회귀모형

- 단순선형회귀모형(simple linear regression analysis)은 1개의 독립변수로 종속변수를 설명하는 선형모형

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, 2, \cdots, n$$

* α, β는 회귀계수, ε_i는 서로 독립이고 평균이 0, 분산이 σ^2인 동일한 분포를 따르는 오차항

- 추정된 회귀계수가 a, b라고 하면 $\hat{y}_i = a + bx_i$로 나타낼 수 있음



$$e_i = y_i - \hat{y}_i$$

최소제곱법

- 최소제곱법(method of least squares)은 잔차의 제곱의 합을 최소로 하는 회귀계수를 찾아 회귀식을 구하는 방법 (a, b에 관한 편미분 이용)

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y_i})^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

$$\Rightarrow \quad b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \quad a = \bar{y} - b\bar{x}$$

- 회귀계수 추정치 b(기울기)와 상관계수 r의 관계

$$b = \frac{s_{XY}}{s_{XX}} = r \cdot \frac{\sqrt{s_{XX}}\sqrt{s_{YY}}}{s_{XX}} = r \cdot \frac{\sqrt{s_{YY}}}{\sqrt{s_{XX}}}$$

회귀직선의 적합도 (표준오차와 결정계수)

- 총변동을 나타내는 제곱합(SST)을 회귀식에 의해 설명된 변동(SSR)과 설명되지 않는 잔차들의 제곱합인 오차제곱합(SSE)으로 나눌 수 있음

$$\text{제곱합: } \sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

$$\begin{array}{lccccc} & \text{SST} & = & \text{SSR} & + & \text{SSE} \\ \text{자유도:} & n-1 & = & 1 & + & n-2 \end{array}$$

- 결정계수 R^2는 총변동 SST에서 설명된 변동 SSR이 차지하는 비로서 회귀직선의 적합도를 나타내는 척도(1에 가까울수록 표본들이 회귀직선 주위에 밀집되어 있음을 의미)

$$R^2=\frac{\text{설명된 변동}}{\text{총변동}}=\frac{\text{SSR}}{\text{SST}}$$

03

단순선형회귀모형의 분석 및 추론

회귀의 분산분석

- 각 제곱합을 자유도로 나누면 SST/(n-1)는 관측값의 표본분산, SSE/(n-2) 오차의 분산 등 일종의 분산형태가 되어 분산분석표 작성이 가능

요인	제곱합	자유도	평균제곱	F비
회귀	SSR	1	$\text{MSR} = \frac{\text{SSR}}{1}$	$F_0 = \frac{\text{MSR}}{\text{MSE}}$
오차	SSE	$n-2$	$\text{MSE} = \frac{\text{SSE}}{(n-2)}$	
전체	SST	$n-1$		

- F비를 토대로 아래와 같이 가설 검정

가설: $H_0 : \beta = 0, \quad H_1 : \beta \neq 0$

검정: $F_0 = \frac{MSR}{MSE} > F_{1,\, n-2,\, \alpha}$이면 H_0를 기각

회귀분석에서의 추론

- 오차항의 정규분포 가정 하에서 $Y_i = \alpha + \beta x_i + \varepsilon_i$ 이므로
  Y_i는 평균이 $\alpha + \beta x_i$이고 분산이 σ^2 인 정규분포를 따르게 됨

- β 에 대한 추정

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$E(b) = \beta,$$

$$Var(b) = \frac{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2 Var(Y_i)\right)}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sigma^2}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

따라서 $$b \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

회귀분석에서의 추론

β 에 대한 추정

β의 $(1-\alpha)\times 100\%$ 신뢰구간 : $[b \pm t_{n-2,\ \alpha/2} \cdot SE(b)]$

$$SE(b) = s \Big/ \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

β 에 대한 검정

귀무가설: $H_0 : \beta = \beta_0$

검정통계량: $t = \dfrac{b-\beta_0}{SE(b)}$

H_0 기각역: 대립가설이 H_1: $\beta < \beta_0$이면 $t < -t_{n-2,\ \alpha}$

대립가설이 H_1: $\beta > \beta_0$이면 $t > t_{n-2,\ \alpha}$

대립가설이 H_1: $\beta \neq \beta_0$이면 $|t| > t_{n-2,\ \alpha/2}$

회귀분석에서의 추론

- 회귀직선의 절편인 모수 α의 추정치 a 는 $a = \bar{Y} - b\bar{x}$ 이므로, 오차항의 분포를 이용하여 다음과 같이 추정
- α에 대한 추정

$$E(a) = E(\bar{Y} - b\bar{x}) = \alpha + \beta\bar{x} - \beta\bar{x} = \alpha,$$

$$Var(a) = Var(\bar{Y} - b\bar{x}) = Var(\bar{Y}) + (\bar{x})^2 Var(b) - 2\bar{x}Cov(\bar{Y}, b)$$

$$= \frac{\sigma^2}{n} + \frac{\bar{x}^2\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} + 0 = \frac{\sigma^2}{n} + \frac{\bar{x}^2\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

따라서 $$a \sim N\left(\alpha, \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)\sigma^2\right)$$

회귀분석에서의 추론

α에 대한 추정

α의 $(1-\alpha)\times 100\%$ 신뢰구간 : $[a \pm t_{n-2,\,\alpha/2} \cdot SE(a)]$

$$SE(a) = s \cdot \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

α에 대한 검정

귀무가설: $H_0 : \alpha = \alpha_0$

검정통계량: $t = \dfrac{\alpha - \alpha_0}{SE(a)}$

H_0 기각역: 대립가설이 H_1: $\alpha < \alpha_0$이면 $t < -t_{n-2,\,\alpha}$

대립가설이 H_1: $\alpha > \alpha_0$이면 $t > t_{n-2,\,\alpha}$

대립가설이 H_1: $\alpha \neq \alpha_0$이면 $|t| > t_{n-2,\,\alpha/2}$

회귀분석에서의 추론

- 임의의 점 $X = x_o$에서의 종속변수 Y 는 평균값 $\mu_{Y|X} = \alpha + \beta x_o$를 가지며 이의 점추정량은 $\hat{y}_0 = a + bx_0$
- $\mu_{Y|X}$ 에 대한 추정

$\mu_{Y|X}$의 $(1-\alpha) \times 100\%$ 신뢰구간 : $[\hat{y}_0 \pm t_{n-2,\,\alpha/2} \cdot SE(\hat{y}_0)]$

$$Var(\hat{y}_0) = Var(a + bx_0) = Var(a) + (x_0)^2 Var(b) + 2x_0 Cov(a,\, b)$$

$$Cov(a,\, b) = Cov\left(\sum_{i=1}^{n} \left(\frac{1}{n} - \bar{x} \frac{(x_i - \bar{x})}{\sum_{i=1}^{n} (x_j - \bar{x})^2} \right) Y_i ,\ \frac{\sum_{i=1}^{n} (x_i - \bar{x}) Y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)$$ 이고

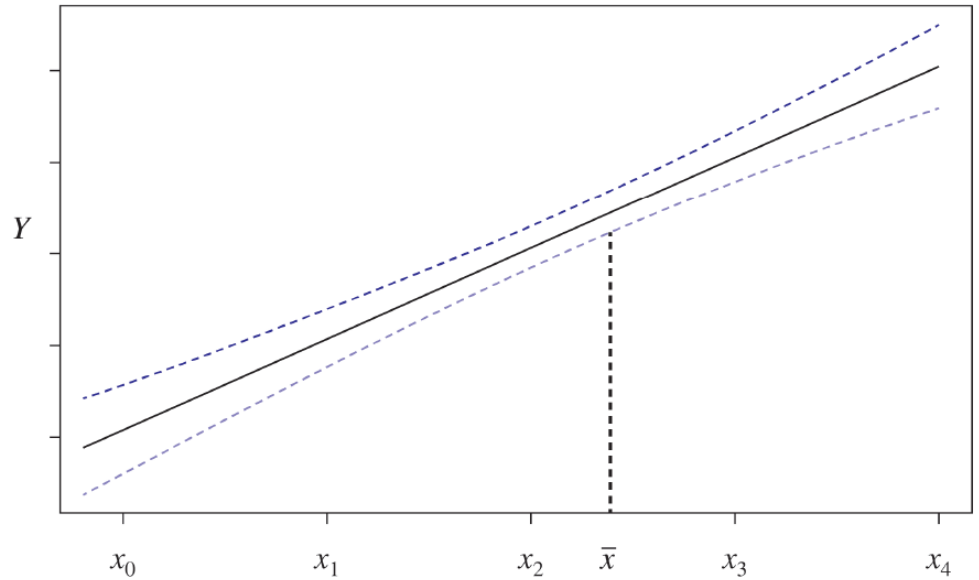
$$Cov(Y_i,\, Y_i) = \sigma^2$$

회귀분석에서의 추론

- $\mu_{Y|X}$ 에 대한 추정

$$SE(\hat{y}_0) = s \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

- 신뢰대는 $x_o = \bar{x}$ 일 때 가장 좁고 멀어질수록 점점 넓어짐



04

R을 이용한 실습

단순선형회귀모형의 적합

- lm 함수는 선형모형을 적합하는 함수

```
x <- c(56, 80, 50, 78, 65, 75, 53, 57, 53, 44)
y <- c(164, 180, 160, 175, 170, 175, 160, 169, 165, 150)
reg <- lm(y ~ x)
summary(reg)
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-5.4997 -1.3318  0.5528  0.6883  4.9094
```

단순선형회귀모형의 적합

- lm 함수는 선형모형을 적합하는 함수

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 126.42286    5.19874  24.318 8.73e-09 ***
x             0.66084    0.08349   7.915 4.71e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.169 on 8 degrees of freedom
Multiple R-squared:  0.8868,    Adjusted R-squared:  0.8726
F-statistic: 62.65 on 1 and 8 DF,  p-value: 4.715e-05
```

정리하기

- 두 변수 간 상호관계는 표본상관계수와 산점도를 이용하여 분석한다.

- 회귀모형은 변수 간의 관계를 나타내는 수학적 모형이다.
  관측값을 이용하여 모형을 추정하고, 이를 통해 변수 간의 관계를 설명하고 예측한다.

정리하기

- 회귀모형에서 서로 관계를 가지고 있는 변수 중,
  다른 변수에 의해 영향을 받는 변수를 종속변수(dependent variable)라 하고,
  종속변수에 영향을 주는 변수를 독립변수(independent variable)라고 한다.

- 단순선형회귀모형은 1개의 독립변수와 종속변수의 관계를
  설명하는 모형이다.

15강

회귀모형 I I

다음 시간 안내