15강

회귀모형 II

통계·데이터과학과 장영재 교수

한국방송통신대학교

목차

01

중선형회귀모형의 적합

중선형회귀모형

- 중선형회귀모형(multiple linear regression analysis)은 독립변수가 2개 이상 포함된 회귀모형
- 일반적으로 중선형회귀모형은 다음과 같이 행렬과 벡터를 이용하여 표현

$$Y = X\beta + \varepsilon$$

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix},\ X = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{bmatrix},\ \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix},\ \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

중선형회귀모형

- 중선형회귀모형의 간단한 사례 : 독립변수 2개와 종속변수 1개

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \quad , \quad i = 1,\ 2,\ \cdots,\ n$$

- 단순선형회귀와 유사한 방식(최소제곱법)을 이용하여 다음을 최소화하는 계수를 추정

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i})^2$$

- 각 계수에 대해 편미분한 뒤 0으로 놓고 연립방정식 해를 구함

중선형회귀모형

- 정규방정식을 만족시키는 계수 : 편회귀 계수(다른 변수 고정)

$$nb_0 + b_1\sum x_{1i} + b_2\sum x_{2i} = \sum y_i$$

$$b_0\sum x_{1i} + b_1\sum x_{1i}^2 + b_2\sum x_{1i}x_{2i} = \sum x_{1i}y_i$$

$$b_0\sum x_{2i} + b_2\sum x_{2i}^2 + b_1\sum x_{1i}x_{2i} = \sum x_{2i}y_i$$

- 오차의 분산 σ^2 의 추정값은 오차제곱합(SSE)을 잔차의 자유도로 나누어 구함

$$s^2 = \frac{1}{n-3}\sum_{i=1}^{n}(y_i - b_0 - b_1x_{1i} - b_2x_{2i})^2$$

<참고> 행렬 표현

벡터의 미분

〈참고〉 벡터의 미분

벡터 $a: n\times 1$, $x: m\times 1$와 행렬 $A: m\times m$이 주어졌을 때, A, a가 x에 관한 함수가 아니라면, 편미분 $\frac{\partial a'x}{\partial x}=\frac{\partial x'a}{\partial x}=a$이고 $\frac{\partial x'Ax}{\partial x}=(A+A')x$이다. 특히 A가 대칭행렬일 경우, $\frac{\partial x'Ax}{\partial x}=2Ax$이다.

<참고> 행렬 표현

최소제곱법

$$\mathrm{SSE} = (Y - X\boldsymbol{\beta})'(Y - X\boldsymbol{\beta}) \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

$$\begin{aligned} \frac{\partial}{\partial \beta}(Y - X\boldsymbol{\beta})'(Y - X\boldsymbol{\beta}) &= \frac{\partial}{\partial \beta}(Y' - \boldsymbol{\beta}'X')(Y - X\boldsymbol{\beta}) \\ &= \frac{\partial}{\partial \beta}(Y'Y - \boldsymbol{\beta}'X'Y - Y'X\boldsymbol{\beta} + \boldsymbol{\beta}'X'X\boldsymbol{\beta}) \\ &= \frac{\partial}{\partial \beta}(Y'Y - 2\boldsymbol{\beta}'X'Y + \boldsymbol{\beta}'X'X\boldsymbol{\beta}) \\ &= -2X'Y + 2X'X\boldsymbol{\beta} = 0 \end{aligned}$$

$$\Rightarrow \quad \boldsymbol{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \vdots \\ \widehat{\beta}_k \end{bmatrix} = (X'X)^{-1}(X'Y)$$

02

중선형회귀모형의 분석 및 추론

회귀직선의 적합도

추정의 표준오차

$$s = \sqrt{\frac{1}{n-k-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

결정계수와 수정된 결정계수(adjusted R-squared)

$$\bar{R}^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

회귀의 분산분석

단순선형회귀모형과 유사한 방식으로 분산분석표 작성

요인	제곱합	자유도	평균제곱	F비
회귀	SSR	k	$\mathrm{MSR}=\dfrac{\mathrm{SSR}}{k}$	$F=\dfrac{\mathrm{MSR}}{\mathrm{MSE}}$
오차	SSE	$n-k-1$	$\mathrm{MSE}=\dfrac{\mathrm{SSE}}{(n-k-1)}$	
전체	SST	$n-1$		

제곱합: $\mathrm{SST}=\mathrm{SSE}+\mathrm{SSR}$

자유도: $n-1=(n-k-1)+k$

F비를 토대로 아래와 같이 가설 검정

$H_0:\ \beta_1=\beta_2=\ \cdots\ =\beta_k=0$

$H_1:$ k개의 β_i 중 적어도 하나는 0이 아니다.

$F>F_{k,\,n-k-1,\,\alpha}$이면 H_0를 유의수준 α하에서 기각

회귀분석에서의 추론

- 회귀계수의 점추정량들의 분포

$$b_i \sim N(\beta_i,\ c_{ii} \cdot \sigma^2) \quad (\text{단},\ i = 0,\ 1,\ \cdots,\ k)$$

c_{ii}는 $(k+1) \times (k+1)$행렬인 $(X'X)^{-1}$의 i번째 대각원소

- 모수 σ^2대신 추정량 s^2을 사용하면 회귀계수에 관한 추론이 가능

점추정량: b_i

표준오차: $SE(b_i) = \sqrt{c_{ii}} \cdot s$

신뢰구간: $b_i \pm t_{n-k-1,\ \alpha/2} \cdot SE(b_i)$

회귀분석에서의 추론

가설검정

귀무가설: $H_0: \beta_i = \beta_{i0}$

검정통계량: $t = \dfrac{b_i - \beta_{i0}}{SE(b_i)}$

H_0 기각역: 대립가설이 $H_1: \beta_i < \beta_{i0}$이면 $t < -t_{n-k-1,\ \alpha}$

대립가설이 $H_1: \beta_i > \beta_{i0}$이면 $t > t_{n-k-1,\ \alpha}$

대립가설이 $H_1: \beta_i \neq \beta_{i0}$이면 $|t| > t_{n-k-1,\ \alpha/2}$

03

회귀진단

회귀진단의 의의

- 회귀모형을 세우고 계수에 대한 추정 및 검정을 실시한 이후에는 적합된 모형이 안정적인지, 가정이 타당한지 세부적으로 검토하는 과정이 필요
- 잔차분석을 통해 가정 위배 여부 검토
- 이상점(outlier)이나 영향점(influential point) 검토
- 독립변수 간의 상관관계를 검토하여 모형의 안정성 검토

잔차분석

- 회귀모형에서 모수에 대한 추론은 오차항 가정에 기초함
- 오차항은 관측될 수 없으므로 추정량인 잔차를 이용하여 가정의 타당성을 검토

회귀분석에서의 가정

A1: 가정된 모형 $y_i = \alpha + \beta x_i + \varepsilon_i$는 옳다.

A2: 오차 ε_i의 평균값은 0이다.

A3: (등분산성) 모든 ε_i의 분산은 σ^2으로 동일하다.

A4: (독립성) 오차 ε_i들은 서로 독립이다.

A5: (정규성) 오차 ε_i들은 정규분포를 따른다.

<산점도>

① 잔차 대 예측값(즉, e_i 대 $\hat{y}_i$): A3

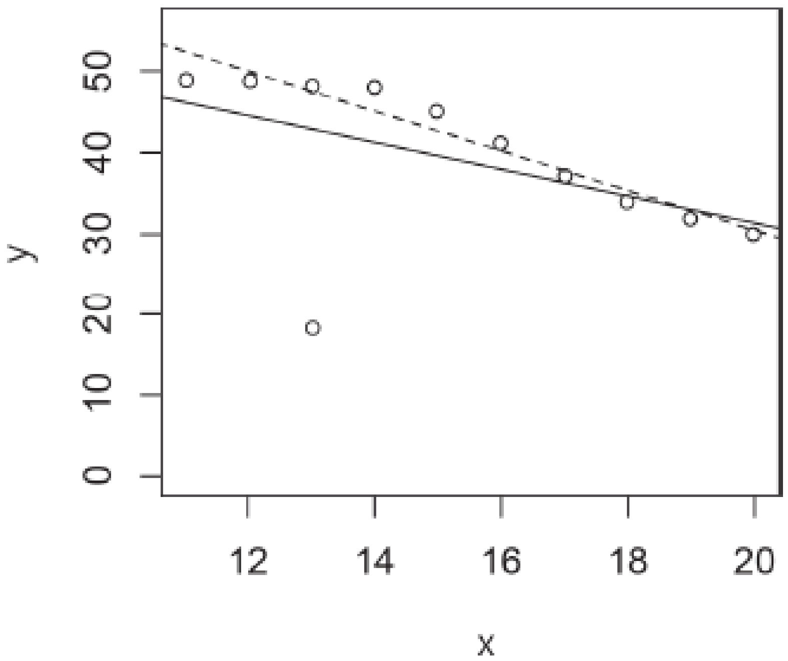
② 잔차 대 독립변수(즉, e_i 대 x_i): A1

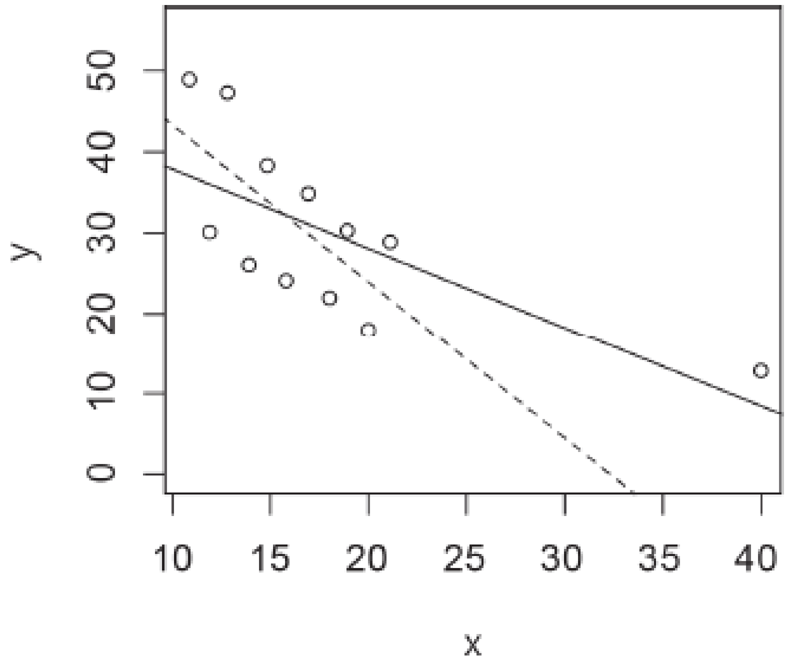
③ 잔차 대 관측순서(즉, e_i 대 i): A2, A4

- A5는 잔차들의 히스토그램이나 정규확률도로 검토

이상점과 영향점

- 이상점(좌측)과 영향점(우측)
- 종속변수의 분포와 독립변수의 분포
- R^2와 직선의 기울기





변수 간 상관성 검토

- 독립변수 간 상관관계가 있을 경우 모형이 불안정
  : 독립변수들 간의 선형관계를 다중공선성(multicollinearity)이라 함

- 분산팽창인수(Variance Inflation Factor: VIF)로 다중공선성 판단

- 더 많은 관측값 수집, 변수 선택 후 모형 적합, 독립변수들의 표준화, 주성분분석 등의 방법을 사용

04

R을 이용한 실습

중선형회귀모형의 적합

lm 함수는 선형모형을 적합하는 함수

```
x1 <- c(4, 6, 6, 7, 8, 9, 9, 9, 11, 12)
x2 <- c(3, 4, 5, 5, 6, 7, 6, 8, 8, 9)
y <- c(38, 42, 46, 47, 50, 53, 52, 56, 58, 62)
reg1 <- lm(y ~ x1 + x2)
summary(reg1)
Call:
lm(formula = y ~ x1 + x2)


Residuals:
     Min        1Q    Median        3Q       Max
-0.90566 -0.30458 -0.02695  0.41442  0.78706
```

중선형회귀모형의 적합

lm 함수는 선형모형을 적합하는 함수

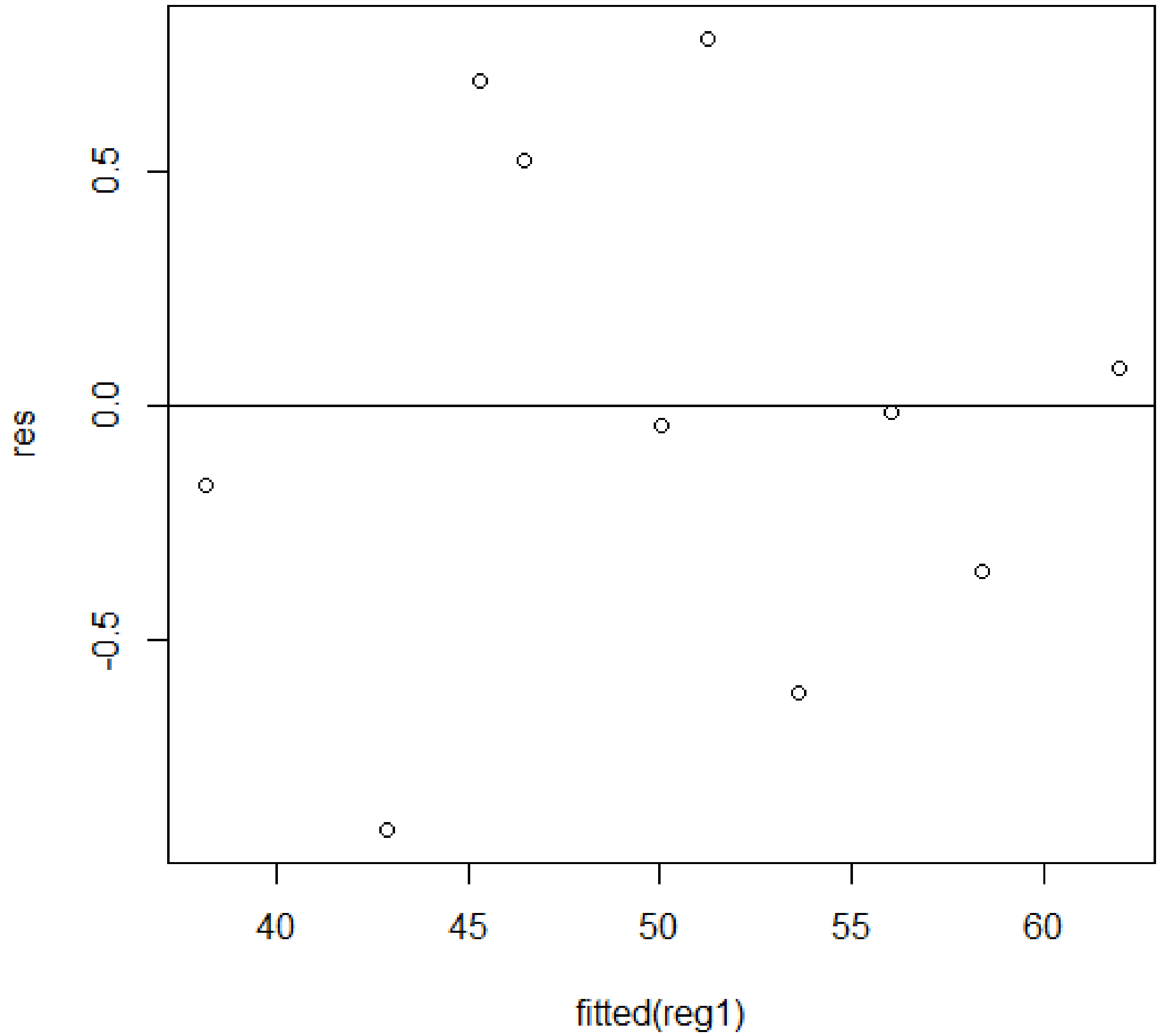
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  26.2911     0.7234  36.343  3.1e-09 ***
x1            1.1698     0.2942   3.976 0.005351 **
x2            2.3989     0.3731   6.430 0.000357 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6249 on 7 degrees of freedom
Multiple R-squared:  0.9944,    Adjusted R-squared:  0.9928
F-statistic: 621.9 on 2 and 7 DF,  p-value: 1.311e-08
```

잔차분석

잔차 산점도

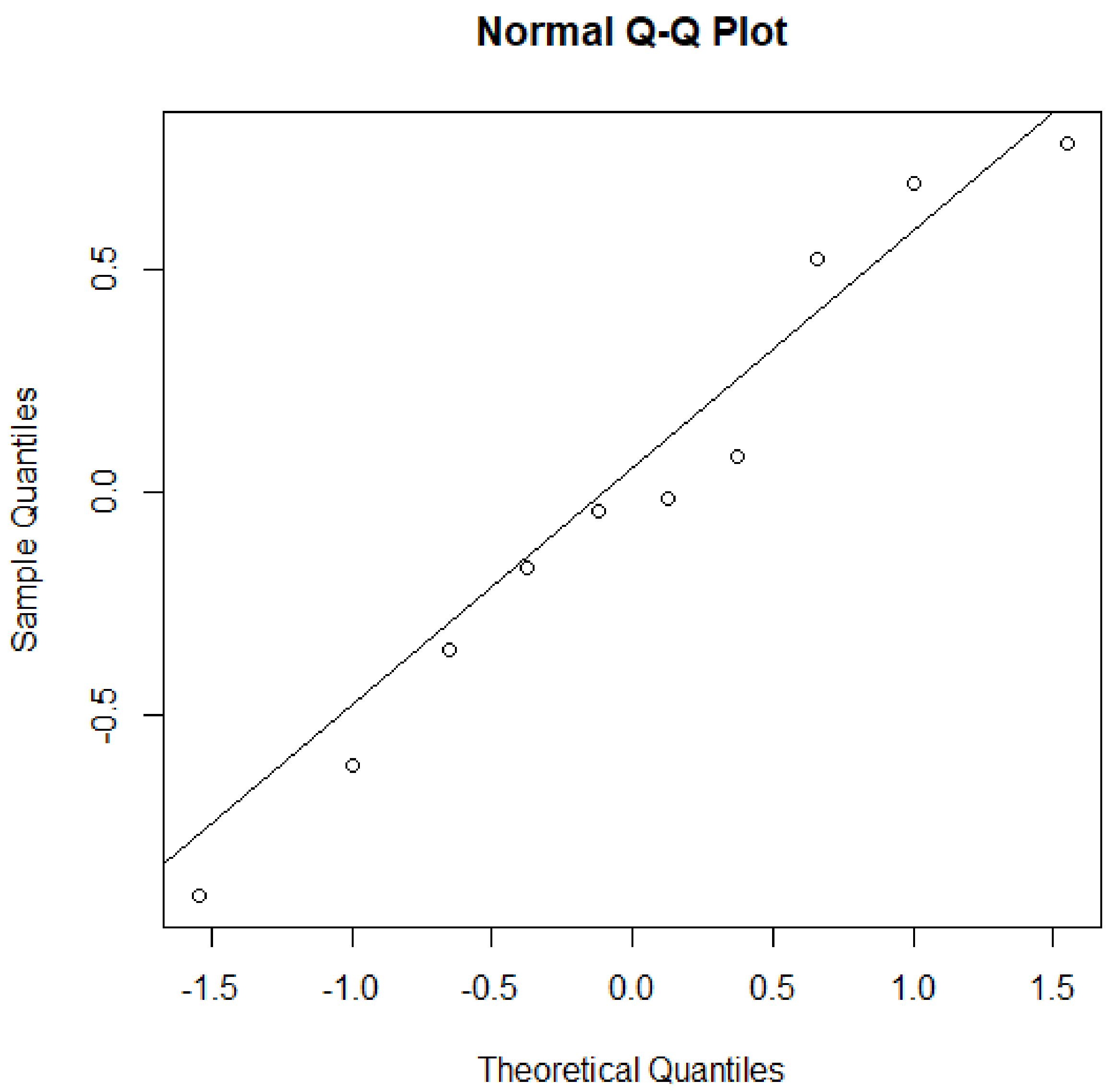
```
res <- resid(reg1)
plot(fitted(reg1), res)
abline(0, 0)
```



잔차분석

정규확률도

```
qqnorm(res)
qqline(res)
```



정리하기

- 중선형회귀모형은 하나의 종속변수와 여러 개의 독립변수 사이의 관계를 나타낸다.
- 회귀모형에서 계수는 잔차의 제곱합을 최소로 하는 값으로 정하며, 회귀직선을 추정한 후에는 추정량의 표준오차와 결정계수 등을 이용하여 그 회귀식이 얼마나 타당한가를 검토해야 한다.
- 회귀진단은 적합된 모형이나 가정에 대해 종합적으로 검토하는 과정이다.

한 학기 동안
수고 많으셨습니다.

한국방송통신대학교