2장 Review

■ 기본개념

$E(Y|X=x) = \beta_0 + \beta_1 x$ → 선형함수

$Var(Y|X=x) = \sigma^2$ → 일정함

– 단순선형회모형

모수(회귀계수) : 자료를근거로추정해야함

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1,2,\cdots,n$$

$$E(\epsilon_i) = 0$$
$$Var(\epsilon_i) = \sigma^2$$
$$Cov(\epsilon_i, \epsilon_j) = 0, \quad i \neq j$$

– 대체모형

추정량 : $\bar{y}$ → β_1의 추정에 영향을받지않음

$(Cov(\hat{\alpha}, \hat{\beta}_1) = 0)$

$$y_i = \alpha + \beta_1 (x_i - \bar{x}) + \epsilon_i$$

새로운설명변수

2

2

2장 Review 모수(회귀계수) β_0, β_1의추정

■ 최소제곱법 : 2차원산점도에서 n개의 점들(자료값)을 잘 설명하는 직선을 구함

– $Q(\beta_0, \beta_1) = \Sigma(y_i - \beta_0 - \beta_1 x_i)^2$를 최소화 하는 β_0, β_1

잔차(점~직선의 수직거리)

– 정규방정식 $(\partial Q/\partial\beta_0 = 0, \ \partial Q/\partial\beta_1 = 0)$ 편미분값=0 인 해

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \hat{\beta}_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

$\hat{\beta}_0$, $\hat{\beta}_1$을 찾는것은 직선을 찾는것과같음

– 추정회귀식, 적합값, 잔차

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad \Rightarrow \quad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

적합값

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

SS : sum of square

MS : mean of square

– σ^2의 추정량

오차의분산

$$\hat{\sigma}^2 = \frac{\Sigma(y_i - \hat{y}_i)^2}{n-2} = \frac{SSE}{n-2} = MSE$$

잔차의제곱합

σ^2의 비편향추정량 : $E(\hat{\sigma}^2) = \sigma^2$

3

3

2장 Review

■ 최소제곱추정량의 특성

추정량의 기댓값=모수

1) 최소제곱추정량 $\hat{\beta}_0$와 $\hat{\beta}_1$은 비편향추정량(unbiased estimator)

$$E(\hat{\beta}_0) = \beta_0 \qquad E(\hat{\beta}_1) = \beta_1$$

2) 최소제곱추정량 $\hat{\beta}_0$와 $\hat{\beta}_1$은 y_i의 선형결합

$$\hat{\beta}_1 = \Sigma k_i y_i \qquad k_i = (x_i - \bar{x})/S_{xx}$$
$$\hat{\beta}_0 = \Sigma m_i y_i \qquad m_i = 1/n - \bar{x} k_i$$

3) 최소제곱추정량 $\hat{\beta}_0$와 $\hat{\beta}_1$은 ϵ_i의 선형결합

$$\hat{\beta}_1 = \beta_1 + \Sigma k_i \epsilon_i \qquad \hat{\beta}_0 = \beta_0 + \Sigma m_i \epsilon_i$$

4) $\hat{\beta}_1$와 $\hat{\beta}_0$의 분산

$$Var(\hat{\beta}_1) = \sigma^2 / S_{xx} \qquad Var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) \qquad Cov(\hat{\beta}_0, \hat{\beta}_1) = -\bar{x} \frac{\sigma^2}{n}$$

① $Var(\hat{\beta}_1)$이 작아지려면 S_{xx}가 커져야함
S_{xx}는 설명변수가 퍼질수록 커짐
→ 정확한추정

② $Var(\hat{\beta}_0)$이 작아지려면 S_{xx}가 커져야함
$\bar{x}^2$ (또는 $\bar{x}$)이 0에가까워야함

③ $Cov(\hat{\beta}_0, \hat{\beta}_1)$은 일반적으로 0이 아님 → β_0, β_1의 추정은 서로영향을 받음
$\bar{x}=0$인경우 $Cov(\hat{\beta}_0, \hat{\beta}_1)=0$ → β_0, β_1의 추정은 서로의 영향을 받지않음

5) 최소제곱법에서의 $\Sigma e_i = 0$, Σe_i^2 은 다른 추정량의 잔차제곱합보다 작음

4

4

2장 Review

6) 가우스-마코브(Gauss-Markoff)정리

y의 선형결합이면서 비편향추정량인 것은 여러개
그 중 최소제곱추정량의 분산이 가장작음

오차의 독립성, 등분산성의 가정하에서 최소제곱추정량이
최량선형비편향추정량(best linear unbiased estimator: BLUE)

$$\tilde{\beta}_1 = \Sigma c_i y_i \quad E(\tilde{\beta}_1) = \beta_1 \qquad Var(\hat{\beta}_1) \le Var(\tilde{\beta}_1)$$
$$\tilde{\beta}_0 = \Sigma d_i y_i \quad E(\tilde{\beta}_0) = \beta_0 \qquad Var(\hat{\beta}_0) \le Var(\tilde{\beta}_0)$$

7) 적합값 $\hat{y}_i$의 합과 관측값 y_i의 합은 같음: $\Sigma y_i = \Sigma \hat{y}_i$

8) x_i, $\hat{y}_i$을 가중값으로 하는 잔차의 가중합은 0: $\Sigma x_i e_i = 0$, $\Sigma \hat{y}_i e_i = 0$

9) 추정회귀선 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$은 반드시 $(\bar{x}, \bar{y})$를 통과

10) 잔차제곱합은 $SSE = SS_{yy} - \hat{\beta}_1 S_{xy}$

5

5

2장 Review 등분산이 아닌경우 모수추정

■ 가중최소제곱법(method of weighted least squares)

- 등분산가정이 만족되지 지 못하는 경우

  : 최소제곱법에 의한 추정량은 최량선형비편향추정량이 아님

  선형비편향추정량은 맞음!

- 가중오차제곱 Q를 최소화하는 회귀계수를 추정하는 방법

최소제곱법에서 최소제곱추정량을 구한것과 같이
① Q를 β_0, β_1으로 편미분해 0으로 놓은 정규방정식 구하기
② 그 해가 가중최소제곱법으로 구한 추정량

$$Q(\beta_0, \beta_1) = \Sigma \omega_i (y_i - \beta_0 - \beta_1 x_i)^2$$

가중값 (0보다 큰 상수) → 모든 가중값이 1인경우는 최소제곱법

- 가중값 ω_i는 반응변수의 분산의 역수에 비례하도록 설정 : 적절히 정해야함

  - n_i개 자료의 평균, $Var(y_i) = \sigma^2 / n_i \propto 1/n_i$: $\omega_i = n_i$
  - n_i개 자료의 합, $Var(y_i) = n_i \sigma^2 \propto n_i$: $\omega_i = 1/n_i$
  - 분산이 x_i에 비례, $Var(y_i) \propto x_i$: $\omega_i = 1/x_i$

6

6

2장 Review 오차의 확률분포를 고려하는경우 : 오차가 정규분포를 따른다고 가정

■ 최대가능도방법(method of maximum likelihood))

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \ \ i = 1, \ldots n$$

$$E(\epsilon_i) = 0, \ \ Var(\epsilon_i)\sigma^2, \ \ Cov(\epsilon_i, \epsilon_j) = 0, i \neq j$$

- 분포에 대한 가정 $\epsilon_i \sim N(0, \sigma^2)$

- 가능도함수 : 결합확률밀도함수와 같은형태

$$L(\beta_0, \beta_1, \sigma^2 | y_1, \cdots, y_n) = f(y_1, \cdots, y_n | \beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} f(y_i)$$

- 최대가능도 추정량: 가능도함수를 최대화하는 β_0, β_1, σ^2

$$\underbrace{\hat{\beta}_{0_{MLE}} = \hat{\beta}_0, \quad \hat{\beta}_{1_{MLE}} = \hat{\beta}_1,}_{\text{최소제곱추정량과 같음}} \qquad \hat{\sigma}^2_{MLE} = \frac{SSE}{n}$$

→ 편향추정량 ($MSE = \frac{SSE}{n-2}$ 가 아님!, 기댓값이 σ^2이 아님)

7

7

연습문제

2.1 (예 1.1 계속) 포도수확량 자료

1장의 (예 1.1)에서 나온 자료에서 열매 개수를 설명변수, 수확량을 반응변수로 하여 다음 물음에 답하라.

2.1.1 최소제곱법에 의해 회귀계수(절편과 기울기)를 추정하고 추정회귀식을 구하라.

2.1.2 회귀계수의 추정량은 y_i의 선형결합의 형태로 표현된다고 하였다. 2.3절의 식 (2.12)와 ~~(2.13)~~에서 계수 k_i와 ~~m_i~~의 값을 구하고 이 계수들을 이용하여 회귀계수의 추정값을 계산하고 2.1.1에서의 결과와 비교하라.

2.1.3 잔차 $e_i = y_i - \hat{y}_i$를 구하고 이들의 제곱합으로 SSE의 값을 구하라.

2.1.4 SSE의 값을 공식 (2.21)에 의해 구하라.

2.1.5 σ^2의 추정값을 구하라.

2.1.6 Σe_i의 값이 0임을 확인하고, $\hat{y}_i$의 값을 x축으로 e_i의 값을 y축으로 하는 산점도를 그리고 잔차가 특정한 패턴을 따르는지 확인하라.

~~**2.1.7** 2.1.2에서 구한 k_i 및 m_i의 값을 이용하여 $\Sigma k_i = 0$, $\Sigma k_i x_i = 1$ 및 $\Sigma m_i = 1$, $\Sigma m_i x_i = 0$이 됨을 수치적으로 확인하라.~~

8

8

연습문제

2.1

$$\sum x_i = 1285.21 \quad \sum y_i = 53.7 \quad \sum x_i^2 = 140168.7 \quad \sum y_i^2 = 248.29 \quad \sum x_i y_i = 5880.88$$

$$S_{xx} = \Sigma(x_i - \bar{x})^2 = \Sigma x_i^2 - (\Sigma x_i)^2/n =$$

$$S_{xy} = \Sigma(x_i - \bar{x})(y_i - \bar{y}) = \Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)/n =$$

$$S_{yy} =$$

2.1.1 $\hat{\beta}_1 =$ $\hat{\beta}_0 =$ $\Rightarrow$ $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x =$

2.1.2 $k_1 = (x_1 - \bar{x})/S_{xx} =$

$k_2 =$

$\vdots$

$k_{12} =$

$$k_1 y_1 + k_2 y_2 + \ldots + k_{12} y_{12} = \quad ? \quad = \hat{\beta}_1$$

9

9

연습문제

2.1.3 $\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 x_1 =$ $e_1 = y_1 - \hat{y}_1 =$

$\hat{y}_2 =$ $e_2 =$

$\vdots$ $\vdots$

$\hat{y}_{12} =$ $e_{12} =$

$SSE = e_1^2 + e_2^2 + \ldots + e_{12}^2 =$

2.1.4 $SSE = S_{yy} - \hat{\beta}_1 S_{xy} =$

2.1.5 $\hat{\sigma}^2 = \dfrac{SSE}{n-2} =$

2.1.6 $\sum e_i = e_1 + \ldots + e_{12} = \quad ? \quad = 0$

산점도: 가로축 $\hat{y}_i$ 세로축 e_i

10

10

연습문제

2.3 Forbes 자료

스코틀랜드 물리학자인 J. D. Forbes는 물의 끓는 온도를 이용하여 해발고도를 알아보고자 하였다. 그는 대기압력을 통해 해발고도를 알 수 있다는 현상을 알고 여러 가지 조건에서 대기압력과 물의 끓는 온도를 동시에 측정하고 그 관계를 규명하였다. 1800년 중반에 기압계는 휴대하기 어려운 기구였으므로 Forbes는 물의 끓는 온도를 이용하여 대기압력을 추정하고 이를 토대로 해발고도를 측정하고자 한 것이다.

물의 끓는 온도(°F)	압력(Hg)	$100 \times \log_{10}$(압력)
194.5	20.79	131.79
212.2	30.06	147.80

자료원 : Weisberg(2014)

2.3.1 압력 $= \beta_0 + \beta_1$(온도) $+ \varepsilon$의 모형을 적합시켜 회귀계수의 추정값 $\hat{\beta}_0$과 $\hat{\beta}_1$, $\hat{\sigma}^2$, e_i, $\hat{y}_i$를 구하라.

2.3.2 $100\log_{10}$(압력) $= \beta_0 + \beta_1$(온도) $+ \varepsilon$의 모형을 적합시켜 회귀계수의 추정값 $\hat{\beta}_0$와 $\hat{\beta}_1$, $\hat{\sigma}^2$, e_i, $\hat{y}_i$를 구하라.

2.3.3 2.3.1과 2.3.2에서의 두 모형을 비교하라.

11

11

연습문제

2.3

2.3.1

$\sum x_i = 3450.2 \quad \sum y_i = 426 \quad \sum x_i^2 = 700759 \quad \sum y_i^2 = 10821 \quad \sum x_i y_i = 86735.5$

$\bar{x} =$ $\qquad \bar{y} =$

$S_{xx} = \Sigma(x_i - \bar{x})^2 = \Sigma x_i^2 - (\Sigma x_i)^2/n =$

$S_{xy} = \Sigma(x_i - \bar{x})(y_i - \bar{y}) = \Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)/n =$

$S_{yy} =$

$\hat{\beta}_1 =$ $\qquad \hat{\beta}_0 =$

$SSE = S_{yy} - \hat{\beta}_1 S_{xy} =$

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i =$

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

12

12

연습문제

2.3.2

$\sum x_i = 3450.2 \quad \sum y_i = 2373.29 \quad \sum x_i^2 = 700759 \quad \sum y_i^2 = 331751.6 \quad \sum x_i y_i = 482141.5$

2.3.3 두 모형의 비교

설명변수 – 온도

1. 반응변수 – 압력
2. 반응변수 – 100 x $\log_{10}$ (압력)

13

13