김은영

제2장 단순선형회귀모형

2.3 회귀계수의 추정

2.3.2 최소제곱추정량의 성질

1. 평균: 불편성

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i - \sum_{i=1}^{n}(x_i-\bar{x})\bar{y}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{S_{xx}} = \sum_{i=1}^{n} w_i y_i$$

여기서, $w_i = \frac{x_i-\bar{x}}{S_{xx}}$

→ $\hat{\beta}_1$은 y_i들의 선형결합

$$\sum_{i=1}^{n} w_i = \sum_{i=1}^{n}\frac{x_i-\bar{x}}{S_{xx}} = \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i-\bar{x}) = 0$$

$$\sum_{i=1}^{n} w_i^2 = \sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{S_{xx}}\right)^2 = \frac{1}{S_{xx}^2}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{S_{xx}}$$

$$\Rightarrow E(\hat{\beta}_1) = E\left(\sum_{i=1}^{n} w_i y_i\right) = \sum_{i=1}^{n} w_i E(y_i) = \sum_{i=1}^{n} w_i(\beta_0+\beta_1 x_i) = \beta_0\sum_{i=1}^{n} w_i + \beta_1\sum_{i=1}^{n} w_i x_i = \beta_1$$

여기서, $\sum_{i=1}^{n} w_i x_i = \sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{S_{xx}}\right)x_i = \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i-\bar{x})x_i = \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x}) = 1$

∴ $E(\hat{\beta}_1) = \beta_1$: $\hat{\beta}_1$은 β_1의 불편 추정량

$$E(\hat{\beta}_0) = E(\bar{y} - \hat{\beta}_1\bar{x}) = E(\bar{y}) - \bar{x}E(\hat{\beta}_1) = \beta_0 + \beta_1\bar{x} - \beta_1\bar{x} = \beta_0$$

여기서, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(\beta_0+\beta_1 x_i+\epsilon_i) = \beta_0+\beta_1\bar{x}+\bar{\epsilon}$

$$E(\bar{y}) = E(\beta_0+\beta_1\bar{x}+\bar{\epsilon}) = \beta_0+\beta_1\bar{x}+E(\bar{\epsilon}) = \beta_0+\beta_1\bar{x}$$

∴ $E(\hat{\beta}_0) = \beta_0$: $\hat{\beta}_0$은 β_0의 불편 추정량

2. 분산

$$Var(\hat{\beta}_1) = Var\left(\sum_{i=1}^{n} w_i y_i\right) = \sum_{i=1}^{n} w_i^2 \, Var(y_i) \quad \Leftarrow \epsilon_i{}'s : \text{독립} \rightarrow y_i{}'s : \text{독립}$$

$$= \sum_{i=1}^{n} w_i^2 \, Var(\beta_0 + \beta_1 x_i + \epsilon_i) = \sigma^2 \sum_{i=1}^{n} w_i^2 = \frac{\sigma^2}{S_{xx}}$$

$$Var(\hat{\beta}_0) = Var(\bar{y} - \hat{\beta}_1 \bar{x}) = Var(\bar{y}) + \bar{x}^2 \, Var(\hat{\beta}_1) - 2\,Cov(\bar{y}, \hat{\beta}_1 \bar{x})$$

$$= Var(\bar{y}) + \bar{x}^2 \, Var(\hat{\beta}_1) - 2\bar{x}\,Cov(\bar{y}, \hat{\beta}_1) = \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{S_{xx}} - 2\bar{x}\,Cov(\bar{y}, \hat{\beta}_1)$$

$$= \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{S_{xx}} = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$$

여기서, $Cov(\bar{y}, \hat{\beta}_1) = Cov\left(\frac{1}{n}\sum_{i=1}^{n} y_i, \sum_{i=1}^{n} w_i y_i\right) = Cov\left(\frac{1}{n} 1^t y_i, w^t y\right) = \frac{1}{n} 1^t Cov(y) w$

$$= \frac{1}{n} 1^t \sigma^2 I w = \frac{\sigma^2}{n} 1^t w = 0, \quad \sum_{i=1}^{n} w_i = 0$$

여기서, $Cov(a^t y, b^t y) = a^t Cov(y) b$

여기서, $Cov(y) = \begin{pmatrix} Var(y_1) & Cov(y_1, y_2) \cdots & Cov(y_1, y_n) \\ & Var(y_2) \ \ \cdots & Cov(y_2, y_n) \\ & \ddots \quad \cdots & \vdots \\ & & Var(y_n) \end{pmatrix}$

$$= \begin{pmatrix} Var(y_1) & 0 & \cdots & 0 \\ & Var(y_2) & \cdots & 0 \\ & \ddots & \cdots & \vdots \\ & & & Var(y_n) \end{pmatrix}$$

$E(\hat{\beta}_0) = \beta_0$	$Var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$
$E(\hat{\beta}_1) = \beta_1$	$Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$

3. 잔차

(1) 잔차의 정의

$$\hat{e}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \quad (i = 1, \cdots, n)$$

(2) 잔차의 성질

① $\sum_{i=1}^{n} \hat{e}_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$ ← 1번째 정규방정식

② $\sum_{i=1}^{n} x_i \hat{e}_i = \sum_{i=1}^{n} x_i (y_i - \hat{y}_i) = \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$ ← 2번째 정규방정식

③ $\sum_{i=1}^{n} \hat{y}_i \hat{e}_i = \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i)\hat{e}_i = \hat{\beta}_0 \sum_{i=1}^{n} \hat{e}_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i \hat{e}_i = 0$

④ $\bar{x}$에서의 적합값 $\hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right) = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i = \bar{y}$

→ 적합선 $y = \hat{\beta}_0 + \hat{\beta}_1 x$는 항상 $(\bar{x}, \bar{y})$를 지난다.

4. 가우스-마르코프 정리(Gauss-Markov Theorem)

회귀모형에서 오차항의 기댓값이 0이고 서로 독립일 때, 최소제곱추정량 $\hat{\beta}_0$과 $\hat{\beta}_1$은 y_i들의 선형함수로 주어지는 β_0와 β_1의 불편추정량들 중에서 제일 작은 분산을 갖는다. 즉, $\hat{\beta}_0$과 $\hat{\beta}_1$은 β_0와 β_1의 각각 최량선형불편추정량(Best Linear Unbiased Estimator)이다.

(증명)

$\boxed{Var(\hat{\beta}_1) \le Var(\hat{\beta}_1^*),\ \text{여기서}\ \hat{\beta}_1^*:\ \beta_1\text{의 선형불편추정량}}$

$\hat{\beta}_1^* = \sum_{i=1}^{n} c_i y_i = \sum_{i=1}^{n} (w_i + d_i) y_i$, 여기서, c_i : 임의의 상수, $d_i = c_i - w_i$, $w_i = \frac{x_i - \bar{x}}{S_{xx}}$

$\hat{\beta}_1^*$의 불편성에 의해

$$E(\hat{\beta}_1^*) = E\left[\sum_{i=1}^{n}(w_i + d_i)y_i\right] = E\left[\sum_{i=1}^{n}(w_i + d_i)(\beta_0 + \beta_1 x_i + \epsilon_i)\right]$$

$$= \beta_0 \sum_{i=1}^{n}(w_i + d_i) + \beta_1 \sum_{i=1}^{n}(w_i + d_i)x_i + \sum_{i=1}^{n} E\left[(w_i + d_i)\epsilon_i\right]$$

$$= \beta_0 \sum_{i=1}^{n} d_i + \beta_1 + \beta_1 \sum_{i=1}^{n} d_i x_i = \beta_1 \qquad \text{여기서,}\ \sum_{i=1}^{n} d_i = 0,\ \sum_{i=1}^{n} d_i x_i = 0$$

$$Var(\hat{\beta}_1^*) = Var\left[\sum_{i=1}^{n}(w_i + d_i)y_i\right] = \sum_{i=1}^{n}(w_i + d_i)^2 \sigma^2,\ \text{여기서}\ Var(y_i) = \sigma^2, y_i's : \text{독립}$$

$$= \sum_{i=1}^{n} w_i^2 \sigma^2 + 2\sum_{i=1}^{n} d_i w_i \sigma^2 + \sum_{i=1}^{n} d_i^2 \sigma^2 = Var(\hat{\beta}_1) + \sum_{i=1}^{n} d_i^2 \sigma^2$$

여기서, $\sum_{i=1}^{n} d_i w_i = \sum_{i=1}^{n} d_i \left(\frac{x_i - \bar{x}}{S_{xx}}\right) = \frac{1}{S_{xx}}\left(\sum_{i=1}^{n} d_i x_i - \bar{x}\sum_{i=1}^{n} d_i\right) = 0$, 여기서 $\sum_{i=1}^{n} d_i x_i = 0, \sum_{i=1}^{n} d_i = 0$

$\therefore\ Var(\hat{\beta}_1^*) \ge Var(\hat{\beta}_1)$

2.3.3 오차분산의 추정

$Var(y_i) = Var(\epsilon_i) = \sigma^2$

σ^2의 추정량: $s^2 = \dfrac{\sum_{i=1}^{n} e_i^2}{n-2}$

잔차는 두 개의 제약을 가진다. 즉, $\sum_{i=1}^{n} e_i = 0$, $\sum_{i=1}^{n} x_i e_i = 0$

2.3.4 최우추정법

$y_i = \beta_0 + \beta_1 x_i + e_i \ \ (i = 1, \cdots, n)$

잔차에 대한 가정: $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$

잔차에 대한 추가 가정: $\epsilon_i \sim i.i.d.\, N(0, \sigma^2)$

$\rightarrow \ y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, $\theta = (\beta_0, \beta_1, \sigma^2)$

$$\mathscr{L}(\theta \mid Y_1, Y_2, \cdots, Y_n) = \prod_{i=1}^{n} f(y_i \mid \theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right]$$

$$= (2\pi)^{-n/2}(\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right]$$

$$\ell(\theta) = \log \mathscr{L}(\theta \mid y) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

당분간 σ^2을 상수라고 가정하자.

$$\max \ell(\theta) \ \Leftrightarrow \min \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

$$\Rightarrow \ \hat{\beta}_{0,MLE} = \ \hat{\beta}_{0,LSE} = \overline{y} - \hat{\beta}_1 \overline{x}$$

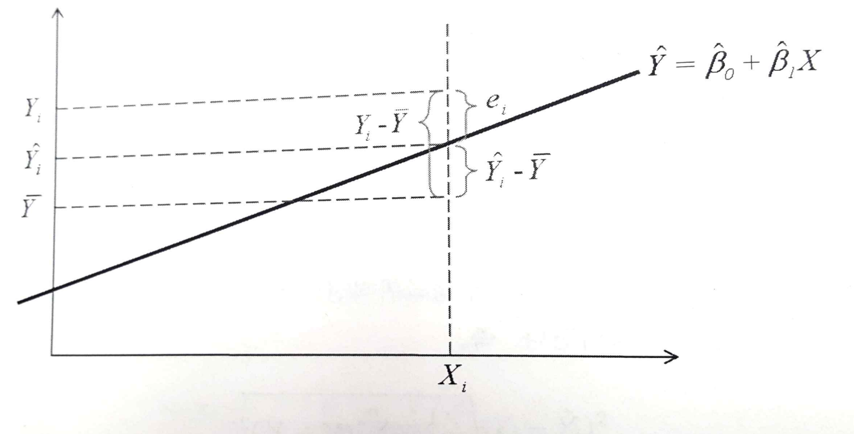
$$\hat{\beta}_{1,MLE} = \ \hat{\beta}_{1,LSE} = \sum_{i=1}^{n} w_i y_i$$

$\hat{\sigma}^2_{MLE}$ 을 구하기 위해

$$\max \ \ \ell(\sigma^2, \hat{\beta}_0, \hat{\beta}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{\partial l(\sigma^2, \hat{\beta}_0, \hat{\beta}_1)}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n} e_i^2 = 0 \ \rightarrow \ \hat{\sigma}^2_{MLE} = \frac{\sum_{i=1}^{n} e_i^2}{n}$$

2.4 회귀직선의 적합도 (Goodness Of Fit: GOF)



(그림 2.6) 세 편차들 간의 관계

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

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

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

SST = SSR + SSE

여기서, $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) = \sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x_i - \hat{\beta}_0 - \hat{\beta}_1 \bar{x})e_i = \sum_{i=1}^{n}\hat{\beta}_1(x_i - \bar{x})e_i$

$$= \hat{\beta}_1 \sum_{i=1}^{n}(x_i - \bar{x})e_i = \hat{\beta}_1 \sum_{i=1}^{n} x_i e_i - \hat{\beta}_1 \bar{x} \sum_{i=1}^{n} e_i = 0$$

- SST: y_i 값들에만 의존
- SSR, SSE: x_i, y_i 값 모두에 의존

적합도 측도: 결정계수 $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$ $\quad 0 < R^2 < 1$

2.5 회귀의 분산분석

세 가지 제곱합을 자유도로 나누면 일종의 분산이 된다. 제곱합의 분할을 이용하여 회귀분석과 관련된 문제를 다루는 것을 “회귀의 분산분석”이라고 한다.

분산분석표 (ANalysis Of VAriance table: ANOVA table)

<표 2.3> 회귀의 분산분석표

요인	제곱합	자유도	평균제곱	F비
회귀	SSR	1	MSR = SSR/1	F_0=MSR/MSE
오차	SSE	$(n-2)$	MSE = SSE/$(n-2)$	
전체	SST	$(n-1)$		