2021.04.17.
김은영

제2장 단순선형회귀모형

2.6 회귀분석에서의 추론

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

$$\epsilon_i \sim i.i.d.\, N(0, \sigma^2)$$

2. 확률변수 $y_i = \beta_0 + \beta_1 x_i + \epsilon_i \sim N(\beta_0 + \beta_1 x_i,\ \sigma^2)$ $(i=1,\ \cdots, n)$

2.6.1 기울기 β_1에 관한 추론

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

여기서, $w_i = \dfrac{x_i-\bar{x}}{S_{xx}}$, $y_i \sim N(\beta_0 + \beta_1 x_i,\ \sigma^2)$

2. $$E(\hat{\beta}_1)=\beta_1, Var(\hat{\beta}_1)=\frac{\sigma^2}{S_{xx}} \ \rightarrow\ \hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right) \ \Rightarrow\ \frac{\hat{\beta}_1-\beta_1}{\sigma/\sqrt{S_{xx}}} \sim N(0,\ 1)$$

여기서, σ^2: 미지의 모수 → $\dfrac{SSE}{\sigma^2} \sim \chi^2(n-2)$를 사용, $\hat{\beta}_1$과 SSE는 서로 독립

3. $$t = \frac{\dfrac{\hat{\beta}_1-\beta_1}{\sigma/\sqrt{S_{xx}}}}{\sqrt{\dfrac{SSE/\sigma^2}{n-2}}} = \frac{\hat{\beta}_1-\beta_1}{s/\sqrt{S_{xx}}} \sim t(n-2) \qquad \text{여기서, } s^2 = \frac{SSE}{n-2}$$

4. $H_0 : \beta_1 = \beta_{10}$의 가설검정

$$\text{검정통계량 } t = \frac{\hat{\beta}_1 - \beta_{10}}{s/\sqrt{S_{xx}}}, \text{ 자유도}=n-2$$

H_1의 형태와 상응하는 기각역

H_1	기각역
$H_1 : \beta_1 > \beta_{10}$	R: $t \geq t_\alpha$
$H_1 : \beta_1 < \beta_{10}$	R: $t \leq -t_\alpha$
$H_1 : \beta_1 \neq \beta_{10}$	R: $\lvert t \rvert \geq t_{\alpha/2}$

2.6.2 절편 β_0에 관한 추론

1. $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$

2. $E(\hat{\beta}_0) = \beta_0$, $Var(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right) \rightarrow \hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\right)$

$$\Rightarrow \frac{\hat{\beta}_0 - \beta_0}{\sigma\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}}} \sim N(0, 1)$$

3. $t = \dfrac{\dfrac{\hat{\beta}_0 - \beta_0}{\sigma\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}}}}{\sqrt{\dfrac{SSE/\sigma^2}{n-2}}} = \dfrac{\hat{\beta}_0 - \beta_0}{s\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}}} \sim t(n-2)$ 여기서, $s^2 = \dfrac{SSE}{n-2}$

4. $H_0 : \beta_0 = \beta_{00}$의 가설검정: 검정통계량 $t = \dfrac{\hat{\beta}_0 - \beta_0}{s\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}}}$, 자유도=$n-2$

H_1의 형태와 상응하는 기각역

H_1	기각역
$H_1 : \beta_0 > \beta_{00}$	R: $t \geq t_\alpha$
$H_1 : \beta_0 < \beta_{00}$	R: $t \leq -t_\alpha$
$H_1 : \beta_0 \neq \beta_{00}$	R: $\lvert t \rvert \geq t_{\alpha/2}$

2.6.3 Y의 평균값에 관한 추론

1. X가 특정한 값 x를 가질 때 반응변수의 평균 $E(Y) = \beta_0 + \beta_1 x$에 대한 추정

(1) $E(Y) = \beta_0 + \beta_1 x$: 추정해야 할 모수

(2) $X = x$에서 $E(Y) = \beta_0 + \beta_1 x$의 점추정치: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

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

$$Var(\hat{y}) = Var(\hat{\beta}_0 + \hat{\beta}_1 x)$$

$= Var(\overline{y} - \hat{\beta}_1 \overline{x} + \hat{\beta}_1 x) = Var\left[\overline{y} + \hat{\beta}_1 (x - \overline{x})\right]$ 여기서, $\hat{\beta}_1$과 $\overline{y}$: 서로 독립

$= Var(\overline{y}) + (x - \overline{x})^2 Var(\hat{\beta}_1)$ 여기서, $(x - \overline{x})$는 상수

$$= \frac{\sigma^2}{n} + \sigma^2 \frac{(x - \overline{x})^2}{S_{xx}} = \sigma^2\left[\frac{1}{n} + \frac{(x - \overline{x})^2}{S_{xx}}\right]$$

$$\rightarrow \quad \frac{\hat{y}-E(y)}{\sigma\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}} \sim N(0,\,1)$$

여기서, σ^2: 미지의 모수

$$\Rightarrow \; t = \frac{\dfrac{\hat{y}-E(y)}{\sigma\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}}}{\sqrt{\dfrac{SSE/\sigma^2}{n-2}}} = \frac{\hat{y}-E(y)}{s\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}} \sim t(n-2)$$

2. X가 특정한 값 x를 가질 때 반응변수가 취하는 새로운 값에 대한 예측

(1) $\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x$, $y-\hat{y}$

(2) $E(y-\hat{y})=0$

$Var(y-\hat{y}) = Var(y) + Var(\hat{y})$ 　　여기서, y와 $\hat{y}$: 서로 독립

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

$$\rightarrow \quad \frac{\hat{y}-y}{\sigma\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}} \sim N(0,\,1)$$

$$\Rightarrow \; t = \frac{\dfrac{\hat{y}-y}{\sigma\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}}}{\sqrt{\dfrac{SSE/\sigma^2}{n-2}}} = \frac{\hat{y}-y}{s\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}} \sim t(n-2)$$

2.7 잔차분석

$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad (i=1,\,\cdots,n)$ 　　여기서, ϵ_i: 오차항, 확률변수, $\epsilon_i \sim i.i.d.\,N(0,\sigma^2)$

$cf)$ e_i: 잔차항, 추정량

< 단순선형회귀모형에 대한 가정 >
1. 선형성: $E(y)=\beta_0+\beta_1 x$
2. 동분산성: $Var(y_i)=Var(\epsilon_i)=\sigma^2 \quad \forall i$
3. 오차항의 정규성: $\epsilon_i \sim i.i.d.\,N(0,\sigma^2)$
4. 독립성: $E(\epsilon_i\epsilon_j)=E(\epsilon_i)E(\epsilon_j) \quad \forall i \neq j$

e_i는 ϵ_i의 실현치로 생각할 수 있기 때문에, 단순선형회귀모형의 가정이 타당한지 확인하는데 잔차를 사용한다.

$e_i = y_i - \hat{y}_i$는 scale-dependent → 표준화 잔차를 정의할 필요가 있다.

$Var(e_i) = s^2(1-h_{ii})$ 여기서, $s^2 = \dfrac{SSE}{n-2}$, $h_{ii} = \dfrac{1}{n} + \dfrac{(x_i - \overline{x})^2}{S_{xx}}$ ← leverage

$$\text{표준화 잔차: } r_i = \frac{e_i}{s\sqrt{1-h_{ii}}}$$

2.8 원점을 지나는 회귀직선

1. 단순선형회귀모형에서 회귀직선이 원점을 지나는 경우 절편의 값이 0이 되므로 기울기의 추정만이 요구된다.

$$y_i = \beta_1 x_i + \epsilon_i \quad (i = 1, \cdots, n) \quad \epsilon_i \sim i.i.d. N(0, \sigma^2)$$

2. β_1에 대한 추정: $Q = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_1 x_i)^2$을 최소화하는 $\hat{\beta}_1$을 추정하자.

$$\frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_1 x_i)x_i = 0 \quad \rightarrow \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

$$E(\hat{\beta}_1) = E\left(\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}\right) = \frac{\sum_{i=1}^{n} x_i E(y_i)}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} x_i(\beta_1 x_i)}{\sum_{i=1}^{n} x_i^2} = \beta_1 \quad \rightarrow \text{ 불편성(unbiasedness)}$$

$$Var(\hat{\beta}_1) = Var\left(\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}\right) = \frac{Var\left(\sum_{i=1}^{n} x_i y_i\right)}{\left(\sum_{i=1}^{n} x_i^2\right)^2} = \frac{\sum_{i=1}^{n} x_i^2 \, Var(y_i)}{\left(\sum_{i=1}^{n} x_i^2\right)^2} = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$$

σ^2에 대한 추정량으로 $s^2 = \dfrac{SSE}{n-1}$을 사용한다.

여기서 자유도: $(n-1)$ ← 하나의 제약이 존재(추정해야 할 모수가 하나)

2.9 상관분석

< x_i들에 대한 가정 >

■ 지금까지의 회귀분석에서는 설명변수를 값이 주어진 상수로 취급했다.

■ 그러나 사회과학 등의 문제에서는 특히 설명변수가 양적변수인 경우 설명변수가 관측되는 모집단은 반응변수의 모집단과 거의 동일하고, 분산도 경우에 따라서는 반응변수의 분산보다 클 수도 있으므로 설명변수가 상수라는 가정은 타당성을 잃게 된다. 설명변수 역시 확률변수로 주어지는 경우에 대한 분석방법을 살펴보자.

2.9.1 상관계수의 추정

x_i와 y_i 모두 확률변수로 가정하면

$$\begin{pmatrix} x_i \\ y_i \end{pmatrix} \sim N_2\left[\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}\right]$$

여기서, $\rho = \dfrac{Cov(x,\ y)}{\sqrt{Var(x)}\ \sqrt{Var(y)}}$: 모상관계수

$$y|x \sim N\left[\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x-\mu_x),\ \sigma_y^2(1-\rho^2)\right]$$

$$E(y|x) = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x-\mu_x) = \mu_y - \rho\frac{\sigma_y}{\sigma_x}\mu_x + \rho\frac{\sigma_y}{\sigma_x}x = \beta_0 + \beta_1 x$$

$$\rightarrow\ \beta_0 = \mu_y - \rho\frac{\sigma_y}{\sigma_x}\mu_x,\ \beta_1 = \rho\frac{\sigma_y}{\sigma_x} \quad \rightarrow\ \sigma_x > 0,\ \sigma_y > 0$$

$\Rightarrow H_0 : \beta_1 = 0 \Leftrightarrow H_0 : \rho = 0$ 검정통계량 $t = \dfrac{\hat{\beta}_1 - 0}{s/\sqrt{S_{xx}}}$

ρ에 대한 추정량으로 γ를 사용한다.

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

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

$$= \frac{S_{xy}}{\sqrt{S_{xx}}\ \sqrt{S_{yy}}}$$

3. γ (ρ의 추정량)와 $\hat{\beta}_1$ (β_1의 추정량) 간의 관계

표본: $\gamma = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{S_{xy}}{S_{xx}} \frac{\sqrt{S_{xx}}}{\sqrt{S_{yy}}} = \hat{\beta}_1 \frac{\sqrt{S_{xx}}}{\sqrt{S_{yy}}}$ $\quad cf)$ 모집단 $\beta_1 = \rho \frac{\sigma_y}{\sigma_x}$

2.9.2 상관계수에 대한 검정

x_i와 y_i 모두 확률변수로 가정하면

$$\begin{pmatrix} x_i \\ y_i \end{pmatrix} \sim N_2 \left[\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix} \right]$$

여기서, $\rho = \frac{Cov(x,\, y)}{\sqrt{Var(x)}\sqrt{Var(y)}}$: 모상관계수

$H_0 : \rho = 0$

검정통계량 $t = \frac{\gamma}{\sqrt{1-\gamma^2} / \sqrt{n-2}} \sim t(n-2)$

[증명]

H_0 하에서 $t = \frac{\hat{\beta}_1}{s / \sqrt{S_{xx}}} \sim t(n-2)$

여기서, $s^2 = \frac{SSE}{n-2} = \frac{(SST - SSR)}{n-2} = \frac{1}{n-2}\left[S_{yy} - \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \right] = \frac{1}{n-2}\left(S_{yy} - \hat{\beta}_1^2 S_{xx} \right)$

여기서, $\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n} \left[(\hat{\beta}_0 + \hat{\beta}_1 x_i) - (\hat{\beta}_0 - \hat{\beta}_1 \bar{x}) \right]^2 = \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 = \hat{\beta}_1^2 S_{xx}$

$$t^2 = \frac{\hat{\beta}_1^2}{s^2 / S_{xx}} = \frac{\hat{\beta}_1^2}{\dfrac{S_{yy} - \hat{\beta}_1^2 S_{xx}}{(n-2) S_{xx}}} = \frac{(n-2)\hat{\beta}_1^2}{\left(\dfrac{S_{yy}}{S_{xx}} - \hat{\beta}_1^2 \right)} = \frac{(n-2)\hat{\beta}_1^2 \dfrac{S_{xx}}{S_{yy}}}{1 - \hat{\beta}_1^2 \dfrac{S_{xx}}{S_{yy}}}$$

$$= \frac{(n-2)\gamma^2}{1-\gamma^2} = \left(\frac{\gamma}{\sqrt{1-\gamma^2} \,/\, \sqrt{n-2}} \right)^2$$

$\rightarrow$ $t = \frac{\gamma}{\sqrt{1-\gamma^2} / \sqrt{n-2}} \sim t(n-2)$