회귀분석

단순선형회귀모형에서의 추론

- 단순선형회귀모형

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad where \quad \varepsilon_i \sim i.i.d.N(0, \sigma^2)$$

- 추정회귀직선

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x$$

ex. $\hat{y} = -1.07 + 2.74x$

- 기울기 $\widehat{\beta_1} = 2.74$의 의미: 복용량을 1단위 늘릴 때 그 효과의 평균지속기간이 2.74일 증가한다는 뜻

- 주의: 위의 추정직선은 주어진 자료로부터 얻은 것이므로 주어진 x값의 구간에서만 그 관계가 유효.

단순선형회귀모형에서의 추론

- 예제. 다음 표는 어떤 알레르기 증세에 효과가 있다고 하는 새로 개발된 약품의 복용량(mg)과 효과가 지속되는 기간(일)을 기록한 자료이다.

복용량(x)	3	3	4	5	6	6	7	8	8	9
지속기간(y)	9	5	12	9	14	16	22	18	24	22

단순선형회귀모형에서의 추론

- 질문

(1) 기울기에 관하여 추정치가 2.74인데 실제로는 4가 아닐까? 기울기가 0이면 x는 y에 전혀 영향을 주지 못하는데 위의 예에서 기울기를 0으로 볼 수는 없을까?

(2) $x^* = 4.5$일 때 $y = 11.26$인데 이 값의 오차는 어느 정도일까?

기울기 β_1에 대한 추론

- $\widehat{\beta_1} = \frac{S_{xy}}{S_{xx}}$ where $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$ and $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

- $E[\widehat{\beta_1}] = \beta_1,\qquad Var[\widehat{\beta_1}] = \frac{\sigma^2}{S_{xx}}$

- $S.E.(\widehat{\beta_1}) = \frac{\sigma}{\sqrt{S_{xx}}} \qquad \Rightarrow \qquad \widehat{S.E.(\widehat{\beta_1})} = \frac{s}{\sqrt{S_{xx}}}$ where $s = \sqrt{\frac{SSE}{n-2}} = \sqrt{MSE}$

- $t = \frac{(\widehat{\beta_1} - \beta_1)}{s/\sqrt{S_{xx}}} \sim t(n-2)$

- $Z = \frac{(\widehat{\beta_1} - \beta_1)}{\sigma/\sqrt{S_{xx}}} \sim N(0,1)$

기울기 β_1에 대한 추론

- β_1의 $100(1-\alpha)\%$ 신뢰구간: $\widehat{\beta_1} \pm t_{\frac{\alpha}{2}}(n-2) \times \frac{s}{\sqrt{S_{xx}}}$

- 가설 $H_0: \beta_1 = \beta_{10}$에 대한 검정(유의수준 α):

  검정통계량: $t = \frac{(\widehat{\beta_1} - \beta_{10})}{s/\sqrt{S_{xx}}}$

  이 t 통계량은 H_0가 맞을 때 자유도가 $n-2$인 t 분포를 따른다.
  각 대립가설에 대한 기각역:

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

예제

- 예제. 다음 표는 어떤 알레르기 증세에 효과가 있다고 하는 새로 개발된 약품의 복용량(mg)과 효과가 지속되는 기간(일)을 기록한 자료이다.

복용량(x)	3	3	4	5	6	6	7	8	8	9
지속기간(y)	9	5	12	9	14	16	22	18	24	22

절편 β_0에 대한 추론

- $\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x} \quad where \quad \widehat{\beta_1} = \frac{S_{xy}}{S_{xx}}$

- $E\left[\widehat{\beta_0}\right] = \beta_0, \qquad Var\left[\widehat{\beta_0}\right] = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$

- $S.E.\left(\widehat{\beta_0}\right) = \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}} \quad \Rightarrow \quad \widehat{S.E.\left(\widehat{\beta_1}\right)} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$

  $where\ s = \sqrt{\frac{SSE}{n-2}} = \sqrt{MSE}$

- $t = \frac{(\widehat{\beta_0} - \beta_0)}{s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \sim t(n-2)$

- $Z = \frac{(\widehat{\beta_1} - \beta_1)}{\sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \sim N(0,1)$

절편 β_0에 대한 추론

- β_0의 $100(1-\alpha)\%$ 신뢰구간: $\widehat{\beta_0} \pm t_{\frac{\alpha}{2}}(n-2) \times s\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}$

- 가설 $H_0: \beta_0 = \beta_{00}$에 대한 검정(유의수준 α):

검정통계량: $t = \frac{(\widehat{\beta_0}-\beta_{00})}{s\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}}$

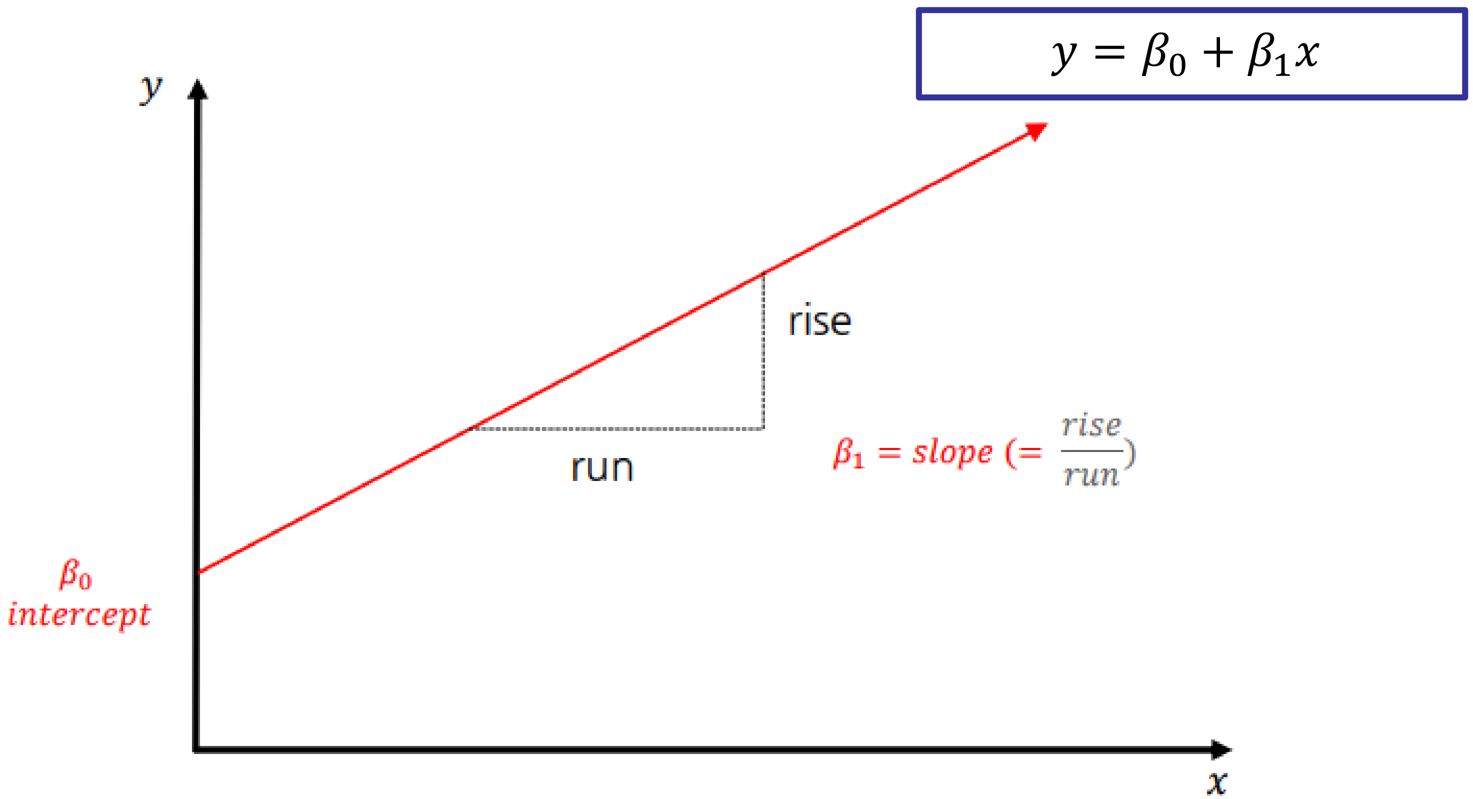
이 t 통계량은 H_0가 맞을 때 자유도가 $n-2$인 t 분포를 따른다.
각 대립가설에 대한 기각역:

$H_1: \beta_0 > \beta_{00}$ $\quad$ $R: t \geq t_\alpha(n-2)$
$H_1: \beta_0 < \beta_{00}$ $\quad$ $R: t \leq -t_\alpha(n-2)$
$H_1: \beta_0 \neq \beta_{00}$ $\quad$ $R: |t| \geq t_{\alpha/2}(n-2)$

평균반응 vs. 반응변수값

- 평균반응: $\beta_0 + \beta_1 x^*$

- 반응변수값 Y: $\beta_0 + \beta_1 x^* + \varepsilon$



평균반응 $\beta_0 + \beta_1 x^*$에 대한 추론

- 추정량: $\widehat{\beta_0 + \beta_1 x^*} = \widehat{\beta_0} + \widehat{\beta_1} x^*$

- $E\left[\widehat{\beta_0} + \widehat{\beta_1} x^*\right] = \beta_0 + \beta_1 x^*,$ $\quad Var\left[\widehat{\beta_0} + \widehat{\beta_1} x^*\right] = \sigma^2\left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right)$

- $S.E.\left(\widehat{\beta_0} + \widehat{\beta_1} x^*\right) = \sigma\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$ $\quad S.E.\left(\widehat{\beta_0 + \beta_1 x^*}\right) = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$

  $$where\ s = \sqrt{\frac{SSE}{n-2}} = \sqrt{MSE}$$

- $t = \dfrac{(\widehat{\beta_0} + \widehat{\beta_1} x^*) - (\beta_0 + \beta_1 x^*)}{s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}} \sim t(n-2)$

평균반응 $\beta_0 + \beta_1 x^*$에 대한 추론

- $\beta_0 + \beta_1 x^*$의 $100(1-\alpha)\%$ 신뢰구간: $(\widehat{\beta_0} + \widehat{\beta_1} x^*) \pm t_{\frac{\alpha}{2}}(n-2) \times s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$

- 가설 $H_0: \beta_0 + \beta_1 x^* = \mu_0$에 대한 검정(유의수준 α):

검정통계량: $t = \frac{(\widehat{\beta_0} + \widehat{\beta_1} x^*) - \mu_0}{s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}}$

이 t 통계량은 H_0가 맞을 때 자유도가 $n-2$인 t 분포를 따른다.
각 대립가설에 대한 기각역:

$H_1: (\beta_0 + \beta_1 x^*) > \mu_0$	$R: t \geq t_\alpha(n-2)$
$H_1: (\beta_0 + \beta_1 x^*) < \mu_0$	$R: t \leq -t_\alpha(n-2)$
$H_1: (\beta_0 + \beta_1 x^*) \neq \mu_0$	$R: \lvert t \rvert \geq t_{\alpha/2}(n-2)$

반응변수값 $\mathrm{Y} = \beta_0 + \beta_1 x^* + \varepsilon$에 대한 추론

- $x = x^*$에서의 반응변수값 Y 예측값의 추정된 표준오차

$$s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

수고하셨습니다.

- 과제 X