Inference : Simple Linear Regression

회귀직선의 유의성 검정

- Model : $y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \;\; i = 1, 2, \ldots, n, \;\; \epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$

- 회귀직선의 유의성 검정 (F-test)
  - 가설 : $H_0 : \beta_1 = 0 \; vs. \; H_1 : \beta_1 \neq 0$
  - 검정통계량 : $F = \dfrac{MSR}{MSE} = \dfrac{SSR/1}{SSE/(n-2)} \sim_{H_0} F(1, n-2)$
  - 검정통계량의 관측값 : F_0
  - 유의수준 α에서의 기각역 : $F_0 \geq F_\alpha(1, n-2)$
  - 유의확률 $= P(F \geq F_0)$

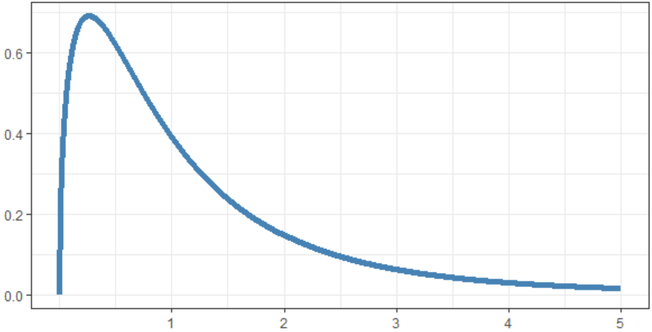
회귀직선의 유의성 검정

■ 회귀직선의 유의성 검정을 위한 분산분석표

요인	제곱합(SS)	자유도(df)	평균제곱(MS)	F_0	유의확률
회귀	SSR	1	$MSR = \frac{SSR}{1}$	$\frac{MSR}{MSE}$	$P(F \geq F_0)$
잔차	SSE	$n-2$	$MSE = \frac{SSE}{n-2}$		
계	SST	$n-1$			

Example

Figure: F분포의 확률밀도함수 그림 : $F(3, 8)$



Example

광고료과 총판매액

- 회귀직선의 유의성 검정 : $H_0 : \beta_1 = 0 \ vs. \ H_1 : \beta_1 \neq 0$

요인	제곱합	자유도	평균제곱	f	유의확률
회귀	313.043	1	313.043	45.24	0.0001487
잔차	55.357	8	$6.92 (= \hat{\sigma}^2)$		
계	368.4	9			

- $F_{0.05}(1, 8) = 5.3177$

회귀계수에 대한 추론

- 모회귀계수(기울기) β_1 에 대한 추론
  - β_1 의 최소제곱추정량 : $\hat{\beta}_1 = \dfrac{S_{(xy)}}{S_{(xx)}}$
  - $\mathrm{E}(\hat{\beta}_1) = \beta_1,\ \mathrm{Var}(\hat{\beta}) = \dfrac{\sigma^2}{S_{(xx)}}$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{(xx)}}\right)$$

회귀계수에 대한 추론

모회귀계수(기울기) β_1 에 대한 추론

- $\widehat{\mathrm{Var}}(\hat{\beta}_1) = \dfrac{MSE}{S_{(xx)}}$, $\hat{\sigma}_{\hat{\beta}_1} = \sqrt{\dfrac{MSE}{S_{(xx)}}}$
- studentized $\hat{\beta}_1$ 의 분포 :

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{S_{(xx)}}} \sim t(n-2), \quad \hat{\sigma} = \sqrt{MSE}$$

- $\hat{\beta}_1$ 의 $100(1-\alpha)\%$ 신뢰구간

$$\hat{\beta}_1 \pm t_{\alpha/2}(n-2)\frac{\hat{\sigma}}{\sqrt{S_{(xx)}}}$$

회귀계수에 대한 추론

- 모회귀계수(기울기) β_1에 대한 추론
  - 가설검정 : $H_0 : \beta_1 = \beta_1^0$
  - 검정통계량 : $T = \dfrac{\hat{\beta}_1 - \beta_1^0}{\hat{\sigma}/\sqrt{S_{(xx)}}} \sim_{H_0} t(n-2)$, 관측값 : t

대립가설	유의확률	유의수준 α 기각역
$H_1 : \beta_1 > \beta_1^0$	$P(T \geq t)$	$t \geq t_\alpha(n-2)$
$H_1 : \beta_1 < \beta_1^0$	$P(T \leq t)$	$t \geq t_\alpha(n-2)$
$H_1 : \beta_1 \neq \beta_1^0$	$P(\lvert T \rvert \geq \lvert t \rvert)$	$\lvert t \rvert \geq t_{\alpha/2}(n-2)$

회귀계수에 대한 추론

- 모회귀계수(절편) β_0에 대한 추론
  - β_0의 최소제곱추정량 : $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
  - $E(\hat{\beta}_0) = \beta_0, \;\; \mathrm{Var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{(xx)}} \right)$

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{(xx)}} \right) \right)$$

회귀계수에 대한 추론

- 모회귀계수(절편) β_0 에 대한 추론
  - studentized $\hat{\beta}_0$ 의 분포 :

$$\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} \sim t(n-2), \quad \hat{\sigma}_{\hat{\beta}_0} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{(xx)}}}$$

  - $\hat{\beta}_0$ 의 $100(1-\alpha)\%$ 신뢰구간

$$\hat{\beta}_0 \pm t_{\alpha/2}(n-2)\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{(xx)}}}$$

회귀계수에 대한 추론

모회귀계수(절편) β_0에 대한 추론

- 가설검정 : $H_0 : \beta_0 = \beta_0^0$
- 검정통계량 : $T = \dfrac{\hat{\beta}_0 - \beta_0^0}{\hat{\sigma}_{\hat{\beta}_0}} \sim_{H_0} t(n-2)$, 관측값 : t

대립가설	유의확률	유의수준 α 기각역
$H_1 : \beta_0 > \beta_0^0$	$P(T \geq t)$	$t \geq t_\alpha(n-2)$
$H_1 : \beta_0 < \beta_0^0$	$P(T \leq t)$	$t \geq t_\alpha(n-2)$
$H_1 : \beta_0 \neq \beta_0^0$	$P(\lvert T\rvert \geq \lvert t\rvert)$	$\lvert t\rvert \geq t_{\alpha/2}(n-2)$

Example

- 광고료와 판매총액
  - $\hat{\beta}_0$ 와 $\hat{\beta}_1$ 의 95% 신뢰구간 $(t_{0.05/2}(8) = 2.306)$

$$\begin{aligned}
\hat{\beta}_0 \pm t_{\alpha/2}\hat{\sigma}_{\hat{\beta}_0} &= -2.270 \pm 2.306 \times \sqrt{6.92} \times \sqrt{\frac{1}{10} + \frac{8^2}{46}} \\
&= -2.270 \pm 2.306 \times 3.21 = (-9.672, 5.132) \\
\hat{\beta}_1 \pm t_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{S_{(xx)}}} &= 2.609 \pm 2.306 \times \frac{\sqrt{6.92}}{\sqrt{46}} \\
&= 2.61 \pm 2.306 \times 0.388 = (1.714, 3.504)
\end{aligned}$$

Example

광고료와 판매총액

- $H_0 : \beta_0 = 0 \ vs. \ H_1 : \beta_0 \neq 0$ 에 대한 가설검정 $(\alpha = 0.05)$
  - 검정통계량 관측값 : $t = \dfrac{\hat{\beta}_0 - 0}{\hat{\sigma}_{\hat{\beta}_0}} = \dfrac{-2.270}{3.21} = -0.707$
  - 기각역 : $|t| \geq t_{0.05/2}(8) = 2.306$
  - 결과 : 기각 못함, 유의확률 $= 0.5$
- $H_1 : \beta_1 = 0 \ vs. \ H_1 : \beta_1 \neq 0$ 에 대한 가설검정 $(\alpha = 0.05)$
  - 검정통계량 관측값 : $t = \dfrac{\hat{\beta}_1 - 0}{\hat{\sigma}_{\hat{\beta}_1}} = \dfrac{2.61}{0.388} = 6.72$
  - 기각역 : $|t| \geq t_{0.05/2}(8) = 2.306$
  - 결과 : 기각!, 유의확률 $= < 0.001$

평균반응예측

- $x = x_0$가 주어졌을 때 평균반응의 예측
  - 평균반응 (mean response) : $\mu_0 = E(Y|x_0) = \beta_0 + \beta_1 x_0$
  - 평균반응 추정량 : $\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$
  - $\mathrm{E}(\hat{\mu}_0) = \mu_0, \;\; \mathrm{Var}(\hat{\mu}_0) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{(xx)}} \right)$

$$\hat{\mu}_0 \sim N\left(\mu_0, \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{(xx)}} \right) \sigma^2 \right)$$

평균반응예측

- $x = x_0$가 주어졌을 때 평균반응의 예측
  - sutentized $\hat{\mu}_0$ 의 분포

$$\frac{\hat{\mu}_0 - \mu_0}{\hat{\sigma}_{\hat{\mu}_0}} \sim t(n-2), \quad \hat{\sigma}_{\hat{\mu}_0} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{(xx)}}}$$

  - $\hat{\mu}_0$의 $100(1-\alpha)\%$ 신뢰구간

$$\hat{\mu}_0 \pm t_{\alpha/2}(n-2)\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{(xx)}}}$$

  - 신뢰대 (confidence band)

Example

광고료와 판매총액

- $\hat{\mu}_0$ 의 $100(1-\alpha)\%$ 신뢰구간

$$\hat{\mu}_0 \pm t_{\alpha/2}(n-2)\hat{\sigma}_{\hat{\mu}_0}$$

- $x_0 = 4$ 인 경우

$$\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = -2.270 + (2.609)(4) = 8.17$$

$$\widehat{\mathrm{Var}}(\hat{\mu}_0) = \hat{\sigma}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{(xx)}}\right) = (6.92)\left(\frac{1}{10} + \frac{(4-8)^2}{46}\right) = 3.10$$

$$\hat{\sigma}_{\hat{\mu}_0} = \sqrt{3.10} = 1.76$$

$$\Rightarrow \hat{\mu}_0 \pm t_{\alpha/2}(n-2)\hat{\sigma}_{\hat{\mu}_0} = 8.17 \pm (2.306)(1.76) = (4.11, 12.23)$$

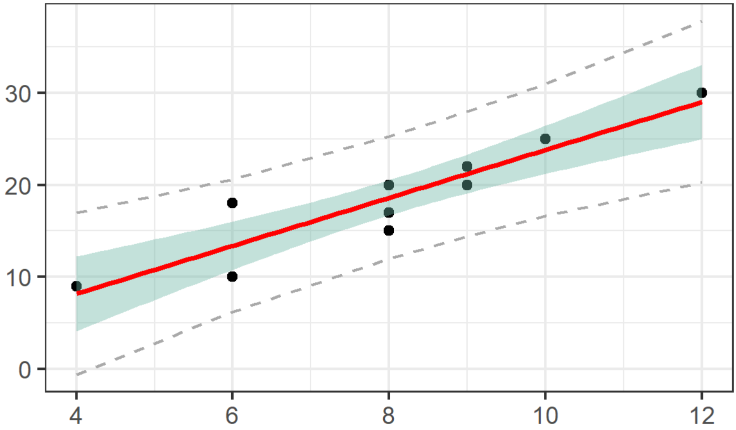
Example

- 광고료와 판매총액

$$
\begin{aligned}
x = 6: &\quad 13.38 \pm (2.306)(1.14) = 13.38 \pm 2.63 = (10.75, 16.01) \\
x = 8: &\quad 18.60 \pm (2.306)(0.83) = 18.60 \pm 1.94 = (16.69, 20.51) \\
x = 9: &\quad 21.21 \pm (2.306)(0.92) = 21.21 \pm 2.12 = (19.09, 23.33) \\
x = 10: &\quad 23.82 \pm (2.306)(1.14) = 23.82 \pm 2.63 = (21.19, 26.45) \\
x = 12: &\quad 29.04 \pm (2.306)(1.76) = 29.04 \pm 4.59 = (24.45, 33, 63)
\end{aligned}
$$

Example

Figure: Confidence Band (신뢰대)



개별적인 y값 예측

- $x = x_0$가 주어졌을 때 $y = y_0$ 예측
  - $y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0$
  - 예측값 : $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$
  - $\mathrm{E}(\hat{y}_0) = \mu_0, \;\; \mathrm{Var}(\hat{y}_0) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{(xx)}}\right)$

$$\hat{y}_0 \sim N\left(\mu_0, \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{(xx)}}\right)\sigma^2\right)$$

개별적인 y값 예측

- $x = x_0$가 주어졌을 때 $y = y_0$ 예측
  - studentized $\hat{y}_0$ 의 분포 :

$$\frac{\hat{y}_0 - y_0}{\hat{\sigma}_{\hat{y}_0}} \sim t(n-2), \quad \hat{\sigma}_{\hat{y}_0} = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{(xx)}}}$$

  - $\hat{y}_0$의 $100(1-\alpha)\%$ 신뢰구간

$$\hat{y}_0 \pm t_{\alpha/2}(n-2)\hat{\sigma}_{\hat{y}_0}$$

모형의 타당성

기본 가정

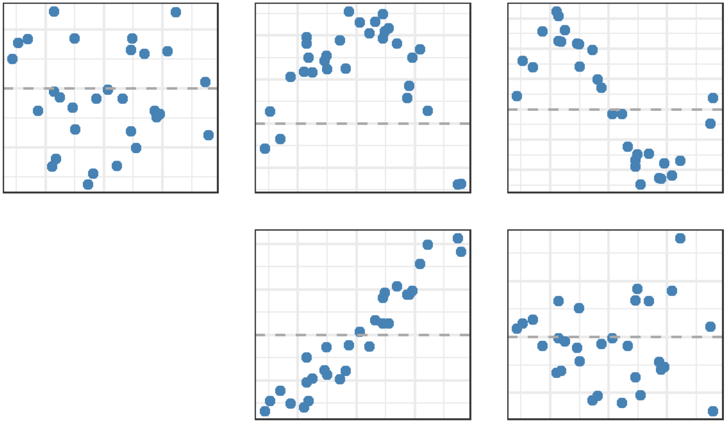
- Linearity (선형성) : $E(Y|X = x) = \mu_{y \cdot x} = \beta_0 + \beta_1 x$
- Homoscedastic (등분산성) : $Var(Y|X = x) = \sigma^2$
- Normality (정규성) : $Y|X = x \sim N(E(Y|X = x), \sigma^2)$
- Independency (독립성) : ϵ's are mutually independent

잔차의 검토

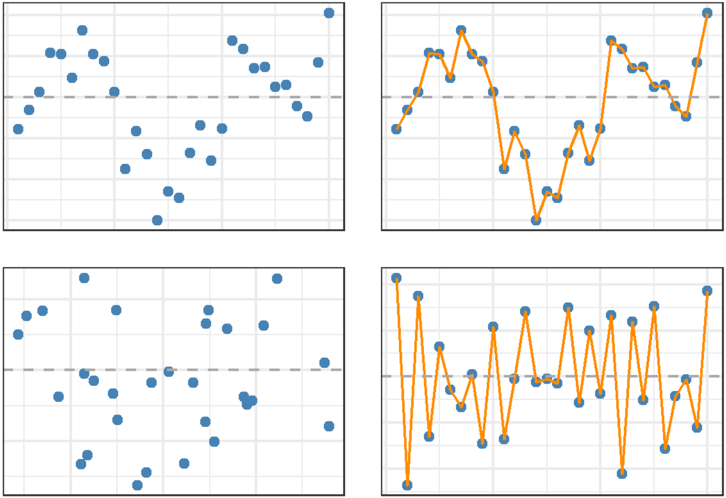
- 잔차(residual) : $e_i = y_i - \hat{y}_i$
- 잔차를 통한 모형의 가정 검토
- 잔차의 산점도 : (x_i, e_i) 또는 $(\hat{y}_i, e_i)$

$$\because \left(\sum_i x_i e_i = 0, \sum_i e_i = 0\right), \left(\sum_i \hat{y}_i e_i = 0, \sum_i e_i = 0\right)$$

잔차의 산점도



오차의 자기 상관



오차의 자기 상관

Durbin-Watson Test

- hypothesis

  H_0 : 오차항들을 독립이다 vs. H_1 : 오차항들은 독립이 아니다.

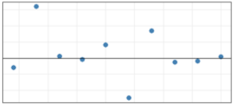
- 검정통계량

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

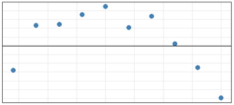
- 검정 $(d_L = d_L(n, p, \alpha), d_U = d_U(n, p, \alpha))$
  - ▹ d or $4 - d < d_L \;\Rightarrow\; H_0$ 기각
  - ▹ d or $4 - d > d_U \;\Rightarrow\; H_0$ 기각못함
  - ▹ $d_L < d$ or $4 - d < d_L \;\Rightarrow\;$ 결정 보류

오차의 자기 상관

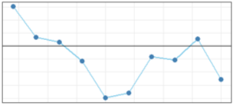
Durbin-Watson Test



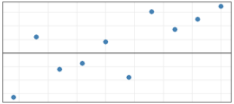
D = 3.013, p-value = 0.921



D = 0.734, p-value = 0.001

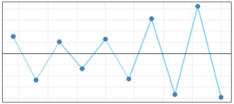


D = 1.066, p-value = 0.015



D = 3.013, p-value = 0.921

D = 2.661 p-value = 0.767



D = 3.507, p-value <0.005