8장. 상관분석과 회귀분석

박명현

서울대학교 학부대학

기울기 β_1에 대한 추론

- 모회귀계수 β_1의 최소제곱추정량 $\hat{\beta}_1$를 다음과 같이 표현할 수 있다.

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- $\hat{\beta}_1$의 기대값과 분산

$$E(\hat{\beta}_1) = \frac{1}{S_{xx}} \sum (x_i - \bar{x}) \cdot E(Y_i|x_i) = \frac{1}{S_{xx}} \sum (x_i - \bar{x})(\beta_0 + \beta_1 x_i)$$
$$= \frac{\beta_0}{S_{xx}} \sum (x_i - \bar{x}) + \frac{\beta_1}{S_{xx}} \sum (x_i - \bar{x}) x_i = \beta_1$$

$$Var(\hat{\beta}_1) = \frac{1}{S_{xx}^2} \sum (x_i - \bar{x})^2 \cdot Var(Y_i|x_i) = \frac{\sigma^2}{S_{xx}^2} \sum (x_i - \bar{x})^2 = \frac{\sigma^2}{S_{xx}}$$

기울기 β_1에 대한 추론

- $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{S_{xx}} \cdot y_i$ 이므로
  독립적으로 정규분포를 따르는 y_i $(i = 1, \cdots, n)$들의 선형결합인
  $\hat{\beta}_1$ 또한 다음의 정규분포를 따르게 된다. (σ^2 unknown)

$$\hat{\beta}_1 \;\sim\; N(\beta_1, \frac{\sigma^2}{S_{xx}}) \tag{1}$$

- Sampling distribution of $\hat{\beta}_1$

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

- $\hat{\beta}_1$의 추정된 분산과 표준오차:

$$s^2(\hat{\beta}_1) = \widehat{Var}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{S_{xx}}, \qquad s(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

β_1에 대한 추론

- β_1에 대한 $100(1-\alpha)\%$ 신뢰구간 (σ^2 unknown)

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

- β_1에 대한 유의성 t 검정
  - 귀무가설: $H_0 : \beta_1 = b$
  - Under H_0, 검정통계량:

$$T = \frac{\hat{\beta}_1 - b}{s(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - b}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} \sim t(n-2)$$

여기서 검정통계량의 관측값을 t라고 하면 다음이 성립한다.

	유의확률 P	유의수준 α의 기각역
(a) $H_1 : \beta_1 > b$	$P = P(T \geq t)$	$t \geq t_\alpha(n-2)$
(b) $H_1 : \beta_1 < b$	$P = P(T \leq t)$	$t \leq -t_\alpha(n-2)$
(c) $H_1 : \beta_1 \neq b$	$P = P(\|T\| \geq \|t\|)$	$\|t\| \geq t_{\alpha/2}(n-2)$

β_1에 관한 추론 예제

Example

예 8.2에서 단순선형회귀모형을 적용할 때, 다음의 추론을 하여라.
(a) 회귀직선의 유의성을 t검정으로 유의수준 1%에서 검정하고, 분산분석표에 의한 F검정과의 관계를 설명하여라.

sol) 검정하고자 하는 가설은 다음과 같다.

$$H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0$$

- 예제 8.2의 풀이 과정에서

$$n = 12, \quad \hat{\beta}_1 = -0.859, \quad S_{xx} = 24.649, \quad \hat{\sigma}^2 = MSE = 0.0289$$

이므로 t 검정통계량의 관측값은

$$t = \frac{\hat{\beta}_1 - 0}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} = \frac{-0.859}{\sqrt{0.0289/24.649}} = 25.087$$

- t분포표에서 $t_{0.005}(10) = 3.169$이므로, 다음과 같이 기각역에 속한다.

$$|25.087| = |t| \geq t_{0.005}(10) = 3.169$$

즉, 유의수준 1%에서 회귀직선이 유의하다고 할 수 있다.

- 이러한 회귀직선의 유의성 t검정은 분산분석표를 이용한 F검정으로도 할 수 있으며,

$$F = \frac{SSR/1}{SSE/(n-2)} = \frac{\hat{\beta}_1^2 S_{xx}}{MSE} = \left(\frac{\hat{\beta}_1 - 0}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} \right)^2 = T^2$$

단순선형회귀분석에서는 F검정과 t검정이 다음의 관계에 있음을 알 수 있다.

$$t_{n-2}^2 = F(1, n-2)$$

유의성 F검정과 T검정비교

- F검정은 회귀모형의 모든 기울기 모수의 유의성에 대한 총제적 검정(overall test)을 수행한다.

  - 일반적으로 절편 β_0는 F검정에서 제외한다.

- 단순회귀모형에서는 β_0를 제외한 회귀계수가 β_1 기울기 하나이므로 $H_0 : \beta_1 = 0$를 검정하는 것이 되어, 개별 검정인 t검정과 총체적 검정인 F 검정의 귀무가설과 대립가설이 동일한 형태가 되며 검정의 결과도 완벽하게 일치한다. 그 관계는 정확히 $t^2 = F$이다.

- 다중회귀에서는 기울기가 여러 개 이므로 t검정과 F 검정이 같지 않다.

Example

(b) 자동차의 가격이 사용년수에 따라 연평균 감소액이 80만원을 초과하는지 유의수준 5%에서 검정하여라.

sol) 자료의 측정 단위가 백만원이므로, 검정하고자 하는 가설은 다음과 같다.

$$H_0 : \beta_1 = -0.8 \quad H_1 : \beta_1 < -0.8$$

- t 검정통계량의 관측값은

$$t = \frac{\hat{\beta}_1 - (-0.8)}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} = \frac{-0.859 + 0.8}{\sqrt{0.0289/24.649}} = -1.723$$

- t분포표에서 $t_{0.05}(10) = 1.812$이므로, 다음의 기각역에 속하지 않는다.

$$t \leq -t_{0.05}(10)$$

- 따라서, 유의수준 5%에서 H_0을 기각할 수 없고, 중고가격의 연평균 감소액이 80만원을 초과한다는 증거는 없다고 할 수 있다.

Example

(c) 모회귀계수 β_1의 95% 신뢰구간을 구하고 그 의미를 해석하여라.

sol) t분포표에서 $t_{0.025}(10) = 2.228$이므로, β_1의 95%신뢰구간은

$$\hat{\beta}_1 \pm t_{0.025}(10)\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = -0.859 \pm (2.228)\sqrt{0.0289/24.649}$$
$$= -0.859 \pm 0.076$$

- 즉, 95% 신뢰구간은

$$-0.935 \leq \beta_1 \leq -0.783$$

이고, 중고차 가격은 연평균 783,000원에서 935,000원 사이로 감소한다고 신뢰수준 95%에서 결론지을 수 있다.

절편 β_0에 관한 추론

- 모회귀계수 β_0의 최소제곱추정량 $\hat{\beta}_0$은 다음과 같이 표현 가능

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \sum_{i=1}^{n} \frac{y_i}{n} - \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \bar{x}$$

- $\hat{\beta}_0$의 기대값과 분산:

$$\begin{aligned} E(\hat{\beta}_0) &= E(\bar{Y} - \hat{\beta}_1 \bar{x}) = E(\frac{1}{n}\sum Y_i) - E(\hat{\beta}_1 \bar{x}) \\ &= \frac{1}{n}\sum E(Y_i|x_i) - E(\hat{\beta}_1)\bar{x} = \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} = \beta_0 \end{aligned}$$

$$\begin{aligned} Var(\hat{\beta}_0) &= \sum \left\{ \frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}} \right\}^2 Var(Y_i|x_i) \\ &= \sigma^2 \left\{ \sum \frac{1}{n^2} - 2\sum \frac{(x_i - \bar{x})}{nS_{xx}} \bar{x} + \sum \frac{(x_i - \bar{x})^2 \bar{x}^2}{S_{xx}^2} \right\} \\ &= \sigma^2 \left\{ \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right\} \end{aligned}$$

절편 β_0에 관한 추론

- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = \sum \frac{y_i}{n} - \frac{\sum(x_i-\bar{x})\cdot y_i}{\sum(x_i-\bar{x})^2}\bar{x} = \sum\left\{\frac{1}{n} - \frac{(x_i-\bar{x})\bar{x}}{S_{xx}}\right\} y_i$ 이므로 $\hat{\beta}_0$ 또한 다음의 정규분포를 따르게 된다.

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

- σ^2을 $\hat{\sigma}^2$으로 대체한 표본분포는 다음과 같다.

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

β_0에 대한 추론

- β_0에 대한 $100(1-\alpha)\%$ 신뢰구간 (σ^2 unknown)

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

- β_0에 대한 유의성 t 검정

- 귀무가설: $H_0 : \beta_0 = a$

- Under H_0, 검정통계량:

$$T = \frac{\hat{\beta}_0 - a}{s(\hat{\beta}_0)} = \frac{\hat{\beta}_0 - a}{\hat{\sigma}\sqrt{n^{-1}+\bar{x}^2/S_{xx}}}$$

이고, 이의 관측값을 t라고 하면 다음이 성립한다.

	유의확률 P	유의수준 α의 기각역
(a) $H_1 : \beta_0 > a$	$P = P(T \geq t)$	$t \geq t_\alpha(n-2)$
(b) $H_1 : \beta_0 < a$	$P = P(T \leq t)$	$t \leq -t_\alpha(n-2)$
(c) $H_1 : \beta_0 \neq a$	$P = P(\|T\| \geq \|t\|)$	$\|t\| \geq t_{\alpha/2}(n-2)$

절편 β_0에 관한 추론의 주의점

- 실제 응용에서는 특수한 상황을 제외하고는 절편 β_0에 대하여 가설검정이나 신뢰구간 추정을 하지 않는다.

  예) 키를 설명변수, 몸무게를 종속변수로 한 회귀식
  - 키가 0일 때의 몸무게 추정값인 β_0의 추정은 실질적인 의미가 없음
  - 관측된 설명변수의 범위를 많이 벗어나게 되면 모형의 타당성이 담보되지 못함

- 그런데 특별한 경우가 아니면 β_0이 통계적으로 비유의적이라 하더라도 회귀모형에 포함시켜 분석을한다.
  - β_0를 회귀모형에서 제외하면 추정된 회귀식이 자료를 잘 설명하지 못함

β_0에 관한 추론 예제

Example

예 8.2에서 단순선형회귀모형을 적용할 때, 모회귀직선의 절편 β_0에 관한 95% 신뢰구간을 구하여라.

sol) 예제 8.2-3의 풀이 과정에서

$$n = 12, \quad \bar{x} = 3.308, \quad S_{xx} = 24.649, \quad \hat{\beta}_0 = 5.133, \quad \hat{\sigma}^2 = 0.0289$$

이고 $t_{0.025}(10) = 2.228$이므로, β_0의 95% 신뢰구간은

$$\hat{\beta}_0 \pm t_{0.025}(10)\ \hat{\sigma}\sqrt{n^{-1} + \bar{x}^2/S_{xx}} = 5.133 \pm 2.526$$

회귀모형을 이용한 예측(prediction)

- 추정된 회귀식을 이용하여 특정한 x_0 값에 대응하는 반응변수 Y의 평균값을 추정 또는 예측해야 하는 경우가 있다.

  - x_0를 우리가 관심있는 설명변수의 특정한 값이라고 하자.

  - 회귀모형에서 $x = x_0$일 때의 예측 가능한 Y의 값은 $\beta_0 + \beta_1 x_0$이고,

    이 값은 $x = x_0$일 때 반응변수 Y값들의 평균값이다.

  - $E(Y_0|x_0)$를 $x = x_0$일 때의 평균반응값이라 하자.

- $E(Y_0|x_0) = \beta_0 + \beta_1 x_0$의 점추정:

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

평균반응 $E(Y_0|x = x_0) = \beta_0 + \beta_1 x_0$에 관한 추론

- $E(Y_0|x_0)$의 구간추정을 위해 $\hat{Y}_0$의 분포에 대해서 알아보자.

$$\begin{aligned}
\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = (\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 x_0 \\
&= \sum_{i=1}^{n} \frac{y_i}{n} - \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \bar{x} + \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} x_0 \\
&= \sum_{i=1}^{n} \left[\frac{1}{n} + \frac{(x_i - \bar{x})}{S_{xx}}(x_0 - \bar{x}) \right] y_i
\end{aligned}$$

- $\hat{Y}_0$은 독립적으로 정규분포를 따르는 y_i들의 선형결합이므로, $\hat{Y}_0$도 정규분포를 따른다고 할 수 있다.

평균반응 $E(Y_0|x = x_0) = \beta_0 + \beta_1 x_0$에 관한 추론

- $\hat{Y}_0$의 기대값과 분산은 다음과 같다.

$$E(\hat{Y}_0) = E(\hat{\beta}_0 + \hat{\beta}_1 x_0) = \beta_0 + \beta_1 x_0$$

$$\begin{aligned} Var(\hat{Y}_0) &= Var(\hat{\beta}_0 + \hat{\beta}_1 x_0) = Var(\bar{y} + \hat{\beta}_1(x_0 - \bar{x})) \\ &= Var(\bar{y}) + Var((x_0 - \bar{x})\hat{\beta}_1) - 2Cov(\bar{y}, (x_0 - \bar{x})\hat{\beta}_1) \\ &= \frac{\sigma^2}{n} + \frac{(x_0 - \bar{x})^2 \sigma^2}{S_{xx}} = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right) \end{aligned}$$

- $\hat{\beta}_0 + \hat{\beta}_1 x_0$를 표준화하고 분모의 σ^2을 그 추정량인 $\hat{\sigma}^2$로 바꾸어 주면 자유도 $(n-2)$인 t분포를 따른다고 알려져 있다.

$$\frac{(\hat{\beta}_0 + \hat{\beta}_1 x_0) - (\beta_0 + \beta_1 x_0)}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)}} \sim t(n-2)$$

평균반응 $E(Y_0|x = x_0) = \beta_0 + \beta_1 x_0$에 관한 추론

- $\beta_0 + \beta_1 x_0$에 관한 $100(1-\alpha)\%$ 신뢰구간:

$$(\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm t_{\alpha/2}(n-2)\ \hat{\sigma}\sqrt{n^{-1} + (x_0 - \bar{x})^2 / S_{xx}}$$

- $\beta_0 + \beta_1 x_0$에 대한 유의성 검정:

  귀무가설 $H_0 : \beta_0 + \beta_1 x_0 = \mu_0$에 관한 검정통계량은

$$T = \frac{(\hat{\beta}_0 + \hat{\beta}_1 x_0) - \mu_0}{\hat{\sigma}\sqrt{n^{-1} + (x_0 - \bar{x})^2 / S_{xx}}}$$

  이고, 이의 관측값을 t라고 하면 다음이 성립한다.

	유의확률 P	유의수준 α의 기각역
(a) $H_1 : \beta_0 + \beta_1 x_0 > \mu_0$	$P = P(T \geq t)$	$t \geq t_\alpha(n-2)$
(b) $H_1 : \beta_0 + \beta_1 x_0 < \mu_0$	$P = P(T \leq t)$	$t \leq -t_\alpha(n-2)$
(c) $H_1 : \beta_0 + \beta_1 x_0 \neq \mu_0$	$P = P(\lvert T\rvert \geq \lvert t\rvert)$	$\lvert t\rvert \geq t_{\alpha/2}(n-2)$

평균반응에 관한 추론 예제[예 8.2]

Example

(a) 사용년수가 $x = 2.5$년인 중고차의 평균 가격이 2,800,000원보다 높은지 유의수준 1%에서 검정하여라.

sol) 검정하고자 하는 가설은 다음과 같다.

$$H_0 : \beta_0 + \beta_1 2.5 = 2.8 \quad H_1 : \beta_0 + \beta_1 2.5 > 2.8$$

- t 검정통계량의 관측값은

$$t = \frac{(\hat{\beta}_0 + \hat{\beta}_1 2.5) - 2.8}{\hat{\sigma}\sqrt{n^{-1} + (2.5 - \bar{x})^2 / S_{xx}}} = 3.293$$

이고, 부록의 t분포표에서 $t_{0.01}(10) = 2.764$이므로, 검정통계량의 값이 기각역에 속하는 것을 알 수 있다.

$$3.293 = t \geq t_{0.01}(10)$$

따라서 $\alpha = 0.01$에서 귀무가설을 기각할 수 있고, 2.5년 지난 중고차의 평균 가격이 2,800,000원보다 높다는 뚜렷한 증거가 있다.

Example

(b) 사용년수가 $x = 1, \ 3, \ 5$일 때에 중고차의 평균 가격의 95% 신뢰구간을 각각 구하고 이를 그래프로 그려 보아라.

- 앞에서의 요약 통계량으로부터, $x = 1$ 일 때 평균 가격의 95% 신뢰구간은 다음과 같이 주어진다.

$$
\begin{aligned}
&(\hat{\beta}_0 + \hat{\beta}_1 1) \pm t_{0.025}(10)\ \hat{\sigma}\sqrt{n^{-1} + (1-\bar{x})^2/S_{xx}} \\
&= (5.133 - 0.859) \pm (2.228)\sqrt{0.0289}\sqrt{1/12 + (1-3.308)^2/24.649} \\
&= 4.274 \pm 0.207
\end{aligned}
$$

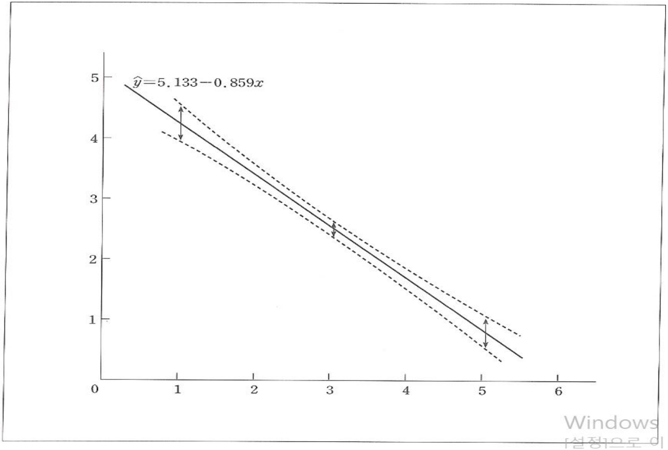
- 같은 방법으로 $x = 3, \ 5$일 때 평균 가격 $\beta_0 + \beta_1 x$에 대한 95% 신뢰구간은 각각 다음과 같이 구할 수 있다.

  $x = 3$ 일 때 2.556 ± 0.112

  $x = 5$ 일 때 0.838 ± 0.169

평균반응의 95% 신뢰구간

- 앞 예제에서 평균반응의 95% 신뢰구간과 상한, 하한을 그려보자.



평균반응의 95% 신뢰구간

- 신뢰구간에 $(x_0 - \bar{x})^2$의 항이 들어 있으므로 신뢰구간의 너비는 $x_0 = \bar{x}$일 때 가장 짧고, $\bar{x}$에서 멀어질수록 폭이 커짐을 알 수 있다.
- 따라서 설명변수 x의 값이 원자료의 범위를 많이 이탈하는 경우 주의를 해야 한다.
- $\bar{x}$에서 너무 멀리 그리고 추정된 회귀식에 이용된 x의 범위를 많이 이탈하는 경우, 그 예측력에는 신뢰성이 떨어진다는 것을 의미한다.

모형의 진단(diagnostic) 및 처방(remedy)

- 최소제곱법으로 추정된 회귀식을 어떻게 평가할 수 있을까?
- 단순회귀추정식 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$이 설명변수와 반응변수 사이의 연관성을 얼마나 잘 설명하고 있는지, 회귀모형 적용의 타당성을 검토하여 보자.
- 우선적으로 산점도(scatter plot)를 그려서 선형회귀모형이 적합한지 판단할 수 있다.

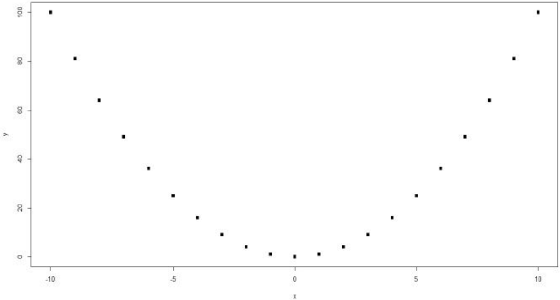
  $\Longrightarrow$통계적 유의성을 점검할 수 있는 수치를 제공하진 못하지만,
  직관적인 판단이 가능하게 해준다.

- 회귀진단 항목

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad (i = 1, 2, \cdots, n) \text{ 에서}$$

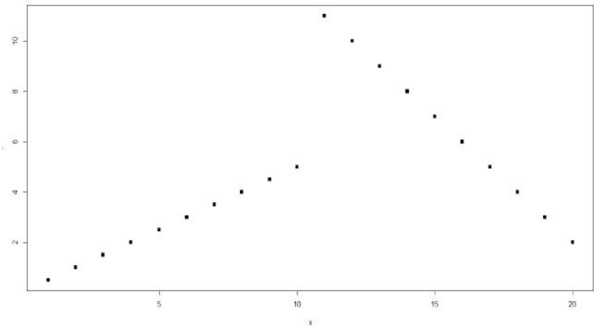
(a) $E(\epsilon_i) = 0$, i.e, $E(Y_i|x_i) = \beta_0 + \beta_1 x_i$ (선형성)
(b) $Var(\epsilon_i) = \sigma^2 > 0$ (등분산성)
(c) $\epsilon_1, \cdots, \epsilon_n$은 서로 독립 (독립성)
(d) $\epsilon_i \sim N(0, \sigma^2)$ (정규성)
(e) 이상치(outlier)나
(f) 영향력이 큰 관찰치(influential observation)의 존재

산점도를 이용한 선형성 진단



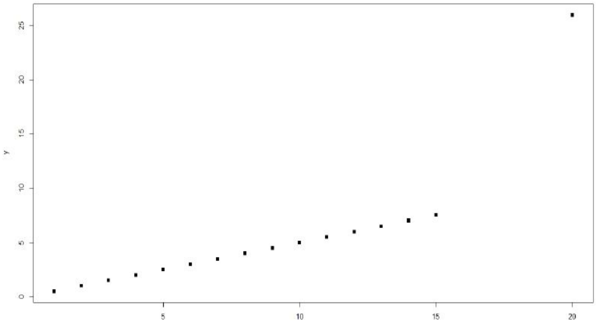
- 처방: 두 변수가 비선형 관계를 보이면 변수변환 등의 기법을 사용하거나 2차 이상의 다항식 또는 로그 함수 등의 비선형 모형(nonlinear model)을 적용

산점도를 이용한 선형성 진단

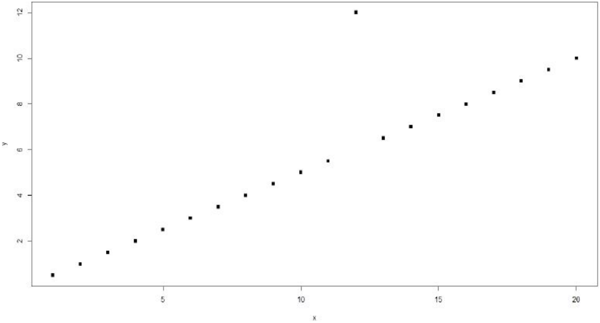


- 처방: 두 개의 그룹으로 나누어 각각 선형 모형을 시도

산점도를 이용한 이상치 진단



- 처방: 이상치가 단순 오기에 의한 실수인지 점검하고, 단순 오기가 아니라면 이상치를 그대로 포함할지, 제거한 후 선형모형을 적용할지 결정



- 처방: 입력 실수인지 확인을 먼저 하고, 영향력을 판단하기 위하여 관측값을 제거한 후 모형을 추정하고 추정치 값들, R^2 등이 얼마나 차이 나는지 비교.

잔차분석(analysis of residual)

- 모형의 검토에서 흔히 다음의 잔차를 이용한다.

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

where $y_i = i$번째 반응변수 관찰값, $\hat{y}_i = i$번째 반응변수 추정값

- 오차에 대한 가정사항: $\epsilon_i \sim_{i.i.d} N(0, \sigma^2)$
- 회귀모형의 가정을 표준화된 오차항을 이용하여 나타내보자.

$$\frac{\epsilon_i - 0}{sd(\epsilon_i)} = \frac{\epsilon_i(= Y_i - (\beta_0 + \beta_1 x_i))}{\sigma} \sim_{i.i.d} N(0, 1), \quad i = 1, \cdots, n$$

- (반)스튜던트화 잔차(semi-studentized residual): 잔차 e_i를 그 표준편차의 추정값인 $\widehat{sd}(e_i)$으로 나눈 값

$$\hat{e}_{st,i} = \frac{e_i}{\widehat{sd}(e_i)} = \frac{e_i(= y_i - \hat{y}_i)}{\hat{\sigma}}, \quad i = 1, \cdots, n$$

잔차분석(analysis of residual)

- 잔차분석:

- 스튜던트화 잔차를 표준화된 오차항의 관측값처럼 생각하여 회귀모형의 가정을 검토할 수 있다.

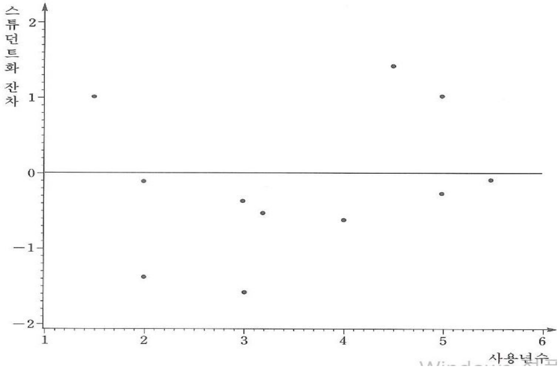
- 스튜던트화 잔차인

$$\hat{e}_{st,i} = \frac{e_i}{\hat{\sigma}}, \quad i = 1, \cdots, n, \quad \longrightarrow \quad \left\{ \frac{y_1 - \hat{y}_1}{\widehat{sd}(e_1)}, \frac{y_2 - \hat{y}_2}{\widehat{sd}(e_2)}, \cdots, \frac{y_n - \hat{y}_n}{\widehat{sd}(e_n)} \right\}$$

들이 표준정규분포에서의 서로 독립인 n개의 관측값과 유사하게 나타나는가를 검토하여, 단순선형회귀모형의 적용 타당성을 알아본다.

- 흔히 좌표평면에 점 $(x_1, \hat{e}_{st,1}), \cdots, (x_n, \hat{e}_{st,n})$을 나타내는 잔차도(residual plot)를 이용하여 분석을 하게 된다.

- 잔차도에서 스튜던트화 잔차들이 다음과 같이 나타나면



1. 선형성: 대략 0에 관하여 대칭적으로 나타남
2. 등분산성: 설명변수의 값에 따른 잔차의 산포가 크게 다르지 않음
3. 독립성: 점들이 특정한 형식을 가지고 나타나지 않음
4. 정규성: 대부분의 점이 ±2의 범위 내에 있음

=> 단순선형회귀모형을 적용하는 것이 타당해 보인다.

단순선형회귀모형의 가정에 어긋나는 경우

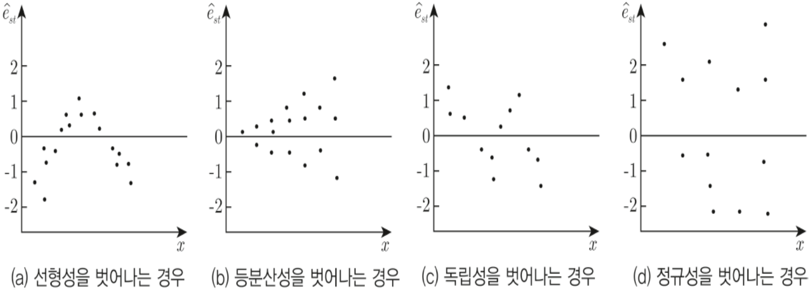


그림 8-10 단순선형회귀모형의 가정에 어긋나는 경우

통계적 추론의 순서

- 단순선형회귀모형을 적용한 통계적 추론시 분석 과정

1. 산점도를 그려서 대략적인 직선 관계를 파악한다.
2. 최소제곱회귀직선과 평균제곱오차를 구하여 잔차분석을 행한다. 이 때, 스튜던트화 잔차들이 대체로 0 에 관해 대칭적으로 나타나고 ± 2이내의 범위에 들어가며 특별한 패턴을 가지지 않음을 확인한다.
3. 잔차분석을 통과하면 신뢰구간이나 검정과 같은 추론을 행한다.

단순선형회귀분석 예제

Example

공동주택의 가구당 월간 전기요금이 월평균 기온에 따라 어떤 변화하는지 알아보기 위해 조사한 결과가 아래 표와 같다.

표 8-6 월평균 기온에 따른 월간 전기요금

(단위 : ℃, 천 원)

기온	전기요금	기온	전기요금	기온	전기요금	기온	전기요금	기온	전기요금
−2.2	14.0	0.8	12.2	8.2	12.9	15.2	11.9	21.8	10.3
−1.9	13.1	1.8	13.0	9.2	11.3	16.3	10.9	22.3	9.1
−1.7	14.4	3.9	11.3	14.2	10.8	21.1	9.6	23.6	7.5
−1.3	13.1	7.0	10.5	14.5	10.0	21.1	8.1	23.6	10.5
−0.7	14.8	8.0	9.7	14.9	9.9	21.5	9.2	24.8	10.0

(a) 단순선형회귀모형 적용이 가능한지 잔차분석을 하여라.

sol) [표 8-6]의 자료에 대하여 산점도를 그려보면 두 변수 사이에 대략 직선관계가 성립함을 알 수 있다.

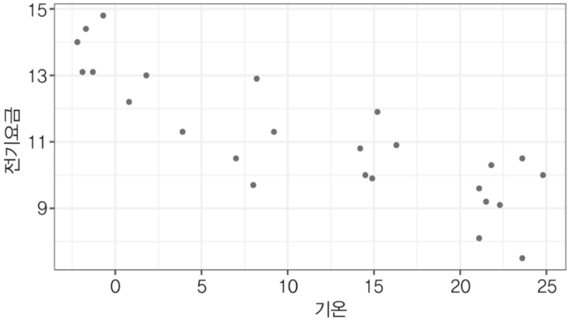


그림 8-11 [표 8-6]에 대한 산점도

- 이 자료에 최소제곱회귀직선을 적합하기 위한 통계량을 계산하면

$$n = 25, \quad \bar{x} = 11.44, \quad \bar{y} = 11.124$$

$$S_{xx} = 2210.44, \quad S_{yy} = 88.8856, \quad S_{xy} = -374.9945$$

- 이로부터 최소제곱회귀직선과 오차분산의 추정값을 구하면 다음과 같고

$$\hat{\beta}_1 = S_{xy}/S_{xx} = -0.17, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 13.06$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 13.06 - 0.17x$$

$$\hat{\sigma}^2 = SSE/(n-2) = \{S_{yy} - S_{xy}^2/S_{xx}\}/(n-2) = 1.0987$$

- 최소제곱회귀직선을 산점도에 그려보면 다음과 같다.

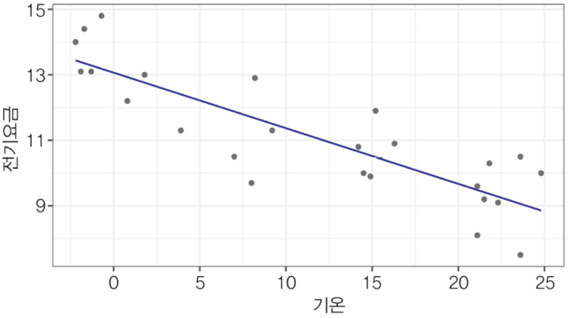


그림 8–12 [표 8–6]에 대한 산점도와 최소제곱회귀직선

- 최소제곱회귀직선 $\hat{y}_i = 13.06 - 0.17x_i$와 $\hat{\sigma} = \sqrt{\hat{\sigma}^2} = 1.0482$를 이용하여 스튜던트화 잔차인 $\hat{e}_{st,i} = \frac{y_i - \hat{y}_i}{\widehat{sd}(Y_i - \hat{Y}_i)}$을 계산하여 잔차도를 그려보았다.
- 잔차도에서 잔차들은 대략 0에 관해 대칭적이며 설명변수의 값에 따른 산포도 비슷하고, 특별한 패턴을 갖고 있지 않으며 모두 ± 2의 범위 이내에 있음을 알 수 있다.
- 이러한 회귀모형을 가정할 때, 회귀직선이 자료의 전체 산포 중에서 얼마나 설명하는가를 알기 위하여 결정계수를 계산하면

$$R^2 = \frac{SSR}{SST} = \frac{S_{xy}^2}{S_{xx}S_{yy}} = 0.7157$$

로서, 전체 산포의 약 72%가 회귀직선에 의하여 설명된다.

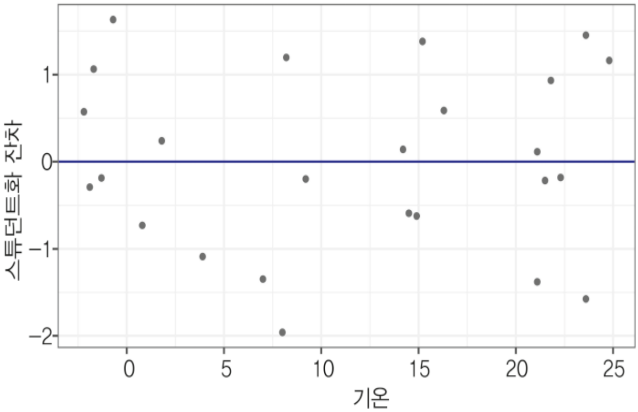


그림 8-13 [표 8-6]에 대한 잔차도

Example

(b) 유의수준 5%에서 회귀직선의 유의성검정을 하여라.

sol) 회귀직선의 유의성을 판단하기 위한 가설은 다음과 같다.

$$H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0$$

- 제곱합을 계산하면

$$SST = S_{yy} = 88.8856$$
$$SSR = S_{xy}^2 / S_{xx} = 63.6165$$
$$SSE = SST - SSR = 25.2691$$

이고 각각의 자유도는 25-1=24, 1, 25-2=23 이다.

- F검정통계량의 관측값은 평균제곱의 비인

$$f = \frac{SSR/1}{SSE/23} = 57.904$$

이고 F분포표에서 $F_{0.05}(1,23) = 4.96$ 이므로,

$$f \geq F_{0.05}(1,23)$$

이 성립한다. 즉, 유의수준 5%에서 회귀직선은 유의하다.

- 분산분석표로 과정을 정리하면 다음과 같다.

요인	제곱합	자유도	평균제곱	F값	유의확률
회귀	$SSR = 63.6165$	1	MSR=63.6165	f=57.904	0.0001
잔차	SSE=25.02691	23	MSE=1.0987		
계	$SST = 88.8856$	24			

Example

(c) 월평균기온이 10일 때, 월간 평균 전기요금에 대한 95%의 신뢰구간을 구하여라.

sol) 월평균기온이 $x = 10$일 때 $E(Y|x) = \beta_0 + \beta_1 10$의 추정값은

$$\hat{\beta}_0 + \hat{\beta}_1 10 = 13.06 + (-0.17)10 = 11.36$$

이 추정값의 표준오차를 계산하면

$$\hat{\sigma}\sqrt{n^{-1} + (x_0 - \bar{x})^2 / S_{xx}} = 0.2121$$

- 한편, t분포표에서 $t_{0.025}(23) = 2.069$이므로 $\beta_0 + \beta_1 10$의 95%신뢰구간은 (단위:1000원) 다음과 같다.

$$(\hat{\beta}_0 + \hat{\beta}_1 10) \pm t_{0.025}(23)\ \hat{\sigma}\sqrt{n^{-1} + (x_0 - \bar{x})^2 / S_{xx}} = 11.36 \pm 0.44$$