8장. 상관분석과 회귀분석

박명현

서울대학교 학부대학

서론

- 자료를 이용하여 변수와 변수 사이의 관계에 대한 통계적 추론을 하고자 한다.

  예) 지능지수와 학업성적, 흡연량과 수명 등

- 상관분석(correlation analysis)

  - 두 변수 사이의 관계 유무 또는 관계의 강도에 대한 연관성을 분석하는 방법.

  - 표본상관계수 r을 이용하여 모상관계수 ρ를 추론.

- 단순선형회귀분석(simple linear regression analysis):

  두 변수 사이의 함수 관계로서 가장 기본적인 형태인 일차함수관계, 즉 직선 관계를 모형으로 하여 분석

- 다중회귀분석(multiple regression analysis):

  두 개 이상의 변수가 한 변수에 영향을 줄 때, 한 변수를 여러 변수의 함수로 나타내어 분석하는 방법

상관분석(correlation analysis)

상관분석

- 모상관계수 ρ: 각 개체의 두 가지 특성을 변수 X, Y로 나타낼 때, 두 변수 사이에 직선관계가 어느 정도 강하게 있는가를 나타냄

$$\rho = Corr(X, Y)$$

- 표본상관계수 r: 모상관계수 ρ의 추정값으로 사용.

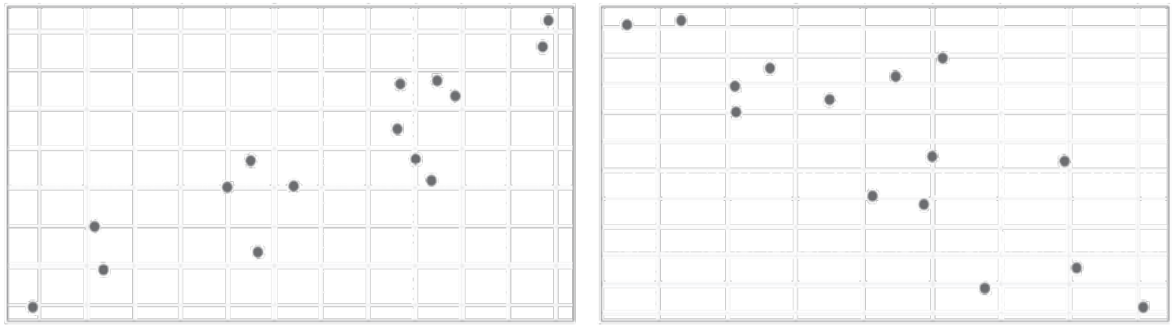
  - 두 특성에 대한 자료가 $(x_1, y_1), \cdots, (x_n, y_n)$ 으로 주어지면

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

1. 항상 $-1 \leq r \leq 1$
2. $r \approx 1$: 두 변수 사이에 기울기가 양수인 직선관계가 강함
3. $r \approx -1$: 두 변수 사이에 기울기가 음수인 직선관계가 강함
4. $r \approx 0$ 두 변수 사이에 직선관계가 없음

상관분석

- 두 변수의 관계에 대한 대략적인 파악은 산점도로 가능



(a) $r = 0.9$ (b) $r = -0.8$

(c) $r = 0$ (d) $r = 0$

그림 8-1 여러 가지 경우의 산점도와 표본상관계수

- r의 값이 0에 가까운 것은 두 변수 사이에 직선관계가 약한 것을 시사하는 것이지 아무런 관계가 없음을 뜻하는 것은 아니다.

표본상관계수 예제

Example (8.1)

다음 자료는 어느 고등학교 학생 중에서 랜덤하게 추출된 20명의 수학 능력 모의시험에서 언어영역과 외국어영역의 점수이다.

학생번호	1	2	3	4	5	6	7	8	9	10
언어 영역	42	38	51	53	40	37	41	29	52	39
외국어 영역	30	25	34	35	31	29	33	23	36	30

학생번호	11	12	13	14	15	16	17	18	19	20
언어 영역	45	34	47	35	44	48	47	30	29	34
외국어 영역	32	29	34	30	28	29	33	24	30	30

이 자료에 대한 산점도를 그리고, 언어영역(x)과 외국어영역(y) 성적 사이의 표본상관계수를 구하고 해석하여라.

sol) 자료로부터 다음의 값을 계산하였다.

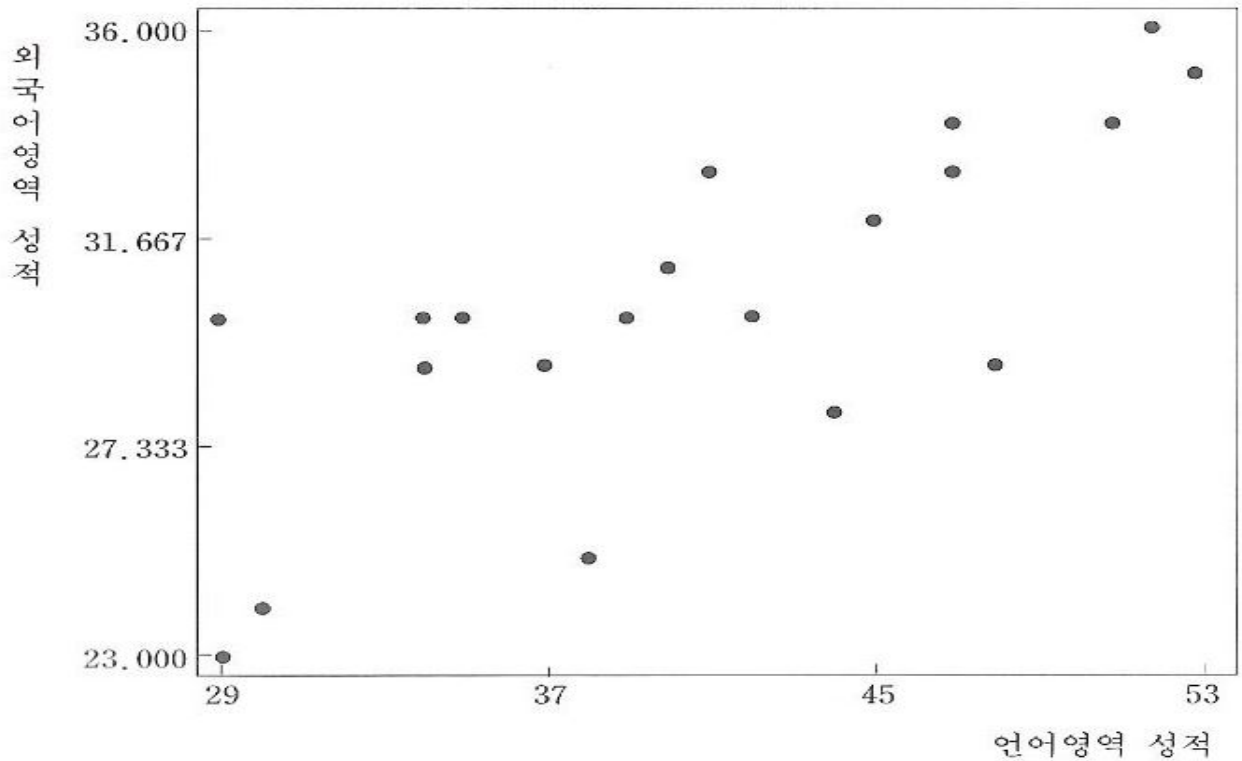
$\bar{x} = 815/20$, $\bar{y} = 605/20$,

$\sum x_i y_i = 25033$, $\sum x_i^2 = 34295$, $\sum y_i^2 = 18533$

- 간편계산식을 이용하여 표본상관계수를 구해보자.

$$
\begin{aligned}
r &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \\
&= \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum x_i^2 - n\bar{x}^2}\sqrt{\sum y_i^2 - n\bar{y}^2}} \\
&= \frac{25033 - n\bar{x}\bar{y}}{\sqrt{34295 - n\bar{x}^2}\sqrt{18533 - n\bar{y}^2}} = 0.757
\end{aligned}
$$

Figure: 언어영역과 외국어영역 성적간의 산점도



- 두 점수 사이에 양의 선형관계가 있고 그 정도는 76% 이다.
- 이는 두 점수간에 존재하는 상관의 성질과 정도만을 의미하며, 두 변수간의 관계를 인과적으로 설명하지는 않는다.

상관관계의 유무에 관한 검정

- 표본상관계수 r을 이용하여 모상관계수 ρ를 추론
- 가정사항: 이변량정규모집단
- 귀무가설 $H_0 : \rho = 0$에 대한 검정통계량은

$$T = \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}}$$

이고, 이의 관측값을 t라고 할 때, 다음이 성립한다.

		유의확률 P	유의수준 α의 기각역
(a)	$H_1 : \rho > 0$	$P = P(T \geq t),\ T \sim t(n-2)$	$t \geq t_\alpha(n-2)$
(b)	$H_1 : \rho < 0$	$P = P(T \leq t),\ T \sim t(n-2)$	$t \leq -t_\alpha(n-2)$
(c)	$H_1 : \rho \neq 0$	$P = P(\lvert T\rvert \geq \lvert t\rvert),\ T \sim t(n-2)$	$\lvert t\rvert \geq t_{\alpha/2}(n-2)$

상관분석 예제

Example

예제 8.1의 경우에 고등학교 학생 전체의 언어영역과 외국어영역 성적이 이변량정규분포를 따른다고 할 때, 이들 성적 사이의 상관관계가 있는가를 유의수준 5%에서 검정하여라.

sol) 검정하고자 하는 가설은

$$H_0 : \rho = 0 \qquad H_1 : \rho \neq 0$$

이고, $n = 20$, $r = 0.757$이므로 검정통계량의 관측값은

$$t = \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}} = \sqrt{20-2}\,\frac{0.757}{\sqrt{1-0.757^2}} = 4.915$$

- t분포표로부터 $t_{0.025}(18) = 2.101$이고

$$|t| = |4.915| \geq 2.101$$

이므로, 유의수준 5%에서 $H_0 : \rho = 0$을 기각한다. 즉, 상관관계가 있다는 뚜렷한 증거가 있다.

회귀분석(regression analysis)

회귀분석

- 회귀분석(regression analysis): 두 변수 사이의 함수 관계를 파악하여 한 변수의 값으로부터 다른 변수의 값에 대한 예측.

$$Y = f(x) + \epsilon$$

- 설명변수(explanatory variable): x, 다른 변수에 영향을 주는 변수
  - 예측변수(predictor variable), 독립변수(independent variable)
- 반응변수(response variable): Y, 설명변수에 영향을 받는 변수
  - 종속변수(dependent variable)
- 회귀분석의 목적: 주어진 자료를 이용하여 설명변수와 반응변수의 관계를 구체적인 함수 $f(x)$ 형태로 나타내고, 설명변수의 값으로부터 반응변수의 값을 예측 $\hat{f}(x)$ 하고자 함

단순선형회귀분석 예시

Example (8.2)

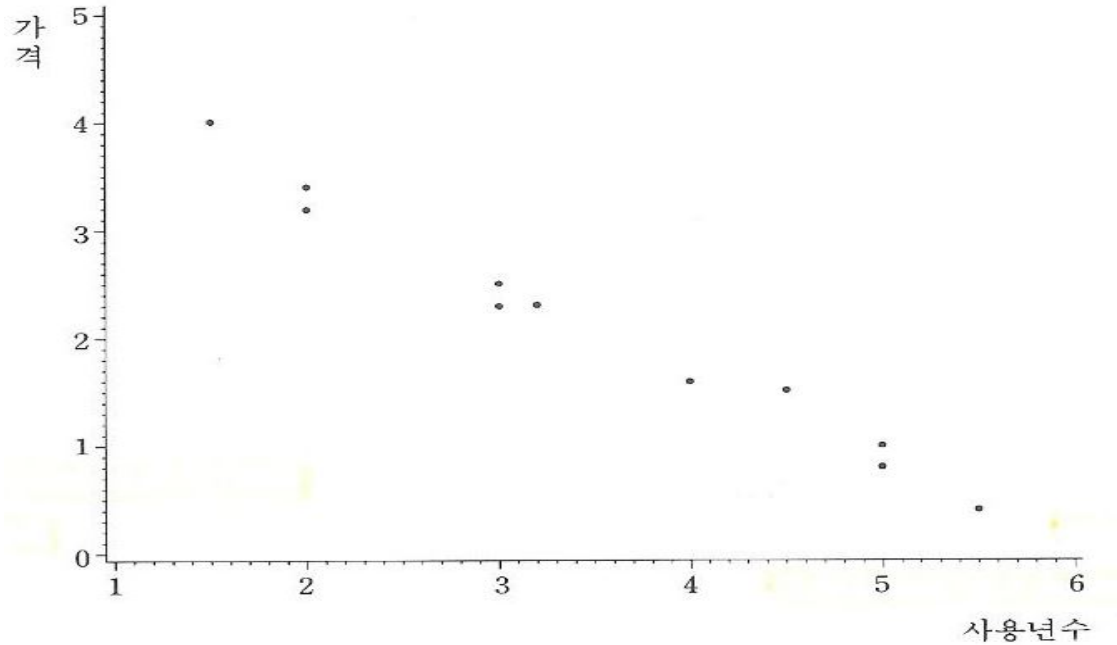
어떤 승용차의 가격이 출고 연도가 지남에 따라 얼마나 떨어지는가를 알아보기 위하여, 이 승용차에 대한 중고판매가격에 대한 표본조사를 한 결과 다음의 자료를 얻었다.

Table: [8.1] 사용년수에 따른 중고차 가격 (단위: 백만 원)

사용년수(x)	1.0	1.5	2.0	2.0	3.0	3.0	3.2	4.0	4.5	5.0	5.0	5.5
중고차가격(y)	4.5	4.0	3.2	3.4	2.5	2.3	2.3	1.6	1.5	1.0	0.8	0.4

- 이 자료로부터 사용년수(x)와 중고차가격(y)의 대략적인 관계의 파악을 위하여 산점도를 그려보았다.
- 설명변수: $(x_1, \cdots, x_{12})$, 반응변수: $(y_1, \cdots, y_{12})$

Figure: 사용년수에 따른 중고차 가격의 변화



- 사용년수가 증가함에 따라 중고차 가격은 점차 감소: $5.1 - 0.8x$
- Y 변수에 있는 변동은 오차항 ϵ으로 설명 가능
- 단순선형회귀모형의 가정: $Y = \beta_0 + \beta_1 x + \epsilon$

단순선형회귀모형(simple linear regression model)

- 설명변수의 값에 대응하는 반응변수의 값이 $\beta_0 + \beta_1 x$주위에 나타난다고 생각될 때, 다음의 선형식을 고려해 보자.

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \cdots, n$$

- Y_i: 종속변수, 반응변수
- x_i: 독립변수, 설명변수, 주어진 상수(constant)값
- 절편 β_0과 기울기 β_1: 미지의 모수(unknown parameters)
- ϵ_i: 오차. Y와 x의 선형관계로 설명할 수 없는 나머지부분

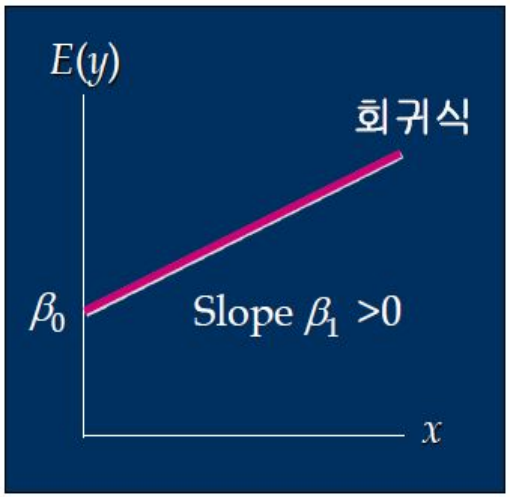
$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

- 모회귀직선: $E(Y_i|x_i) = \beta_0 + \beta_1 x_i$; a function of x_i
  - 설명변수에 따른 반응변수의 평균값을 나타냄

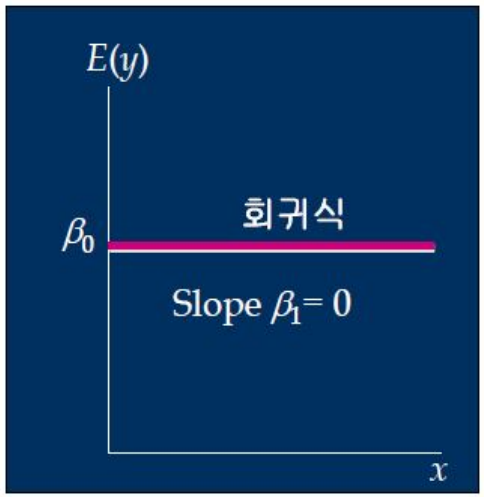
모회귀계수(절편 β_0, 기울기 β_1)

- β_0: 미지의 모수이며 y축의 절편으로 $x = 0$일 때 y의 값이다.
- β_1 : 미지의 모수이며 기울기로 x가 1단위 증가할 때 y의 변동량으로 해석한다.

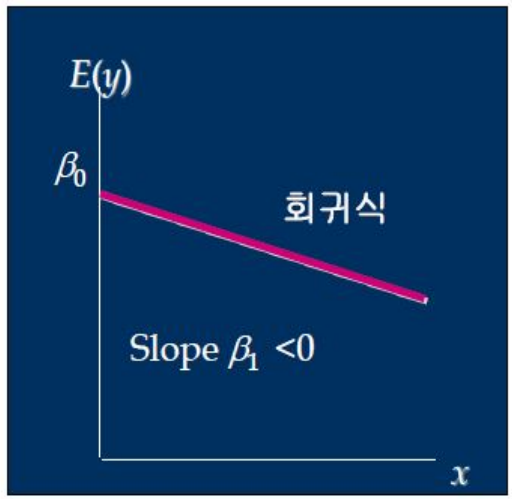
- 양의 선형관계 (β_1>0)



- 연관성 없음(β_1 =0)



- 음의 선형관계(β_1 <0)



오차항 ϵ_i

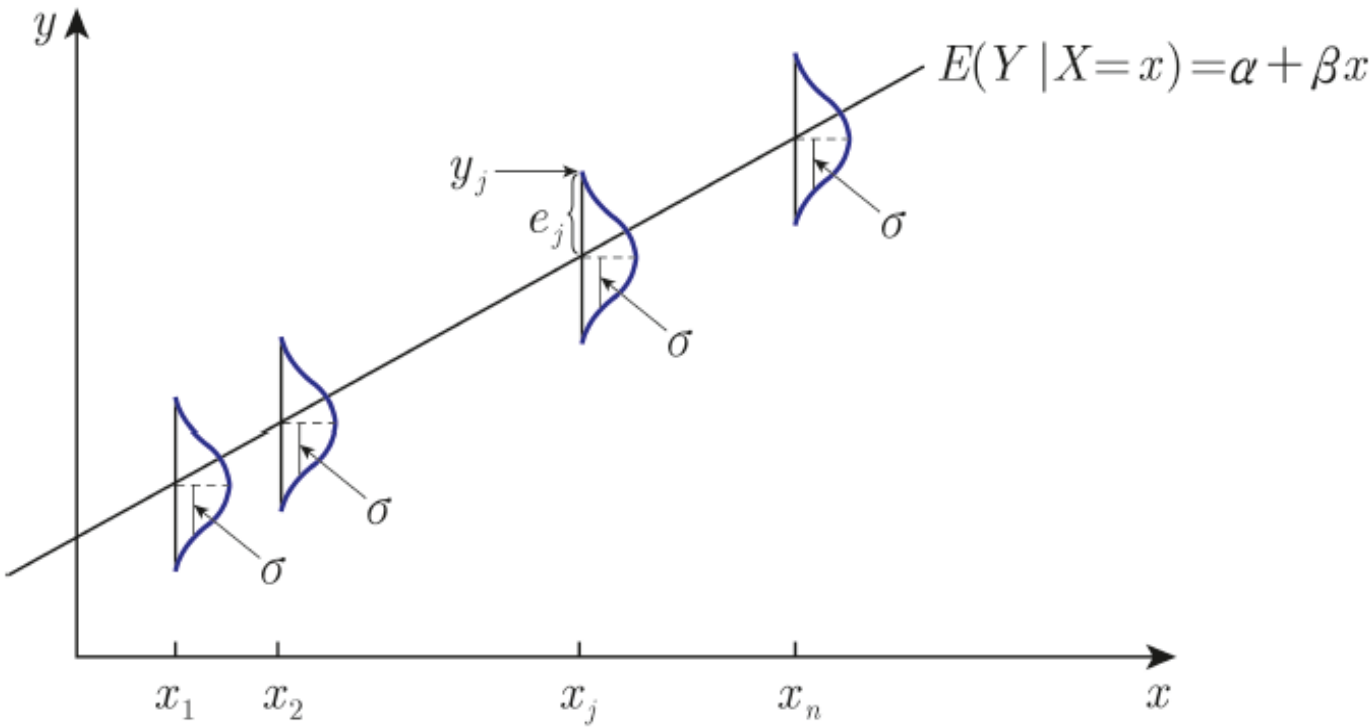


그림 8–4 단순선형회귀모형

- $\epsilon_i \sim_{i.i.d} N(0, \sigma^2)$
- $\epsilon_1, \cdots, \epsilon_n$은 서로 독립적으로 관측된다고 가정

오차항 ϵ_i의 가정사항

- 오차항 ϵ_i의 가정사항 $(i = 1, 2, \cdots, n)$

$$\epsilon_i \sim_{i.i.d} N(0, \sigma^2) \ (\sigma^2 unknown)$$

1) the mean of the error variable ϵ_i is 0 $(E(\epsilon_i) = 0)$
2) the variance of ϵ_i is $\sigma^2 (Var(\epsilon_i) = \sigma^2 > 0)$
3) the errors ϵ_i and ϵ_j are independent (for all $i, j; i \neq j$)
4) ϵ_i is normally distributed

단순선형회귀모형의 가정사항

- $\epsilon_i \sim_{i.i.d} N(0, \sigma^2) \longrightarrow Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ (σ^2 *unknown*)

  1) $E(Y_i|x_i) = \alpha + \beta_1 x_i$; a function of x_i: 선형성
     - $E(Y_i|x_i) = E(\beta_0 + \beta_1 x_i + \epsilon_i) = E(\beta_0 + \beta_1 x_i) + E(\epsilon_i) = \beta_0 + \beta_1 x_i$

  2) $Var(Y_i|x_i) = \sigma^2$; the same for all levels of X: 등분산성
     - $Var(Y_i|x_i) = Var(\beta_0 + \beta_1 x_i + \epsilon_i) = Var(\epsilon_i) = \sigma^2$

  3) Y_i and Y_j are independent (for all $i, j; i \neq j$) : 독립성
     - $Cov(Y_i|x_i, Y_j|x_j) = Cov(\beta_0 + \beta_1 x_i + \epsilon_i, \beta_0 + \beta_1 x_j + \epsilon_j) = Cov(\epsilon_i, \epsilon_j) = 0$

  4) Y_i is normally distributed: 정규성

- Y, ϵ: 확률변수($r.v$)
- x_i는 자료에서 주어진 값. non-random, fixed values
- β_0, β_1, σ^2: unknown parameters to be estimated

모회귀계수(절편 β_0, 기울기 β_1)의 추정

Theorem (최소제곱법(method of least squares))

단순선형회귀모형 $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i (i = 1, \cdots, n)$에서 오차의 제곱합인

$$SS(\beta_0, \beta_1) = \sum_{i=1}^{n} \{y_i - (\beta_0 + \beta_1 x_i)\}^2 = \sum_{i=1}^{n} \epsilon_i^2$$

을 최소로 하는 β_0, β_1의 값을 그 추정값으로 한다.

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{(\beta_0, \beta_1)} \sum_{i=1}^{n} \epsilon_i^2$$

- 최소제곱법에 의한 β_0, β_1의 추정량 $\hat{\beta}_0$, $\hat{\beta}_1$을 구하려면, 오차의 제곱합 $SS(\beta_0, \beta_1)$을 β_0, β_1에 대하여 각각 편미분하여 얻은 식을 으로 놓은 연립방정식을 풀면된다.

최소제곱법

- 이러한 연립방정식을 정규방정식(normal equation)이라고 하며 (1), (2)와 같이 주어진다.

$$\begin{cases} \frac{\partial SS}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \\ \frac{\partial SS}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0, \end{cases}$$

$$\begin{cases} \sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i = 0 & (1) \\ \sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0 & (2) \end{cases}$$

- 이 때, (1), (2)의식을 다음의 계산 공식을 사용하여 계산하면 편리하다.

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\bar{y}$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})y_i = \sum_{i=1}^{n} x_i(y_i - \bar{y})$$

최소제곱법

- (1),(2)에 각각 $\sum x_i$, n을 곱하면 다음과 같고

$$\begin{cases} \sum x_i \sum y_i - n\beta_0 \sum x_i - \beta_1 (\sum x_i)^2 = 0 & (3) \\ n\sum x_i y_i - n\beta_0 \sum x_i - n\beta_1 \sum x_i^2 = 0 & (4) \end{cases}$$

(3)식에서 (4)식을 빼주면 다음과 같은 방정식을 얻을 수 있다.

$$\sum x_i \sum y_i - n\sum x_i y_i - \beta_1 (\sum x_i)^2 + n\beta_1 \sum x_i^2 = 0$$

- 마지막으로 β_1로 묶어준 다음에

$$\beta_1 (n\sum x_i^2 - (\sum x_i)^2) = n\sum x_i y_i - \sum x_i \sum y_i$$

최소제곱법

- 정리하면 다음과 같이 $\hat{\beta}_1$을 구할 수 있다.

$$\hat{\beta}_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{\sum x_i y_i - \bar{x}\sum y_i}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}$$
$$= \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

- $\hat{\beta}_0$은 $\sum_{i=1}^n y_i - n\beta_0 - \beta_1 \sum_{i=1}^n x_i = 0$의 (1)번식을 n으로 나누면

$$\bar{y} - \beta_0 - \beta_1 \bar{x} = 0$$

따라서 $\hat{\beta}_0$은 다음과 같이 구한다.

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

최소제곱법

- 이렇게 얻게 되는 추정량 $\hat{\beta}_0$, $\hat{\beta}_1$를 단순선형회귀모형에서 β_0와 β_1의 최소제곱추정량(least squares estimator)이라 부르고, 다음과 같이 정리된다.

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

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

- 이로부터 구해지는 직선은 최소제곱회귀직선(least squares regression line)이라 부르고 다음과 같다.

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

(The fitted or estimated regression line)

최소제곱법 예제

Example

예제 8.2 (13p)에서, 사용년수를 x로 하고 중고차 가격을 반응변수 Y로 하여 단순선형회귀모형을 가정하자. 이 때, 표의 자료로부터 β_0와 β_1의 최소제곱추정값을 구하고 산점도에 최소제곱회귀직선을 그려 보아라.

sol) 표의 자료로부터 필요한 통계량의 값들을 계산하면

$$\bar{x} = 39.7/12 = 3.308, \quad \bar{y} = 27.5/12 = 2.292$$

$$S_{xy} = \sum x_i y_i - n\bar{x}\bar{y} = 69.81 - 12 \cdot 3.308 \cdot 2.292 = -21.169$$

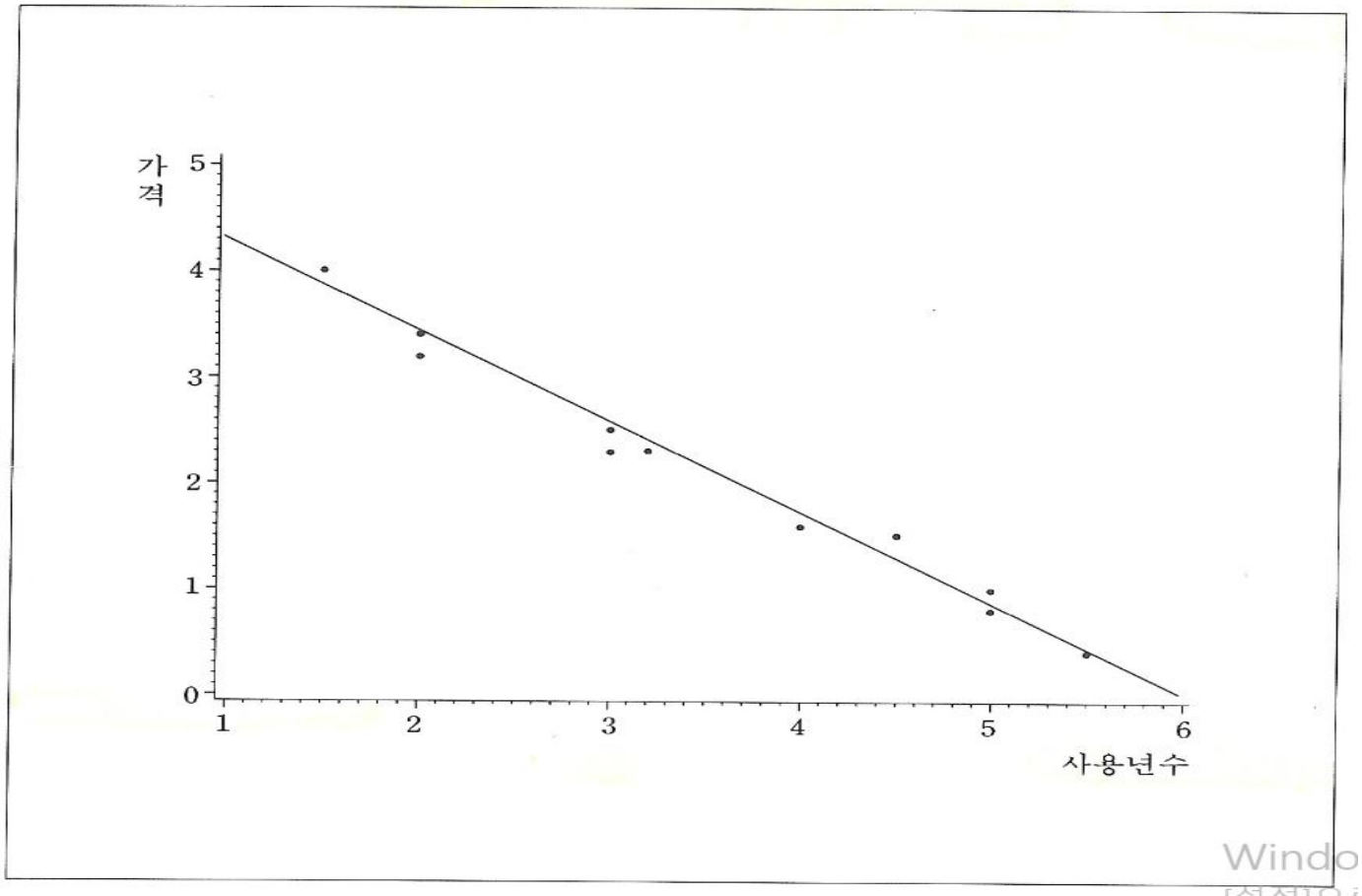
$$S_{xx} = \sum x_i^2 - n\bar{x}^2 = 155.99 - 12 \cdot 3.308^2 = 24.649$$

- 이로부터 모회귀계수의 최소제곱추정값을 계산하면 다음과 같고

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = -0.859, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 5.133$$

- 최소제곱회귀직선은 $\hat{y} = 5.133 - 0.859x$ 이다.

Figure: 표 8.1의 자료에 대한 산점도와 최소제곱회귀직선



- 산점도를 통해 단순선형회귀모형에 의한 적합이 어느 정도 사용년수에 따른 중고차 가격의 변화를 잘 설명해주고 있음을 알 수 있다.
- $x_2 = 1.5$

$E(y_i|x_i)$와 $\hat{y}_i$

- 종속변수 Y를 다음의 두 가지 방식으로 설명할 수 있다.

  $y_i = \beta_0 + \beta_1 x_i + \epsilon_i = E(y_i|x_i) + \epsilon_i$ (ideal model)

  $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i = \hat{y}_i + e_i$ (based on data estimates)

- 예제 8.2에서 $x_2 = 1.5$일 때, 종속변수의 기대값을 구해보자.

$$E(Y_2|x_2) = 5.1 - 0.8 \times 1.5 = 3.9 \neq (y_2 = 4), \quad \epsilon_4 = 0.1$$

  $x_2 = 1.5$일 때, 종속변수의 추측값을 구해보자.

$$\hat{y}_i = 5.133 - 0.859 \times 1.5 = 3.8485 \neq (y_2 = 4), \quad e_2 = 0.1555$$

- $\hat{y}_i$은 반응변수의 평균인 $E(Y_i|x_i) = \beta_0 + \beta_1 x_i$의 추정값이다.

$$\hat{y}_i = \widehat{E(Y_i|x_i)}$$

잔차(residual) e_i

- 잔차(residual): 설명변수 x_i에 대응되는 추측값 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$과 실제 관측값 y_i의 차이.

  - 추정된 회귀식이 반응 변수 관찰치를 설명하지 못하고 있는 부분.

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \quad (i = 1, 2, \cdots, n)$$

  - e_i is small $\rightarrow$ Our fitted model is good
  - e_i is large $\rightarrow$ We need another model

- 잔차 e_i 들은 오차항 ϵ_i 의 관측값인 것처럼 생각할 수 있고, 오차분산 σ^2의 크기에 따라 크거나 작게 나타난다.
- 따라서, 오차분산 $Var(\epsilon_i) = \sigma^2$에 대한 정보를 제공해주는 잔차들의 제곱합을 이용하여 오차분산을 추정하자.

오차분산 σ^2의 추정

- 오차제곱합(error sum of squares, SSE)

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

- 오차제곱합(SSE)을 이것의 자유도 $(n-2)$로 나눈 평균제곱오차(mean squared error, MSE)을 이용하여, σ^2을 추정할 수 있다.

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-2} = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

즉, $E(MSE) = \sigma^2$이다.

제곱합의 분해(decomposition)

- 최소제곱회귀직선에 의한 단순선형회귀모형의 적합이 관측 결과를 얼마나 잘 설명하고 있는가에 대하여 생각해 보자.

- 반응변수의 관측값 $y_1, \cdots, y_n$이 그 평균인 $\bar{y}$와 떨어진 정도인 $y_i - \bar{y}$를 다음과 같이 두 편차의 합으로 나타낼 수 있다.

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

- $y_i - \bar{y}$: 총편차(total deviation)
- $y_i - \hat{y}_i$: 잔차로서 오차항에 기인하는 편차
- $\hat{y}_i - \bar{y}$: 설명변수 x를 고려한 회귀직선에 의하여 설명되는 편차

제곱합의 분해(decomposition)

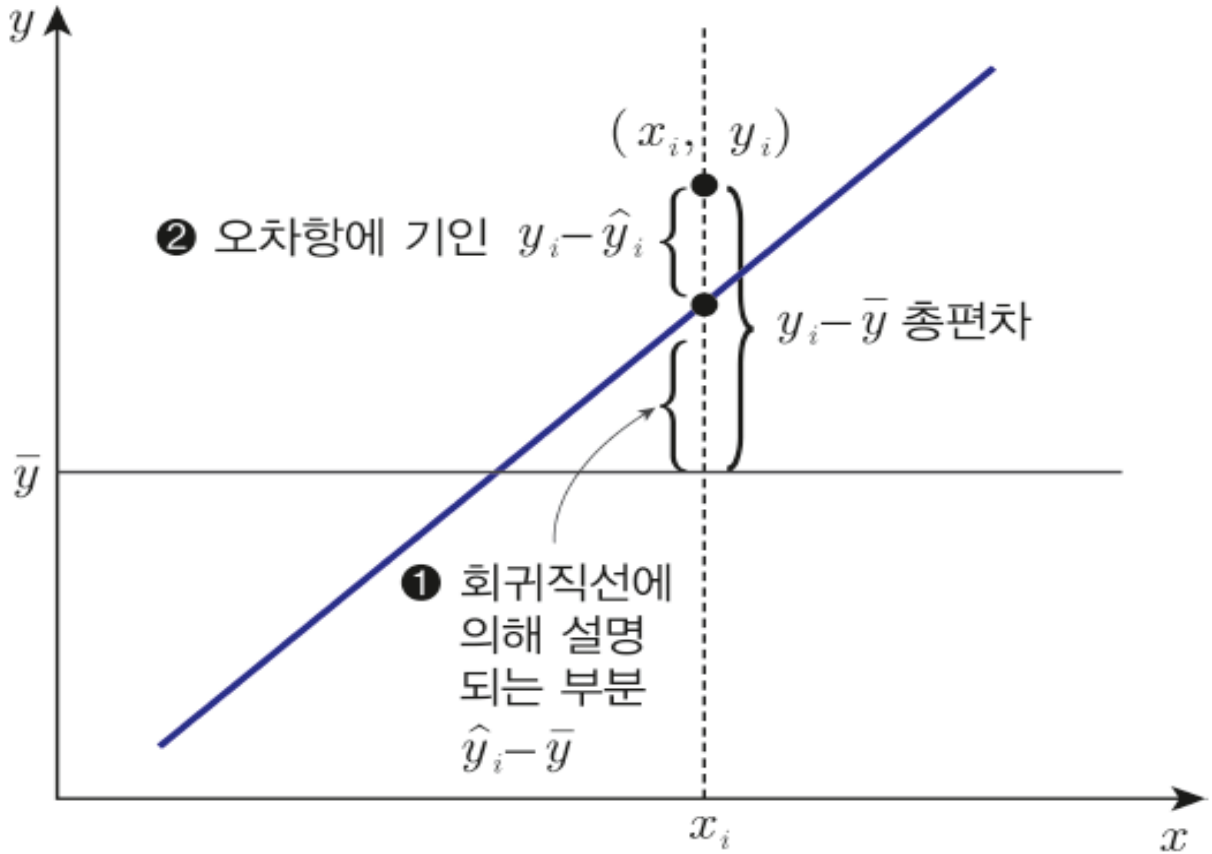


그림 8–6 총편차 $y_i - \bar{y}$의 분해

제곱합의 분해(decomposition)

- 총편차의 분해식의 양변을 제곱하여 합하면 다음이 성립함을 알 수 있다.

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

- 총편차의 제곱합=잔차의 제곱합+회귀로 설명되는 편차의 제곱합
- 자료의 변동= 오차에 기인하는 부분+회귀직선에 기인하는 부분

- SST: 총 제곱합(total sum of squares)
- SSE: 오차 제곱합(error sum of squares)
- SSR: 회귀 제곱합(regression sum of squares)

총제곱합의 분해

- 제곱합: $SST = SSE + SSR$

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

- d.f = number of information - limiting conditions
- $n - 1$ d.f:

  (n observed values $y_i - \bar{y}$) - (1 constraint $\sum(y_i - \bar{y}) = 0$)
- $n - 2$ d.f:

  (n observed values $y_i - \hat{y}_i$) - (2 d.f are lost $\hat{\beta}_0, \hat{\beta}_1$ to obtain $\hat{y}_i$)
- 1 d.f:

  (We know 2 parameters $\hat{\beta}_0, \hat{\beta}_1$) - (1 d.f is lost, $\sum(\hat{y}_i - \bar{y}) = 0$)

평균제곱(mean square)과 제곱합(sum of squares)

- 평균제곱(mean square): 대응하는 제곱합을 d.f 로 나누어 준 것.
  - 오차 평균제곱: $MSE = \frac{SSE}{n-2}$
  - 회귀 평균제곱: $MSR = \frac{SSR}{1}$

- Useful calculation formula

$$SST: \quad \sum_{i=1}^{n}(y_i - \bar{y})^2 = S_{yy}$$

$$SSR: \quad \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n}\{\hat{\beta}_1(x_i - \bar{x})\}^2 = \hat{\beta}_1^2 S_{xx} = \frac{S_{xy}^2}{S_{xx}}$$

$$SSE: \quad \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = S_{yy} - \hat{\beta}_1^2 S_{xx}$$

결정계수(coefficient of determination) R^2

- 결정계수: 총제곱합 SST 중에서 회귀제곱합 SSR이 차지하는 비율

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- 자료 전체의 변동을 나타내는 SST 중에서 회귀에 의하여 설명되는 SSR이 크면, 회귀모형이 관측결과를 잘 설명

- 결정계수 R^2은 표본상관계수의 제곱이다.

$$R^2 = \frac{SSR}{SST} = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \left\{ \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} \right\}^2 = r^2$$

- $0 \leq R^2 \leq 1$
  - $R^2 \approx 1$: 추정된 회귀식이 총 변동량의 많은 부분을 설명
  - $R^2 \approx 0$: 추정된 회귀식이 총 변동량을 적절하게 설명하지 못함

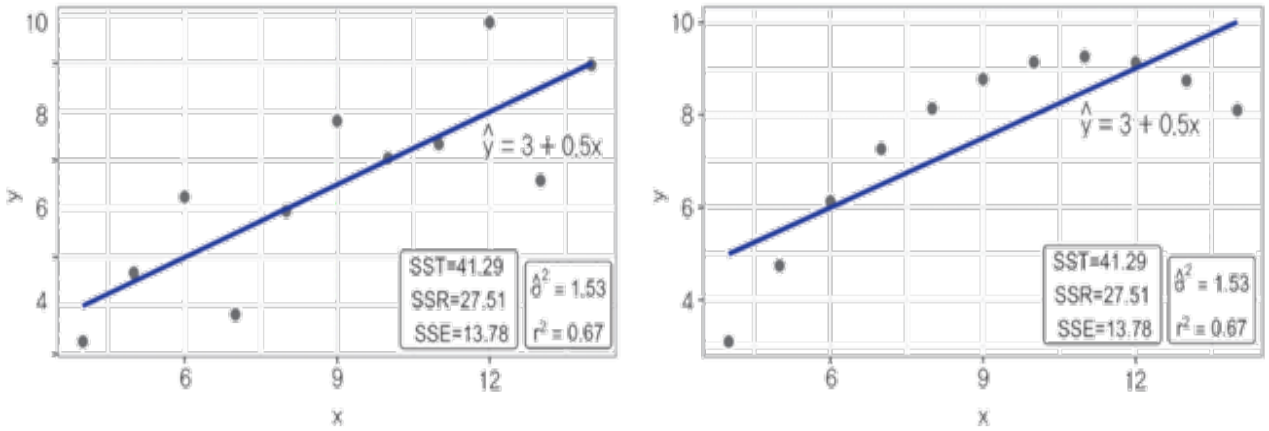
회귀분석시 주의점

- 다음의 표의 자료를 이용하여, 회귀분석을 한 결과를 비교해 보자.

표 8-3 모든 통계량이 거의 일치하는 두 조의 자료[10]

자료 A	x	10	8	13	9	11	14	6	4	12	7	5
	y	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68
자료 B	x	10	8	13	9	11	14	6	4	12	7	5
	y	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74

- (a), (b)는 SST, SSE, SSR, R^2 등의 통계량 값이 거의 일치한다.



(a) 자료 A의 최소제곱회귀직선과 통계량

(b) 자료 B의 최소제곱회귀직선과 통계량

그림 8-7 통계량은 같지만 관계가 다른 [표 8-3]의 자료

- 그러나 두 자료는 전혀 다른 관계를 나타내고 있다.(산점도)
- 단순히 통계량만에 의한 해석보다는 산점도 or 잔차의 검토가 필요하다.

단순회귀분석에서의 추론

- 다음의 단순선형회귀모형의 가정이 타당할 때

$$\begin{cases} Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \\ \epsilon_i \sim_{i.i.d} N(0, \sigma^2) \quad (i = 1, \cdots, n) \end{cases}$$

- 회귀직선이 의미가 있는지 검정하기 위해서 다음의 가설을 고려해 보자.

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

- 만일 $\beta_1 = 0$이라면 설명변수와 반응변수사이에 선형적인 관계가 없다는 것이 된다.
- 회귀직선이 유의하여 선형적인 연관성이 있다면, $\beta_1 \neq 0$일 것이다.

회귀직선의 유의성 F 검정

- 귀무가설 H_0가 사실일 때 다음의 F 검정통계량은 자유도가 $1,\ n-2$인 F분포를 따른다고 알려져 있다.

$$F = \frac{SSR/1}{SSE/(n-2)} \sim F(1, n-2)$$

pf) by Cochran's Theorem

$\frac{SSR}{\sigma^2}$, $\frac{SSE}{\sigma^2}$ are independently distributed chi-square random variables with $1,\ (n-2)$ degrees of freedom

$$\frac{\frac{SSR}{\sigma^2}/1}{\frac{SSE}{\sigma^2}/(n-2)} = \frac{SSR/1}{SSE/(n-2)} \sim F(1, n-2)$$

회귀직선의 유의성 F 검정

- 제곱합: $SST = SSE + SSR$
- 회귀직선이 유의하다면 자료의 산포를 나타내는 SST 중에서 SSR이 차지하는 비중이 커지고 SSE가 차지하는 비중은 작아질 것이다.

  즉, SSR/SSE이 커질수록 증거가 강해진다.

- 이때, F검정의 대립가설은 양측방향($H_1 : \beta_1 \neq 0$)이나, β_1의 값이 음수나 양수로 큰 값 모두 F값이 큰 값으로 나오므로 오른쪽단측 검정을 한다.

회귀직선의 유의성 F 검정

- 가설: $H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0$
- 검정통계량(under H_0):

$$F = \frac{SSR/1}{SSE/(n-2)} \sim F(1, n-2)$$

이러한 F 검정통계량의 관측값이 f이면 다음이 성립한다.

유의확률

$$P = P\{F \geq f\}, \quad F \sim F(1, n-2)$$

유의수준 α의 기각역

$$f \geq F_\alpha(1, n-2)$$

- 분산분석표(analysis of variance table)

요인	제곱합	자유도	평균제곱	F값	유의확률
회귀	SSR	1	MSR=SSR/1	f=MSR/MSE	$P\{F \geq f\}$
잔차	SSE	$n-2$	MSE=SSE/$(n-2)$		
계	SST	$n-1$			

단순선형회귀분석 예제

Example

예제 8.2 (13p)에서 단순선형회귀모형을 적용할 때, 오차분산 σ^2의 추정값, 결정계수를 구하고, 또한 회귀직선의 유의성 F 검정을 하여라.

sol) 표 8.1의 자료로부터 $S_{xy} = -21.169$, $S_{xx} = 24.649$, $S_{yy} = 18.469$
따라서, 잔차제곱합 SSE, 평균제곱오차 MSE는 다음과 같다.

$$SSE = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 18.469 - (-21.169)^2/24.649 = 0.289$$

$$MSE = \hat{\sigma}^2 = 0.289/10 = 0.0289$$

- 결정계수의 값은

$$R^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \frac{(-21.169)^2}{24.649 \cdot 18.469} = 0.9844$$

회귀직선이 전체 자료의 산포 중 98%를 설명해 준다는 뜻이고, 단순선형회귀모형에 의한 적합이 관측 결과를 잘 설명해주고 있다.

- 검정하고자 하는 가설은 다음과 같다.

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

각 제곱합을 계산하면

$$SST = 18.469, \;\; SSR = S_{xy}^2 / S_{xx} = (-21.169)^2 / 24.649 = 18.180$$

- 이로부터 평균제곱과 F 검정통계량의 관측값을 구하면

$$MSR = SSR/1 = 18.180, \;\; MSE = SSE/(12-2) = 0.0289$$

$$f = MSR/MSE = 18.180/0.0289 = 629.76$$

- F분포표에서 $F \sim F(1, 10)$일 때 $P(F \geq 629.76) < 0.001$ 이므로 회귀직선의 유의성에 대한 증거가 뚜렷하다고 할 수 있으며, 이러한 검정 과정을 분산분석표로 나타내면 다음과 같다.

요인	제곱합	자유도	평균제곱	F값	유의확률
회귀	$SSR = 18.180$	1	MSR=18.180	f=629.76	< 0.001
잔차	SSE=0.289	10	MSE=0.0289		
계	$SST = 18.469$	11			