

회귀분석이란 - 1 . pdf . .

일반통계학 제 8장 상관분석과 회귀분석

통계학 2016.1학기 정혜영

8-1 상관분석

1. 상관분석 VS 회귀분석

1.1 상관분석 이변량 정규분포 가정

- (1) 상관계수는 두 변수의 직선관계가 얼마나 강하고 어떤 방향인지를 나타냄
- (2) 상관분석은 두 변수의 상관계수를 분석함으로써, 두 변수 사이의 연관성을 분석
- (3) 표본상관계수를 이용하여 모상관계수에 대해 **추론**

1.2 회귀분석

- (1) 두 변수 사이의 함수관계($y = f(x)$)를 **분석**하여 한 변수값으로부터 다른 변수값에 대한 예측
- (2) 단순회귀분석
 - 두 변수 사이의 직선관계($y = ax + b$)를 모형으로 하여 분석
- (3) 중회귀분석
 - 두 개 이상의 변수가 한 변수에 영향을 줄 때,
한 변수를 여러 변수의 함수($y = a_0 + a_1x_1 + \dots + a_kx_k$)로 나타내어 분석

2. 모상관계수 및 표본상관계수

2.1 모상관계수

(1) (X, Y) 가 확률변수일 때, 모상관계수 ρ 는 다음과 같이 정의된다.

$$\rho = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}$$

(2) 성질

- 두 확률변수의 직선관계가 얼마나 강하고 어떤 방향인지를 나타내는 척도
- 모집단의 분포가 좌표평면에서 양의 방향이면 양의 값을, 음의 방향이면 음의 값을 가진다.
- $\text{Corr}(aX + b, cY + d) = \text{sign}(ac)\text{Corr}(X, Y)$

2.2 표본상관계수

(1) $(x_1, y_1), \dots, (x_n, y_n)$ 이 랜덤표본 $(X_1, Y_1), \dots, (X_n, Y_n)$ 의 관측값일때, 표본상관계수

r 은 다음과 같이 정의된다.

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

8-1 상관분석

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

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}, \quad S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n}$$

(2) 성질

$$-1 \leq r \leq 1$$

- r의 값은 관측값이 직선 관계에 가까울수록, -1(기울기가 음의 방향) 또는 1(기울기가 양의 방향)에 가깝게 주어진다.

- 표본상관계수의 값이 0에 가깝다고 해서 두 변수 간에 아무런 관계가 없음을 의미하지 않는다. 선형관계는 없다.

$$sd(r) = \sqrt{\frac{1-r^2}{n-2}} \quad T = \frac{r}{sd(r)}$$

3. ρ 에 관한 추론 (이변량 정규모집단 가정)

(1) 모집단의 분포가 이변량 정규분포인 경우에 귀무가설 $H_0: \rho = 0$ 에서 다음이 성립한다.

$$T = \sqrt{n-2} \frac{r}{\sqrt{1-r^2}} \sim t(n-2)$$

8-1 상관분석

(2) ρ 에 대한 가설검정

1. 가설

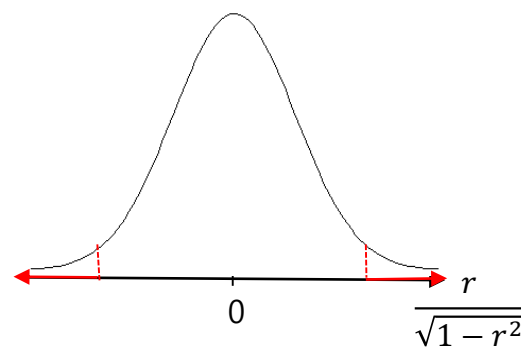
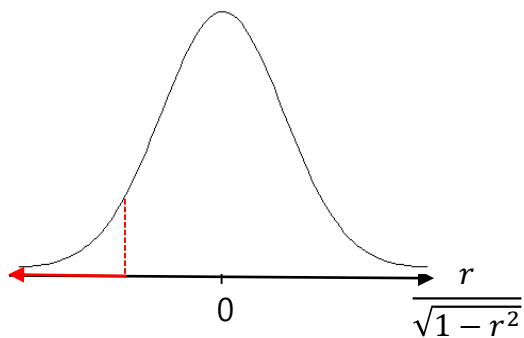
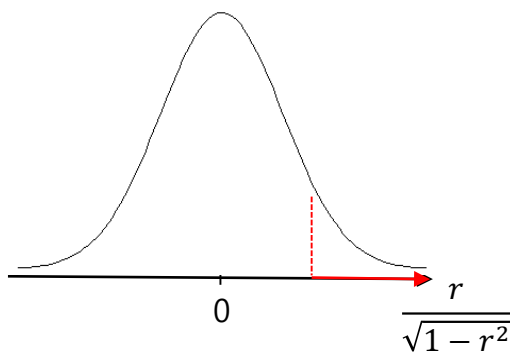
$$H_0: \rho = 0 \text{ vs } H_1: \rho > 0 \quad H_0: \rho = 0 \text{ vs } H_1: \rho < 0 \quad H_0: \rho = 0 \text{ vs } H_1: \rho \neq 0$$

2. 검정통계량

$$\frac{r}{\sqrt{1-r^2}} \sim t(n-2)$$

3. 기각역

$$H_0: \rho = 0 \text{ vs } H_1: \rho > 0 \quad H_0: \rho = 0 \text{ vs } H_1: \rho < 0 \quad H_0: \rho = 0 \text{ vs } H_1: \rho \neq 0$$



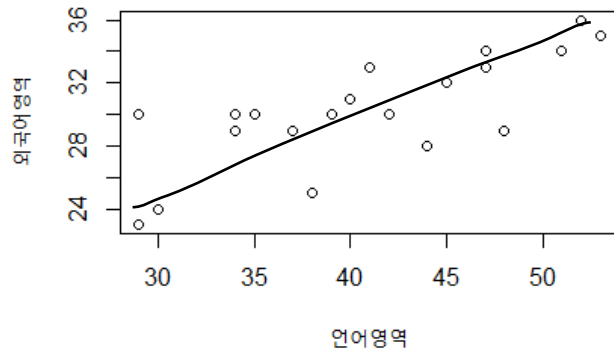
8-1 상관분석

(예제) 어느 고등학교 학생 중에서 랜덤하게 추출된 20명의 수학능력 모의시험에서 언어 영역과 외국어 영역의 점수이다.

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

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

언어영역과 외국어영역 성적간의 산점도



표본상관계수 $r = 0.7567$ 이고 산점도 상에서도 선형관계가 보이므로 이 표본의 결과를 가지고 실제 그 고등학교 학생 전체의 언어영역과 외국어 영역 성적사이의 상관관계가 있는지 유의수준 5%에서 검정하여라.

1. 가설 $H_0: \rho = 0$ vs $H_1: \rho \neq 0$

2. 검정통계량 계산 $t = \sqrt{n-2} \frac{r}{\sqrt{1-r^2}} = 4.915$

3. 기각역 $t_{0.025}(18) = 2.101 < 4.915$

유의확률 $2P(t > 4.915) = 0.0022 < 0.05$ 이므로 귀무가설을 기각할 만한 증거가 된다. 따라서 두 성적간에는 유의수준 5%에서 상관관계가 있다는 뚜렷한 증거가 있다.

8-2 단순회귀분석의 모형과 적합

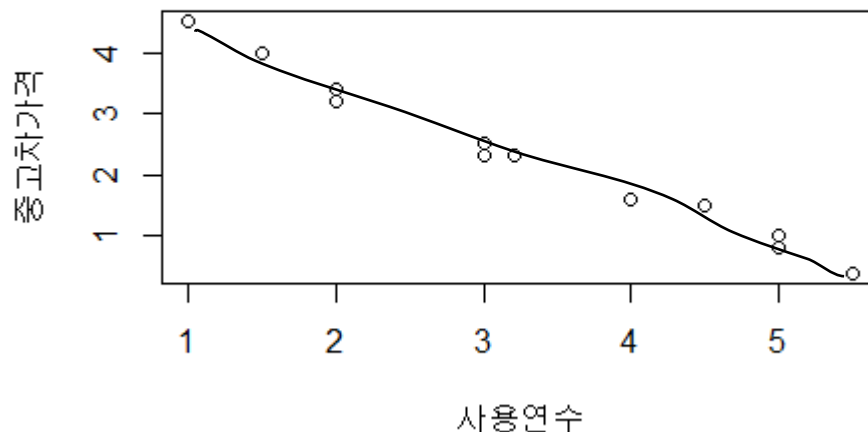
싶은 알고싶다.

1. 단순선형회귀모형

1.1 예 - 사용연수에 따른 중고차 가격

사용연수	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	4.5	4.0	3.2	3.4	2.5	2.3	2.3	1.6	1.5	1.0	0.8	0.4

산점도



$\pm \varepsilon$ (오차)

$$\Rightarrow (\text{중고차 가격}) = \alpha + \beta \times (\text{사용연수})$$

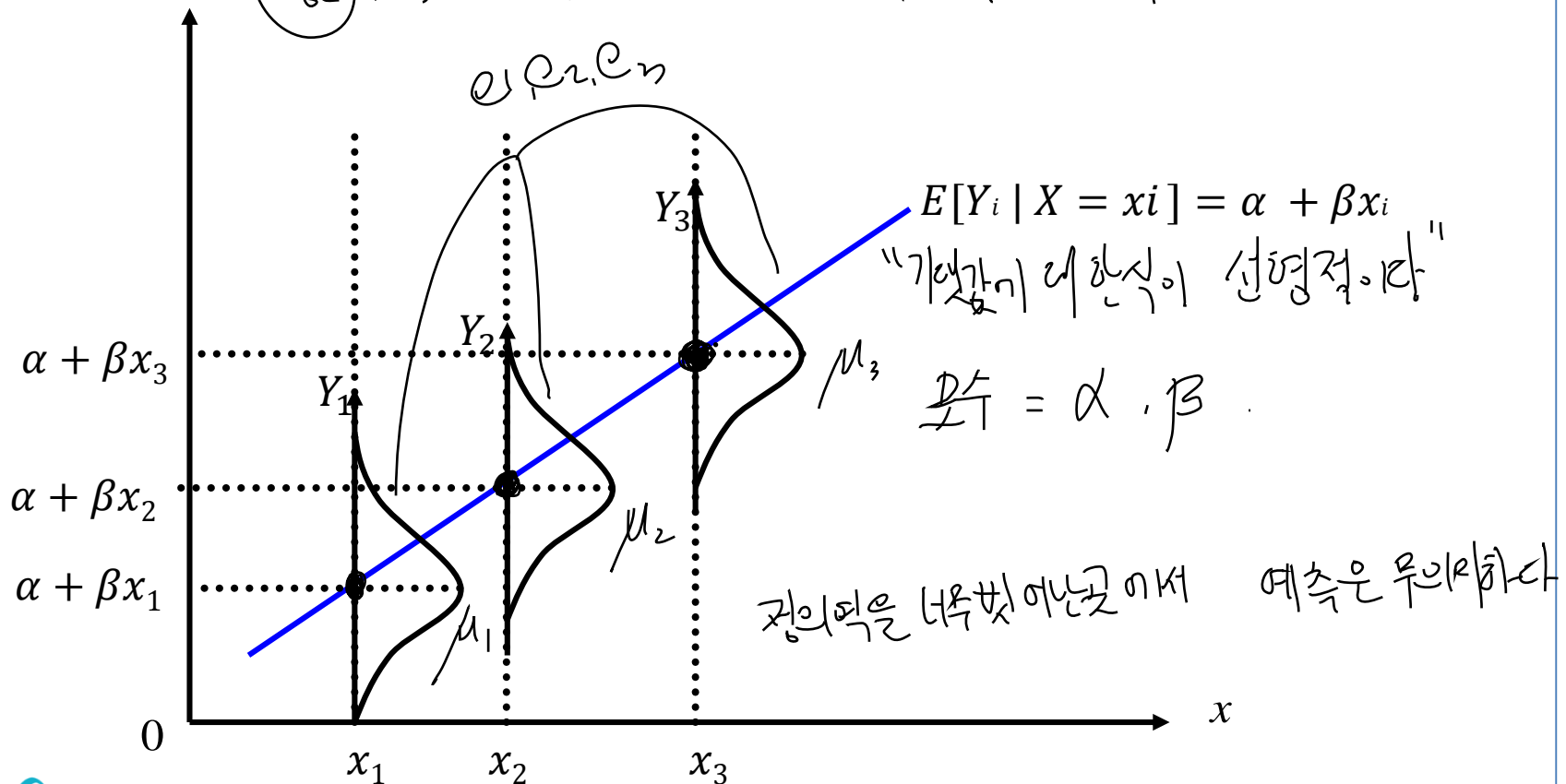
8-2 단순회귀분석의 모형과 적합

1.2 단순 선형 회귀 모형 : $Y_i = \alpha + \beta x_i + e_i$

e_i 는 오차

- 오차를 나타내는 e_i 가 등분산을 가지는 확률변수이므로 반응변수 Y_i 도 확률변수가 된다. 설명변수 x_i 에 대응하는 반응변수의 값이 $\alpha + \beta x_i$ 주위에 나타날 때, 반응변수의 값 y_i 는 확률변수 Y_i 의 관측값으로 생각할 수 있다.

"평균값은 선형성을 가진다고 할 수 있다"가 가정.



8-2 단순회귀분석의 모형과 적합

- 단순선형회귀모형의 기본 가정

(1) $E(e_i)=0$, 즉, $E[Y_i | X = x_i] = \alpha + \beta x_i$ (선형성)

(2) $Var(e_i)=Var(e_i)=\dots=Var(e_n)=\sigma^2 > 0$ (등분산성)

(3) e_1, e_2, \dots, e_n 은 서로 독립 (독립성)

((3)에 의해 Y_1, Y_2, \dots, Y_n 도 서로 독립적으로 관측된다고 가정)

- 설명변수에 따른 반응변수의 평균값을 나타내는 직선 $y = \alpha + \beta x$ 를 모회귀직선, α 와 β 를 모회귀계수, σ^2 을 오차분산이라고 한다. 이제 모회귀직선을 추정하기 위해 최소제곱법을 사용하자.

- 최소제곱법 (method of least squares)

$Q(\alpha, \beta) = \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2$ 을 최소로 하는 α, β 를 $\hat{\alpha}, \hat{\beta}$ 이라 표시하고 α, β 의 최소제곱 추정량(least squares estimator)이라고 부른다.

$$Q(\alpha, \beta) = \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2 \quad \alpha, \beta \text{ ଅନୁକ୍ରମେ } \frac{\partial}{\partial \alpha}, \frac{\partial}{\partial \beta}$$

05/11/2016

$$\left(\frac{\partial Q(\alpha, \beta)}{\partial \alpha} = \sum_{i=1}^n 2(y_i - \alpha - \beta x_i)(-1) = 0 \right.$$

$$\left. \frac{\partial Q(\alpha, \beta)}{\partial \beta} = \sum_{i=1}^n 2(y_i - \alpha - \beta x_i)(-x_i) = 0 \right.$$

$$\left(-2n\bar{y} + 2n\alpha + 2\beta n\bar{x} = 0 \right.$$

$$\left. 2n\bar{xy} + 2\alpha n\bar{x} + 2\beta n\bar{x}^2 = 0 \right.$$

$$\left(\beta\bar{x} + \alpha - \bar{y} = 0 \Rightarrow \alpha = \bar{y} - \beta\bar{x} \right.$$

$$\left. \beta\bar{x}^2 + \alpha\bar{x} - \bar{xy} = 0 \right.$$

$$\left[\begin{array}{l} \beta\bar{x}^2 + (\bar{y} - \beta\bar{x})\bar{x} - \bar{xy} = 0 \\ \beta(\bar{x}^2 - (\bar{x})^2) + (\bar{x}\bar{y} - \bar{xy}) = 0 \\ \beta = \frac{\bar{xy} - \bar{x} \cdot \bar{y}}{\bar{x}^2 - \bar{x}^2} = \frac{S_{xy}/n}{S_{xx}/n} = \frac{S_{xy}}{S_{xx}} \end{array} \right]$$

$$\beta = \frac{S_{xy}}{S_{xx}}, \quad \alpha = \bar{y} - \beta\bar{x}$$

8-2 단순회귀분석의 모형과 적합

$$\sum y_i - \hat{y} = 0$$

$$\sum x_i (y_i - \hat{y}) = 0$$

잔차

- 정규방정식 (normal equation)

$$\begin{cases} -2 \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta} x_i) = 0 \\ -2 \sum_{i=1}^n x_i (y_i - \hat{\alpha} - \hat{\beta} x_i) = 0 \end{cases} \Rightarrow \begin{cases} n\hat{\alpha} + \sum_{i=1}^n x_i \hat{\beta} = \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i \hat{\alpha} + \sum_{i=1}^n x_i^2 \hat{\beta} = \sum_{i=1}^n x_i y_i \end{cases}$$

위 정규방정식에 의해 α, β 의 최소제곱추정량 $\hat{\alpha}, \hat{\beta}$ 를 다음과 같이 얻게 된다.

- 최소제곱 추정값과 최소제곱회귀직선

평균 (\bar{x}, \bar{y}) 를 지낸다.

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

최소제곱회귀직선 $\hat{y} = E(\widehat{Y_i} | x_i) = \hat{\alpha} + \hat{\beta}x = \bar{y} + \hat{\beta}(x - \bar{x})$

y의 변동성 $\rightarrow \alpha, \beta$ 의 변동성

- 간단 계산 공식

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

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



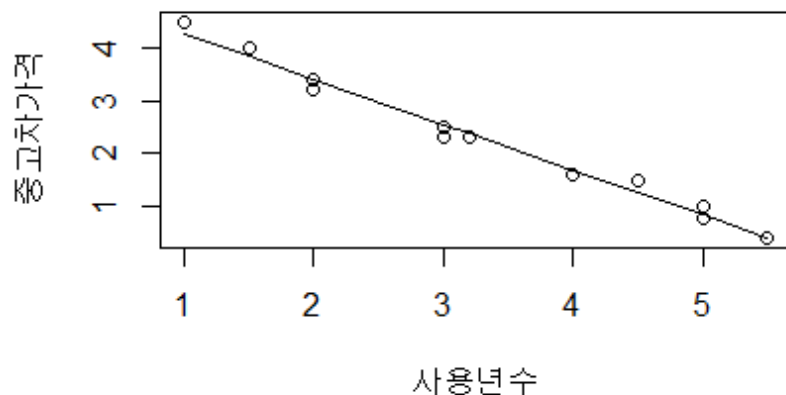
8-2 단순회귀분석의 모형과 적합

(예제)

사용연수(x), 중고차 가격(Y)로 하여 단순선형회귀모형을 가정하자. 아래의 표를 이용하여 $E(Y|x) = \alpha + \beta x$ 에 대한 최소제곱추정값 $\hat{\alpha}, \hat{\beta}$ 을 구하고 최소제곱회귀직선을 구하여라.

사용연수	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	4.5	4.0	3.2	3.4	2.5	2.3	2.3	1.6	1.5	1.0	0.8	0.4

$$\hat{y} = E(\widehat{Y_i} | x_i) = 5.133 - 0.859x$$



8-2 단순회귀분석의 모형과 적합

- 잔차 (residual) e_i 의 추정값으로 쓴다.

$$\hat{e}_i = \text{실제 관측값}(y_i) - \text{추측값}(\hat{y}_i) \quad (i = 1, 2, \dots, n)$$

이 값이 작을수록 최소제곱회귀직선에 의한 적합이 실제 관측 결과를 잘 설명해주는 것이다.

- 잔차제곱합(residual sum of squares or error sum of squares)

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

- 평균제곱오차 (mean squared error) - 오차분산의 추정

$$\hat{\sigma}^2 = \text{MSE} = \frac{SSE}{n - 2}$$

$$E(\hat{\sigma}^2) = \sigma^2 \rightarrow e_i \text{의 분산}$$

(예제) 중고차 가격 예제에서 오차분산 σ^2 의 추정값을 구하여라.

8-2 단순회귀분석의 모형과 적합

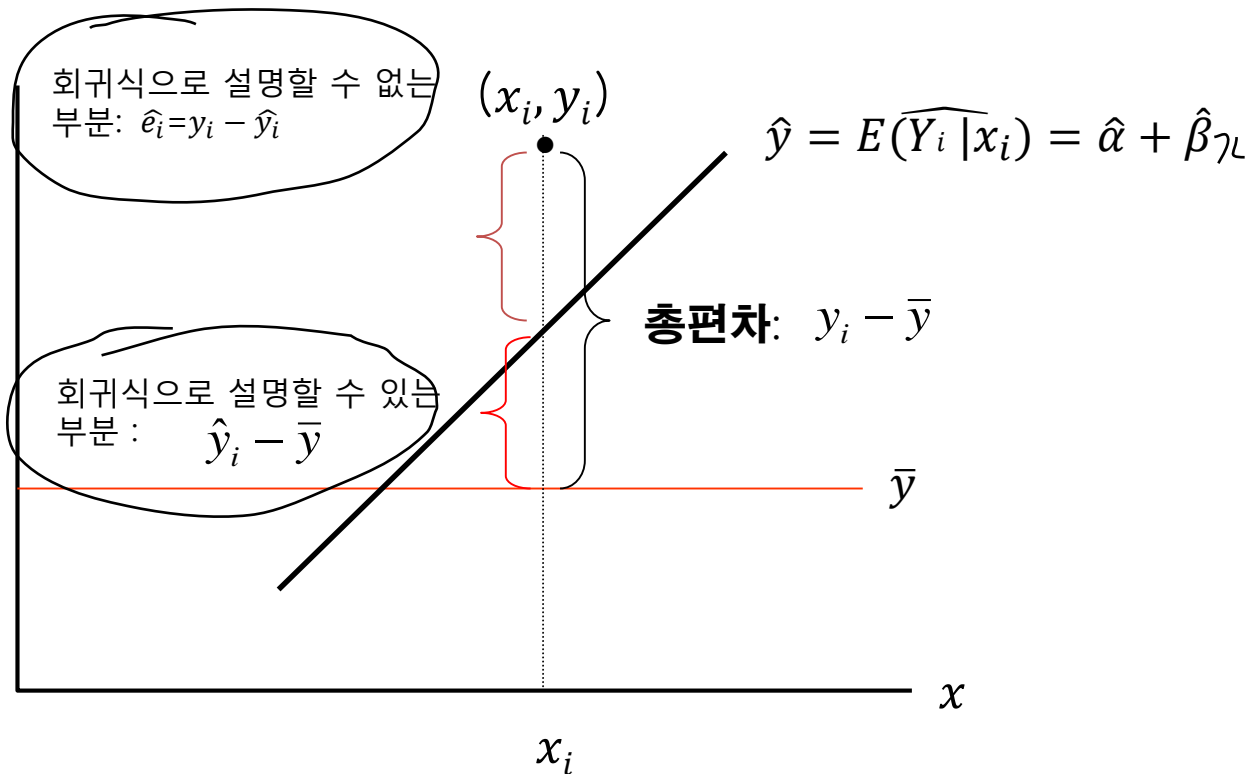
- 총편차 $y_i - \bar{y}$ 의 분해

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

$y_i - \hat{y}_i$: 오차항에 기인하는 편차(잔차)

$\hat{y}_i - \bar{y}$: 회귀직선에 기인하는 편차

회귀선 주변에 점이 많이 있을 바람.



8-2 단순회귀분석의 모형과 적합

- 총제곱합의 분해

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

$$\underbrace{\text{총제곱합}}_{\text{SST}(n-1)} = \underbrace{\text{잔차제곱합}}_{\text{SSE}(n-2)} + \underbrace{\text{회귀제곱합}}_{\text{SSR}(1)}$$

※ 괄호안은 자유도임.

- 결정계수(coefficient of determination)

자료 전체의 흩어진 정도를 나타내는 SST 중에서 회귀선에 설명되는 부분인 SSR이 차지하는 비중이 크면 회귀모형이 관측결과를 잘 설명해주는 것이다. 따라서

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^n (\hat{\alpha} + \hat{\beta}x - \bar{y})^2 = \sum_{i=1}^n (\bar{y} - \hat{\beta}\bar{x} + \hat{\beta}x - \bar{y})^2 = \sum_{i=1}^n \{\hat{\beta}(x - \bar{x})\}^2 = (S_{xy})^2 / S_{xx}$$

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

$$r^2 = \frac{SSR}{SST} = \frac{(S_{xy})^2}{S_{xx}S_{yy}} = \left(\frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} \right)^2 : \text{표본상관계수의 제곱}$$

따라서, 결정계수가 1에 가까울수록 산점도에서 점들이 직선 주위에 밀집되어 나타나게 되어 회귀선에 의한 설명이 잘 됨을 뜻한다. ✎

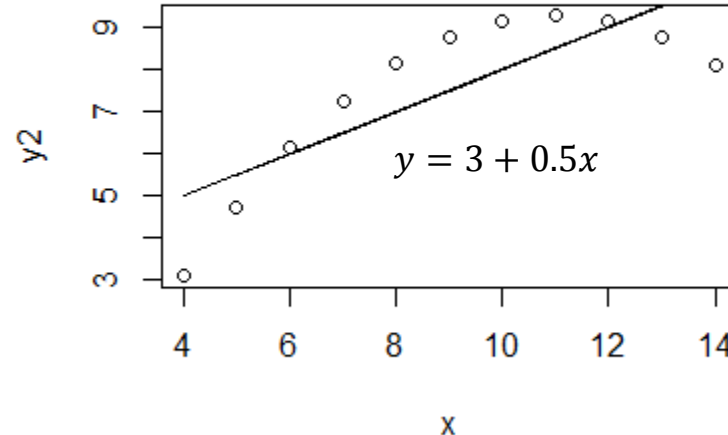
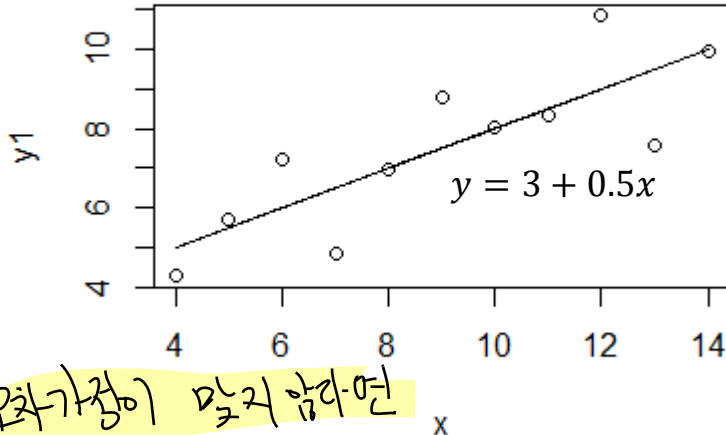
8-2 단순회귀분석의 모형과 적합

(예제) 중고차 가격에 단순회귀모형을 적용할 때, 결정계수를 구하고 이를 해석하여라.

- 회귀분석에서 통계량의 계산값에 의한 결론의 문제점

상관계수의 단점을 갖고 있다.

통계량으로만 판단하면 안된다.



오차가 많이 많지 않으면
회귀분석이 의미가 없다.

SST=41.27
SSR=27.51
SSE=13.76

$\hat{\sigma}^2 = 1.53$
 $r^2 = 0.67$

산점도가 다른 자료임에도 통계량의 값이 모두 일치! 산점도 혹은 잔차의 검토가 회귀분석의 과정에서 매우 중요함!!

8-3 단순회귀분석에서의 추론

1. 단순 선형 회귀 모형 : $Y_i = \alpha + \beta x_i + e_i, \quad i = 1, 2, \dots, n$

❖ 오차항의 가정이 아래와 같을 때, 모 회귀계수에 대한 추론이 가능

- (1) 선형성
- (2) 등분산성
- (3) 독립성

(4) 정규성 ($e_i \sim N(0, \sigma^2)$) \longrightarrow 추론 가능 \Rightarrow 추론 가능

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

$$SST = SSE + SSR$$

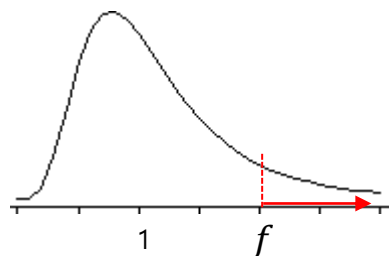
$$\sim \chi^2(n-1) \quad \sim \chi^2(n-2) \quad \sim \chi^2(1)$$

2. 회귀직선의 유의성 검정

(1) 가설 - $H_0: \beta = 0$ vs $H_1: \beta \neq 0$

회귀평의비 X 의비 0
 (2) 귀무가설 H_0 가 사실일 때, $F = \frac{SSR/1}{SSE/(n-2)} \sim F(1, n-2)$

(3) 기각역



❖ 회귀직선이 유의하면 총제곱합 중에서 회귀제곱합이 차지하는 비중은 커지고 잔차제곱합이 차지하는 비중은 작아질 것이다.

$SSR \gg SSE$ 이면 기각!!

오차가 설명하는 부분보다

회귀선이 설명하는 부분이 많다면

8-3 단순회귀분석에서의 추론

3. 분산분석표

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	$n - 1$			

(예제) 자동차의 중고차가격 예제에서 단순회귀모형을 적용할 때, 분산분석표를 만들고 회귀직선의 유의성 검정을 하여라. (단, $S_{xy} = -21.169$, $S_{xx} = 24.649$, $S_{yy} = 18.469$. $f_{0.05}(1,10) = 4.9646$, 유의확률 $= 2.311e-10$)

회귀	SSR	1	$(S_{xy})^2 / S_{xx} = 18.18$	$MSR = 18.18$
잔차	SSE	$n - 2$	$S_{yy} - (S_{xy})^2 / S_{xx} = 0.288$	$MSE = 0.024$
계	SST	$n - 1$		$f = 757.5$

8-3 단순회귀분석에서의 추론

x_i 는 숫자다.

4. 모회귀계수 β 에 관한 추론

$$E(\hat{\beta}) = E\left(\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) = E\left(\frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) = \frac{\sum_{i=1}^n (x_i - \bar{x})E(y_i)}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(\alpha + \beta x_i)}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta$$

($\sum_{i=1}^n (x_i - \bar{x}) = 0$, $\sum_{i=1}^n (y_i - \bar{y}) = 0$) 이므로

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n (x_i - \bar{x})y_i, \quad \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (x_i - \bar{x})x_i$$

$$\text{Var}(\hat{\beta}) = \text{Var}\left(\frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{(\sum_{i=1}^n (x_i - \bar{x})^2)^2} \text{Var}(y_i) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma^2}{S_{xx}}$$

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right) (\because y_i \text{의 정규성})$$

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

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

$\hat{\sigma}^2 = \frac{SSE}{n-2} = MSE$

$$\hat{\beta} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

계량기

8-3 단순회귀분석에서의 추론

- β 에 대한 $100(1-\alpha)\%$ 신뢰구간

$$\left(\hat{\beta} - t_{\alpha/2}^{(n-2)} \frac{\hat{\sigma}}{\sqrt{S_{xx}}} < \beta < \hat{\beta} + t_{\alpha/2}^{(n-2)} \frac{\hat{\sigma}}{\sqrt{S_{xx}}} \right)$$

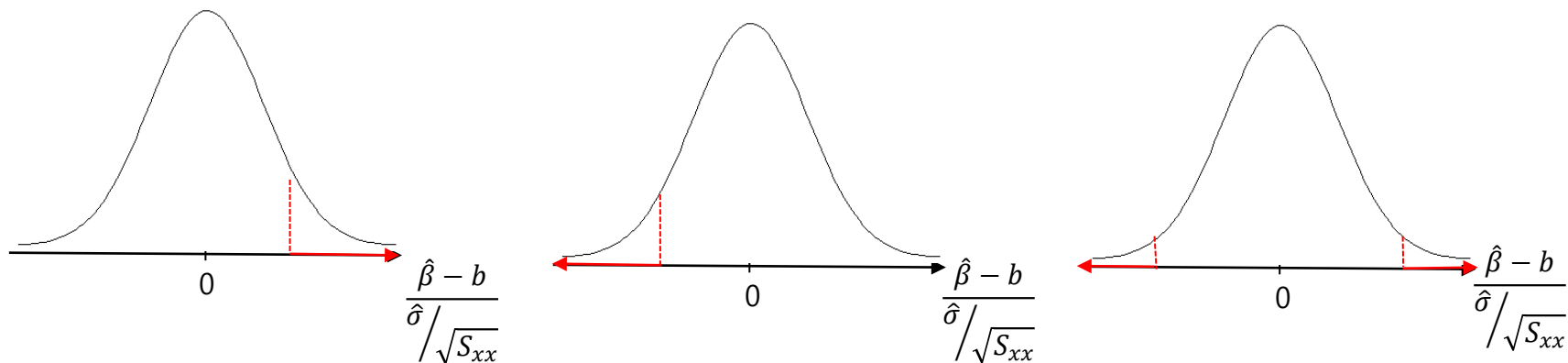
- 가설검정

$$1. H_0: \beta = b \text{ vs } H_1: \beta > b \quad H_0: \beta = b \text{ vs } H_1: \beta < b \quad H_0: \beta = b \text{ vs } H_1: \beta \neq b$$

$$2. \text{검정통계량 } \frac{\hat{\beta} - b}{\hat{\sigma} / \sqrt{S_{xx}}} \sim t(n-2)$$

- 3. 기각역

$$1. H_0: \beta = b \text{ vs } H_1: \beta > b \quad H_0: \beta = b \text{ vs } H_1: \beta < b \quad H_0: \beta = b \text{ vs } H_1: \beta \neq b$$



8-3 단순회귀분석에서의 추론

(예제) 자동차의 중고차가격 예제에서 단순회귀모형을 적용할 때, 다음 추론을 하여라.

(1) 회귀직선의 유의성을 t검정으로 유의수준 1%에서 검정하고 분산분석표에 의한 F검정과
의 관계를 설명하여라. ($t_{0.005}(10) = 3.169$)

1) 가설 $H_0: \beta = 0$, $H_1: \beta \neq 0$

2) 유의수준 $\alpha = 0.05$

3) 검정통계량

$$\frac{\hat{\beta} - \beta}{\hat{\sigma} / \sqrt{S_{xx}}} \sim t(n-1)$$

4) 계산 $\frac{-0.859}{0.154 / \sqrt{24.649}} = -21.693$

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

8-3 단순회귀분석에서의 추론

(2) 자동차의 가격이 사용년수에 따라 연평균 감소액이 80만원을 초과하는지 유의수준 5%에서 검정하여라. ($t_{0.05}(10) = 1.812$)

$$H_0: \beta = -80$$

$$\beta = -80$$

$$H_1: \beta > 80$$

$$\beta < 80$$

??
~

(3) 모회귀계수 β 의 95% 신뢰구간을 구하고 그 의미를 해석하여라.

8-3 단순회귀분석에서의 추론

$$\alpha + \beta x \rightarrow \hat{\alpha} + \hat{\beta} x \quad (x \text{는 주어진 값})$$

추정치

- 평균반응 $E[Y | X = x] = \alpha + \beta x$ 에 관한 추론

$$E[\hat{\alpha} + \hat{\beta}x] = \alpha + \beta x, \quad \text{Var}(\hat{\alpha} + \hat{\beta}x) = \left\{ \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} \right\} \hat{\sigma}^2$$

$\hat{\alpha} + \hat{\beta}x$ 은 정규분포를 따른다. 오차분산을 추정하게 되면 다음과 같은 t 포를 따르게 된다.

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

$$E[\hat{y}|x] = \alpha + \beta x$$

$$\sigma \rightarrow \hat{\sigma}$$

$\Rightarrow t$ 분포

(1) $\alpha + \beta x$ 에 관한 $100(1 - \alpha)\%$ 신뢰구간

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

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

(2) $H_0: \alpha + \beta x = \mu_0$ 에 관한 검정통계량 $\frac{\hat{\alpha} + \hat{\beta}x - \mu_0}{\sqrt{\left(\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} \right) \hat{\sigma}^2}} \sim t(n-2)$ $\frac{E[\hat{y}|x] - \mu_0}{\text{표준 오차}}$

8-3 단순회귀분석에서의 추론

- 모회귀직선 α 에 관한 추론은 $\alpha + \beta x$ 에 관한 추론에서 $x = 0$ 인 경우에 해당하므로

$$\frac{\hat{\alpha} - \alpha}{\sqrt{\left(\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}\right)\hat{\sigma}^2}} \sim t(n-2)$$

(1) α 에 관한 $100(1 - \alpha)\%$ 신뢰구간

$$\hat{\alpha} - t_{\alpha/2}^{(n-2)} \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\hat{\sigma}^2} < \alpha < \hat{\alpha} + t_{\alpha/2}^{(n-2)} \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\hat{\sigma}^2}$$

(2) $H_0: \alpha = a$ 에 관한 검정통계량 $\frac{\hat{\alpha} - a}{\sqrt{\left(\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}\right)\hat{\sigma}^2}} \sim t(n-2)$

8-3 단순회귀분석에서의 추론

(예제) 비료의 양과 수확량에 대한 통계치들이 아래와 같이 주어져 있다.

단, $\bar{y} = 120, \bar{x} = 10, n = 40$, 단, $S_{xy} = 81, S_{xx} = 81, S_{yy} = 100$

(1) 최소제곱회귀 직선식을 써라.

$$110 + 1x$$

$$\hat{\beta} = S_{xy} / S_{xx} = 1, \quad \hat{\alpha} = \bar{y} - \hat{\beta} \bar{x} = 110$$

(2) 오차분산 추정값을 써라. 결정계수를 구하고 해석하라.

$$\hat{\sigma}^2 = \frac{SSE}{n-2} = 0.5$$

$$r^2 = \frac{SSR}{SST}$$

$$= 0.81$$

(3) 분산분석표를 만들어라.

회귀선이 81%를 설명

8-3 단순회귀분석에서의 추론

$$110 + x$$

$$E(Y|X=x) = 110 + x$$

(4) 비료의 양 $x=12$ 일 때 평균 수확량에 대한 95% 신뢰구간을 구하여라. ($t_{0.025}(38) = 2.204$)

$$\left(110 + 12 \pm 2.204 \cdot \sqrt{\left(\frac{1}{40} + \frac{2^2}{81} \right) 0.5^2} \right)$$

(5) 회귀직선의 유의성에 대한 가설을 세우고 유의수준 5%에서 t-검정 하여라.

$$H_0: \beta = 0 \quad H_1: \beta \neq 0 \quad (t)$$

(6) 비료의 양 $x=15$ 일 때 평균 수확량이 130인지 아닌지에 대해 유의수준 5%에서 검정하여라.

$$H_0: \alpha + \beta \cdot 15 = 130$$

$$H_1: \text{---}$$

8-3 단순회귀분석에서의 추론

(예제)

X	2	3	4	5	6
Y	4	4	6	6	10

단, $\sum X_i Y_i = 134, \sum X_i = 20, \sum Y_i = 30, \sum X_i^2 = 90, \sum Y_i^2 = 204$

$$S_{xy} = 14, S_{xx} = 10, S_{yy} = 24$$

(1) SSR, SSE, SST, 오차항의 분산 추정치(MSE)를 구하여라.

$$S_{yy}^2 / S_{xx} = 19.6$$

$$SST - SSR = 4.4$$

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-2} = 1.46$$

(2) 결정계수와 상관계수를 구하여라.

$$r^2 \rightarrow r$$

$$\frac{SSR}{SST} = \frac{19.6}{24} = 0.816$$

$$r = 0.90$$

(3) 분산분석표를 작성하여라.

SSR

1

SSE

n-2

SST

n-1

8-4 단순회귀분석에서의 잔차분석

1. 단순회귀분석 적용의 순서

- (1) 산점도를 이용하여 직선관계 확인
- (2) 단순회귀모형 적합 및 잔차분석
- (3) 잔차분석을 통과하면, 신뢰구간 및 검정과 같은 추론 및 예측 시행

2. 단순선형회귀모형 $Y_i = \alpha + \beta x_i + e_i$, $i = 1, 2, \dots, n$ 의 기본 가정

- (1) $E(e_i)=0$, 즉, $E[Y_i | X = x_i] = \alpha + \beta x_i$ (선형성)
- (2) $Var(e_i)=Var(e_1)=\dots=Var(e_n)=\sigma^2 > 0$ (등분산성)
- (3) e_1, e_2, \dots, e_n 은 서로 독립 (독립성)
- (4) $e_i \sim N(0, \sigma^2)$ (정규성)

→ 추론

❖ 잔차 $\hat{e}_i = y_i - \hat{y}_i$ 을 이용하여 위의 가정을 검토해야만 회귀분석에 대한 추론이 의미가 있게 된다.

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

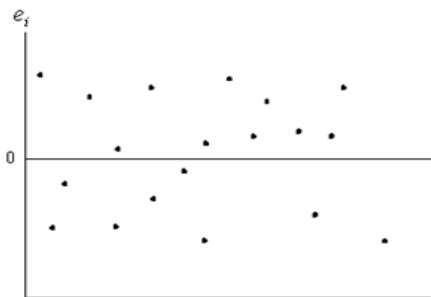
8-4 단순회귀분석에서의 잔차분석

- 잔차도 (residual plot) : 설명변수와 스튜던트화 잔차 ($\frac{\hat{e}_i}{sd(\hat{e}_i)}$, $\hat{e}_i = y_i - \hat{y}_i$)를 산점도로 나타낸 것으로 잔차도가 다음과 같은 성질을 갖고 있다면 단순선형회귀모형 적용이 타당하다고 간주한다.

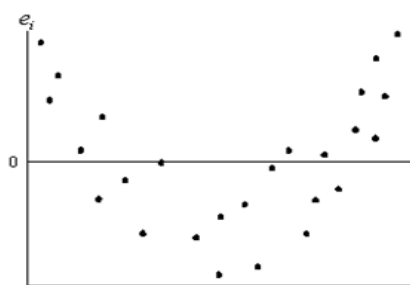
- ① 대략 0에 관하여 **대칭적**으로 나타나고,
- ② 설명변수의 값에 따른 **잔차의 산포가 크게 다르지 않고**,
- ③ 점들이 특정한 형식을 가지고 나타남이 없으며,
- ④ **거의 모든 점이 (-2, 2)의 범위 내에** 나타나야 한다.

→ 약 95%

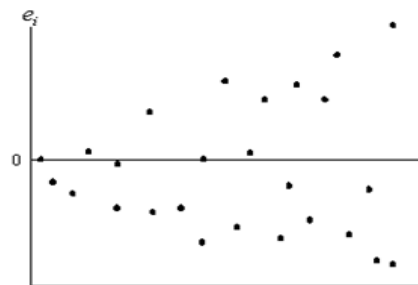
* 결정계수가 높게 나타난다 할지라도 잔차의 가정에 어긋난다면 잘못된 분석이므로 가정을 만족시킨 후 다시 분석해야 한다.



정상



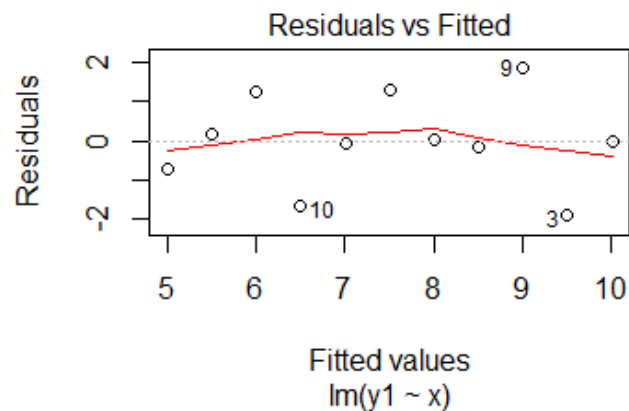
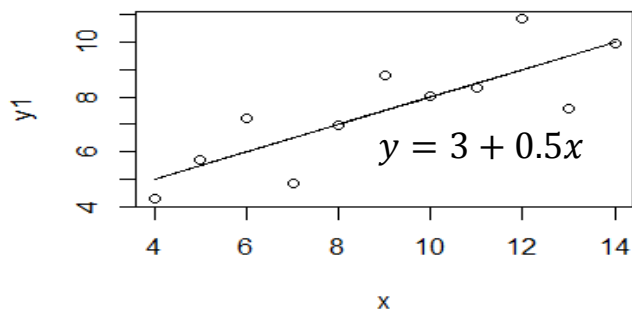
비선형관계



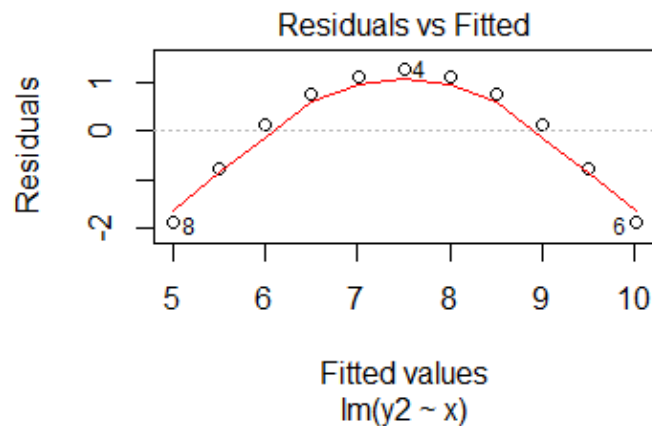
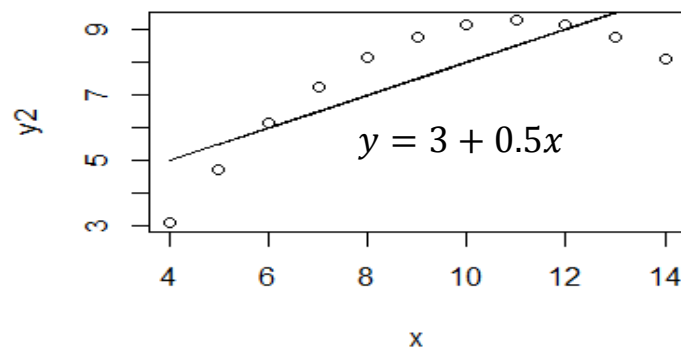
분산증가

8-4 단순회귀분석에서의 잔차분석

(예제) 동일한 선형회귀모형 결과에 대한 잔차분석



잔차에 특정한 패턴이 없음. 따라서 선형회귀모형에 대한 해석은 타당함.



잔차에 특정한 패턴이 보이므로 선형회귀모형에 대한 해석은 적절치 못함.

8-5 중회귀분석

- 중회귀모형 : 반응변수의 변화를 설명하기 위해 **두개 이상의 설명변수**를 고려하는 모형

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + e_i, e_i \sim N(0, \sigma^2) \text{ 이고 서로 독립 } (i = 1, 2, \dots, n)$$

- 회귀계수의 추정을 위해 최소제곱방법 사용
- 총제곱합의 분해

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2, \text{ 단, } \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_k x_{ki}$$

$$\begin{array}{ccccc} \text{총제곱합} & = & \text{잔차제곱합} & + & \text{회귀제곱합} \\ SST(n-1) & & SSE(n-k-1) & & SSR(k) \end{array}$$

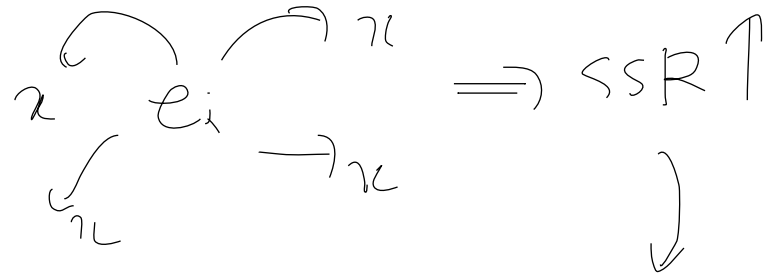
- 평균제곱오차 (mean squared error) - 오차분산의 추정

$$\hat{\sigma}^2 = \text{MSE} = \frac{SSE}{n - k - 1}$$

$$\text{결정계수 } r^2 = \frac{SSR}{SST}$$

- 수정된 결정계수(adjusted R-square) : 결정계수는 설명변수가 많아지면 커지는 경향이 있다. 따라서 결정계수를 독립변수의 수에 따라서 수정할 필요가 있다

$$R_{adj}^2 = 1 - \frac{n-1}{n-k-1} \frac{SSE}{SST}$$



8-5 중회귀분석

- 분산분석표

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

- 회귀모형의 유의성 검정 (F검정)

(1)가설 - $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$ vs $H_1: \text{not } H_0$

(2)귀무가설 H_0 가 사실일 때, $F = \frac{\frac{SSR}{k}}{\frac{SSE}{(n-k-1)}} \sim F(k, n - k - 1)$

- 회귀모형의 유의성 검정 (t검정)

각 계수 β_i 에 대한 유의성을 개별적으로 검정

8-5 중회귀분석

- 중회귀분석 결과

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.114e+02	6.297e+01	1.768	0.085979 .
Brain	2.060e+00	5.634e-01	3.657	0.000856 ***
Height	-2.732e+00	1.229e+00	-2.222	0.033034 *
Weight	5.599e-04	1.971e-01	0.003	0.997750

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.79 on 34 degrees of freedom

Multiple R-squared: 0.2949, Adjusted R-squared: 0.2327

F-statistic: 4.741 on 3 and 34 DF, p-value: 0.007215

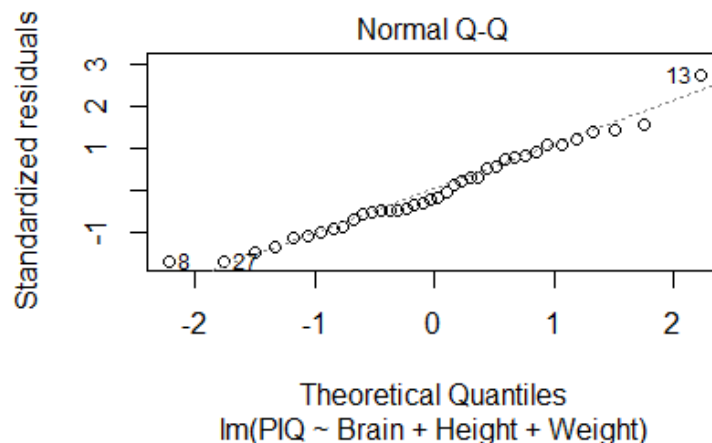
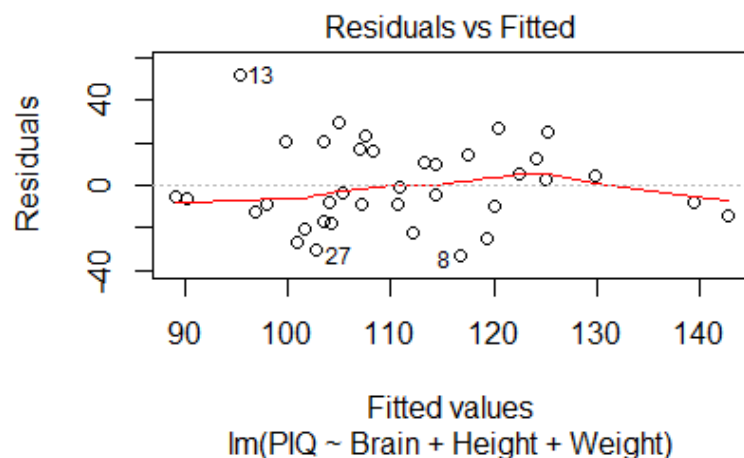
-모형의 유의성 : F검정 결과 유의수준 0.05보다 0.007215는 더 희박하므로 귀무가설 $H_0: \beta_1 = \beta_2 = \beta_3 = 0$ 을 기각할 만한 증거가 된다. 따라서 추정된 모형은 적합하다.

-추정된 회귀모형 : $\widehat{PIQ} = 111.4 + 2.06 \text{ Brain} - 2.732 \text{ Height} + 0.0005599 \text{ Weight}$

-각 추정된 계수에 대한 해석 : 다른 변수가 고정된 상태에서 키가 1단위 증가함에 따라 평균IQ는 2.732만큼 감소한다.

8-5 중회귀분석

- 잔차분석 결과



적합된 모형의 잔차도를 확인해본 결과 잔차의 값이 매우 큰 관측치가 몇 개 존재하기는 하지만 특별한 패턴이 관측되지는 않았다. 그리고 정규 분위수 그래프에서도 정규분포를 벗어난다는 뚜렷한 증거는 발견되지 않았다. 따라서 주어진 자료에 대한 중회귀모형의 적용은 타당함을 알 수 있고, 적용된 모형을 통한 추론은 의미가 있다고 할 수 있다.

8-5 중회귀분석

(예제) A회사 대리점 중 $n = 10$ 개를 뽑아 월매출액과 관할구역의 인구수, 가구당 월평균 수입을 조사한 표가 아래와 같다. 회귀모형의 적합성을 유의수준 5%에서 검정하여라.

i	1	2	3	4	5	6	7	8	9	10
월매출액(y_i)	2.0	1.3	2.4	1.5	0.6	2.0	1.0	2.0	1.3	0.9
인구수(x_{1i})	3.0	1.1	3.5	2.5	0.6	2.8	1.3	3.3	2.0	1.0
월평균수입(x_{2i})	3.2	3.0	3.6	2.6	1.9	3.5	2.1	3.3	2.8	2.3

추정된 회귀식 : $\hat{y} = -0.4517 + 0.3067x_1 + 0.4584x_2$

요인	제곱합	자유도	제곱평균	F비
회귀	①	③	⑤	⑦
잔차	0.1075	④	⑥	
계	②	9		

$S_{yy} = 3.060, F(2, 7, 0.05) = 4.737$

$S_{yy} = 3.060$

①/③

$10 - 1$

⑤/⑥

$10 - 2 - 1$
2

$0.1075 / ④$