8장. 상관분석과 회귀분석

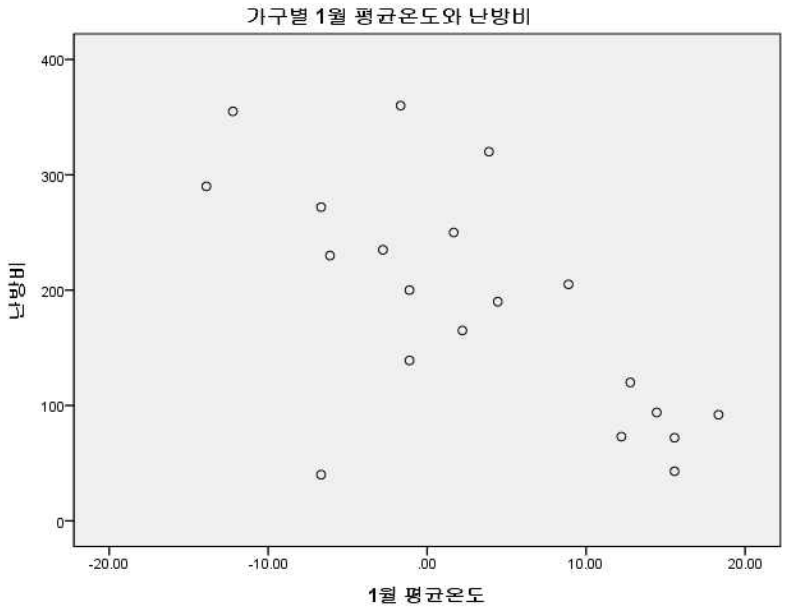
1. 서론

예) 1월의 평균온도(℃)와 난방비 지출액(단위:백원)과의 관계

평균온도(x)	1.67	-1.67	2.22	15.56	18.33	-1.11	-12.22	-13.89	-6.11	12.78
난방비(y)	250	360	165	43	92	200	355	290	230	120

평균온도(x)	12.22	8.89	-6.67	3.89	15.56	-6.67	14.44	4.44	-2.78	-1.11
난방비(y)	73	205	400	320	72	272	94	190	235	139

- 서로 상관이 있을까?
- 서로 상관이 있다면 평균온도로부터 난방비 지출액을 예측할 수 있을까?



- 자연 또는 사회현상의 규명에 있어서 관련된 변수들간의 상호관련성을 수학적인 함수의 형태로서 찾고자 함
- 지능지수와 학업성적, 흡연량과 수명, 공정온도와 제품의 강도, 초등학교 아동들의 학년에 따른 평균신장의 변화 등...
- 한 변수의 변화로부터 다른 변수의 변화를 예측, 또한 관심있는 변수의 최적값은 이 변수에 영향을 주는 다른 변수가 어떤 값을 취할 때 얻어질 것인가를 결정하는데 도움을 줌

● 상관분석(correlation analysis)과 회귀분석(regression analysis) : 표본에 나타난 두 변수 사이의 관계를 이용하여 모집단에서의 관계를 추측 또는 예측하는 통계적 추론의 방법

- 관측값 : $(x_1,\ y_1),\ (x_2,\ y_2),\cdots,(x_n,\ y_n)$
- 예) ① x : 영어성적, y : 수학성적
  ② x : 공정온도, y : 강도

③ x : 진통제의 분량, y : 진통지속시간

- 산점도, 표본상관계수 : 두 변수사이의 관계를 파악하기 위해 자료를 요약하는 방법

2. 상관분석

1) 상관분석(correlation analysis)

(1) 상관계수는 두 변수의 직선관계가 얼마나 강하고 어떤 방향인지를 나타냄

(2) 상관분석은 두 변수의 상관계수를 분석함으로서, 두 변수 사이의 연관성을 분석

(3) 표본상관계수를 이용하여 모상관계수에 대하여 추론

2) X와 Y의 모상관계수

(1) (X, Y)가 확률변수일 때, 상관계수 ρ는 다음과 같이 정의된다.

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

(2) 성질

① 두 확률변수 사이에 직선관계가 얼마나 강하고 어떤 방향인지를 나타내는 측도

② 모집단의 분포가 좌표평면에서 양의 방향이면 양의 값을, 음의 방향이면 음의 값을 가진다.

③ $Corr(aX+b, cY+d) = sign(ac)Corr(X, Y)$

④ $-1 \leqq \rho \leqq 1$

⑤ 어떤 직선 주위에 밀집되어 나타날수록 -1 또는 1에 가깝게 주어진다.

3) 표본상관계수

- 두 변수 x, y의 선형관계를 나타내는 측도 = Pearson의 표본상관계수
- 랜덤표본 $(X_1, Y_1), \cdots, (X_n, Y_n)$에 대한 관측값이 $(x_1, y_1), \cdots, (x_n, y_n)$ 일 때, 표본상관계수 r은 다음과 같이 정의된다.

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

(1) 제곱합과 곱의 합의 기호와 간편계산법 :

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

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

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

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

(2) 표본상관계수의 성질

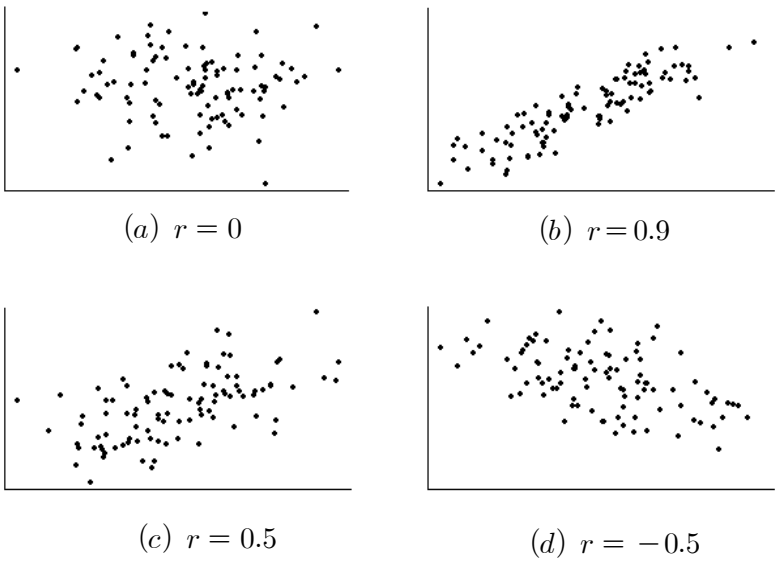
① $-1 \leqq r \leqq 1$

② r의 값은 관측값이 직선 관계에 가까울수록, -1(기울기가 음의 방향) 또는 1(기울기가 양의 방향)에 가까움.

③ $|r|$의 값이 1에 가까울수록 강한 상관관계, $|r|$의 값이 0에 가까울수록 약한 상관관계

④ 해석 시 주의사항 : 만약 $r = 0$ 이라면 두 변수사이의 직선관계가 약함을 뜻하는 것이지, 반드시 두 변수 사이의 관계가 없음을 뜻하는 것은 아니다.

(3) 여러 가지 경우의 산점도와 표본상관계수



$(a)\ r = 0$ $(b)\ r = 0.9$

$(c)\ r = 0.5$ $(d)\ r = -0.5$

- 그림 8.1, 예 8-1

4) ρ 에 관한 추론(이변량정규모집단의 경우)

(1) 이론적 배경

모집단의 분포가 이변량정규분포인 경우에 $H_0 : \rho = 0$이면 다음이 성립한다.

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

(2) 상관관계의 유무에 관한 검정(이변량정규모집단의 경우)

귀무가설 $H_0 : \rho = 0$에 대한 가설검정

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

② 검정통계량의 관측값 : $t = \sqrt{(n-2)}\,\frac{r}{\sqrt{1-r^2}}$

③ 가설검정

대립가설	유의수준 α의 기각역	유의확률
① $H_1 : \rho > 0$	$t \geqq t_\alpha(n-2)$	$p = P(T \geqq t)$
② $H_1 : \rho < 0$	$t \leqq -t_\alpha(n-2)$	$p = P(T \leqq t)$
③ $H_1 : \rho \neq 0$	$\lvert t \rvert \geqq t_{\alpha/2}(n-2)$	$p = P(\lvert T \rvert \geqq \lvert t \rvert)$

- 검정통계량의 관측값 t가 주어진 유의수준 α에서의 기각역에 포함되거나, 유의확률(P-값)이 유의수준 α보다 작으면 귀무가설 H_0를 기각한다.(즉, 대립가설 H_1을 채택한다.)
- 예제 8.1

3. 단순회귀분석의 모형과 적합

1) 회귀분석(regression analysis)

- 변수들간의 함수적인 관련성을 규명하기 위하여 어떤 수학적 모형을 가정하고, 이 모형을 측정된 변수들의 자료로부터 추정하는 통계적 분석방법
- 추정된 모형을 사용하여 필요한 예측을 하거나 관심있는 통계적 추정과 검정을 실시함
- 회귀분석의 목적 : 설명변수와 반응변수의 관계를 구체적인 함수 형태($y = f(x)$)로 나타내고, 설명변수의 값으로부터 반응변수의 값을 예측하는 것으로 두 변수 사이의 함수관계를 분석.
- x를 설명변수(explanatory variable) 또는 독립변수(independent variable)
  y를 반응변수(response variable) 또는 종속변수(dependent variable)

(2) 단순회귀분석(simple regression analysis)

- 두 변수간의 관계를 일차함수(직선관계)로 모형화하여 분석 : $y = \alpha + \beta x$

(3) 중회귀분석(multiple regression analysis)

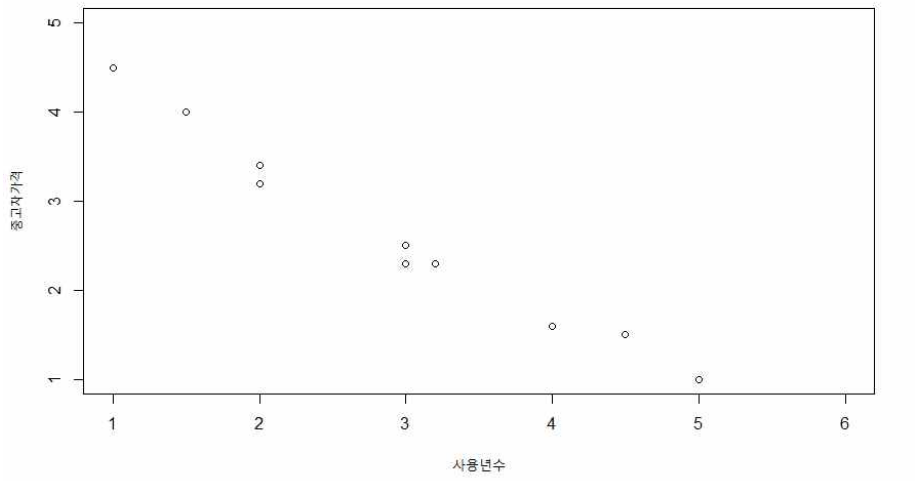
- 두 개 이상의 변수가 한 변수에 영향을 줄 때 한 변수를 여러 변수의 함수로 나타내어 분석 : $y=\beta_0+\beta_1 x_1+\cdots+\beta_k x_k$

예) 사용년수에 따른 중고차 가격의 변화

① 자료

사용년수	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

② 산점도



$\Rightarrow$ (중고차가격) $=\alpha+\beta\times$(사용년수)

설명변수, 독립변수 : 사용년수, 반응변수, 종속변수 : 중고차가격

● 회귀분석의 첫단계 : 주어진 자료를 이용하여 산점도를 그려 두 변수의 관계를 살펴봄

2) 단순선형회귀모형 또는 직선회귀모형

- 설명변수가 1개인 회귀모형으로 설명변수와 반응변수 사이에 직선관계가 있는 모형

(1) 자료구조 및 용어

① 자료구조 : $(x_1, Y_1),(x_2, Y_2),\cdots,(x_n, Y_n)$

② $(x_1,\cdots,x_n)$: 설명변수(독립변수) 즉, 두 변수가 있을 때, 다른 한 변수에 영향을 주는 변수

③ $(Y_1,\cdots,Y_n)$: 반응변수(종속변수) 즉, 두 변수가 있을 때, 다른 한 변수에 영향을 받는 변수

※ 관측값 : $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$

(2) 단순선형회귀모형 : 독립변수의 정해진 값 $x_1,x_2,\cdots,x_n$에서 측정되는 종속변수 $Y_1,Y_2,\cdots,Y_n$에 대하여 다음의 관계식이 성립한다고 가정하자.

$$Y_i=\alpha+\beta x_i+e_i,\quad i=1,2,...,n$$

α, β : 회귀모수(모회귀계수) (α: 상수항, β: 기울기)

$x_1,\cdots,x_n$: 설명변수(독립변수), $y_1,\cdots,y_n$: 반응변수(종속변수)

$e_1,\cdots,e_n$: 오차항으로 서로 독립인 $N(0,\sigma^2)$ 확률변수-선형성, 독립성, 등분산성

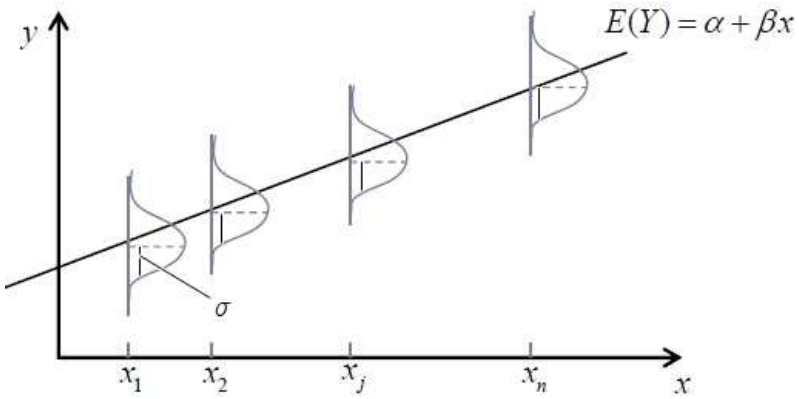
① $E(e_i)=0$, 즉 $E(Y_i|x_i)=\alpha+\beta x_i$, $(-\infty<\alpha<\infty,\ -\infty<\beta<\infty)$(선형성)

② $Var(e_1)=Var(e_2)=\cdots=Var(e_n)=\sigma^2>0$(등분산성)

③ $e_1,e_2,\cdots,e_n$은 서로 독립(독립성)

- 설명변수의 값이 $x_1,x_2,\cdots,x_n$일 때 대응하는 $Y_1,Y_2,\cdots,Y_n$은 서로 독립적으로 관측된다고 가정
- 그림 8-4 : $Y_i\sim N(\alpha+\beta x_i,\ \sigma^2)$

※ **단순회귀모형에서 α, β 가 미지, $x_1,x_2,\cdots,x_n$일 때 대응하는 $Y_1,Y_2,\cdots,Y_n$의 관측값을 사용하여, x에 따른 Y의 기댓값(평균) $E(Y|x)=E(Y)=\alpha+\beta x$를 추정하는 것이 단순회귀분석의 일차적인 작업**



(3) 모회귀직선(population regression line) : $y=\alpha+\beta x$

3) 최소제곱법

- **모회귀직선, 모회귀계수 추정 : 표본을 선출하여 추정**
- **이제, 위 모회귀직선의 식에서 모회귀계수값 α, β을 어떻게 추정하는 것이 타당할 것인가 ?**

  단순선형회귀모형과 모회귀직선의 근본적인 차이는 e_i, 이 항을 최소로 하는 것이 중요함.

  그런데 이 값은 양, 음 두가지가 있으므로 분산의 개념과 비슷하게 생각하여 제곱을 한다.

● **최소제곱법(least squares method)**

단순회귀모형 $Y_i = \alpha + \beta x_i + e_i$에서 오차의 제곱합

$Q(\alpha,\beta) = \sum_{i=1}^{n}(e_i)^2 = \sum_{i=1}^{n}\{ Y_i - (\alpha+\beta x)\}^{\ 2}$를 최소가 되도록 α와 β를 추정하는 방법을 **최소제곱법**이라 하고, 이때 얻어지는 추정량 $\hat{\alpha}$와 $\hat{\beta}$를 각각 α와 β의 **최소제곱추정량(Least squares estimator)**이라 한다.

(1) 최소제곱추정량의 유도

① $Q(\alpha,\beta)$를 α, β에 대하여 편미분 한 후 얻은 식을 0으로 놓으면 다음과 같은 연립방정식을 얻는다.

i) $\frac{\partial}{\partial\alpha} Q(\alpha,\beta) = 2\sum_{i=1}^{n}\{ Y_i - (\alpha+\beta x_i)\}(-1) = 0 \ \Rightarrow \ \sum_{i=1}^{n} Y_i = n\alpha + \beta\sum_{i=1}^{n} x_i$

ii) $\frac{\partial}{\partial\beta} Q(\alpha,\beta) = 2\sum_{i=1}^{n}\{ Y_i - (\alpha+\beta x_i)\}(-x_i) = 0 \ \Rightarrow \ \sum_{i=1}^{n} x_i Y_i = \alpha\sum_{i=1}^{n} x_i + \beta\sum_{i=1}^{n} x_i^2$

위 연립방정식을 "정규방정식"이라 함.

② 위 연립방정식을 풀면 α,β**의 최소제곱추정량**을 얻을 수 있다.

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{x}, \quad \hat{\beta} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(Y_i-\overline{Y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$$

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

- 제곱합 기호 : $S_{xx} = \sum_{i=1}^{n}(x_i-\overline{x})^2 = \sum x_i^2 - n\overline{x}^2 = \sum x_i^2 - (\sum x_i)^2/n$

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

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

(3) 최소제곱 회귀직선 : $\widehat{E(Y|x)} = \widehat{E(Y)} = \hat{y} = \hat{\alpha} + \hat{\beta}x$

- 여기서 $\hat{y_i}$는 반응변수 평균 $E(Y_i|x_i) = \alpha + \beta x_i$의 추정값, 즉 독립변수인 x가 주어졌을 때, y의 기대값을 의미하며 편의상 $\widehat{E(Y_i|x_i)}$대신 $\hat{y_i}$으로 나타냄.

- 최소제곱회귀직선은 다음과 같이 변형되어 사용되기도 함

$$\begin{aligned}\hat{y} = \hat{\alpha} + \hat{\beta}x &= (\hat{\alpha} + \hat{\beta}x) + \hat{\beta}(x - \bar{x}) \\ &= (\bar{y} - \hat{\beta}\bar{x} + \hat{\beta}\bar{x}) + \hat{\beta}(x - \bar{x}) \\ &= \bar{y} + \hat{\beta}(x - \bar{x})\end{aligned}$$

- 예제 8.2 : 사용년수에 따른 중고차 가격자료를 이용하여 단순회귀직선을 추정하여라.

사용년수	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

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

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

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2 = \sum x_i^2 - (\sum x_i)^2/n = 155.99 - (39.7)^2/12 = 24.649$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - (\sum x_i)(\sum y_i)/n = 69.81 - (39.7)(27.5)/12 = -21.169$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} = -21.169/24.649 = -0.859$$

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 2.292 - (-0.859)(3.308) = 5.133$

4) 잔차(residual)

- 적합된 회귀직선이 실제 관측 결과를 잘 설명해 주는가? 최소제곱회귀직선이 반응변수 값의 변화를 얼마나 잘 설명하는가?

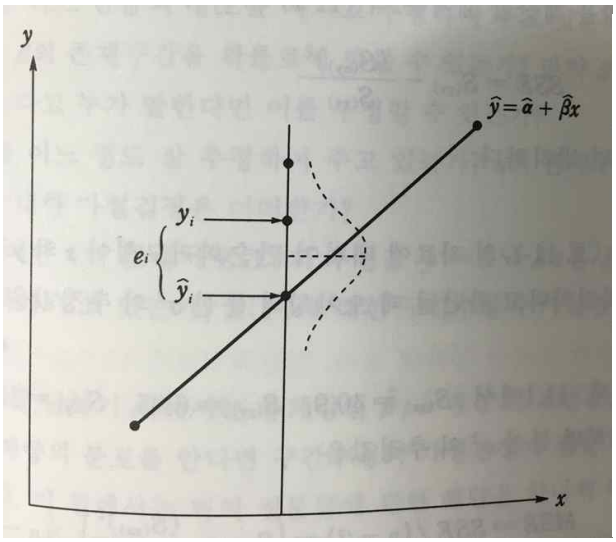
- 단순회귀모형 $Y_i = \alpha + \beta x_i + \epsilon_i$에서 실제 관측값 Y_i와 추정량 $\hat{Y}_i = \hat{\alpha} + \hat{\beta} x_i$의 차

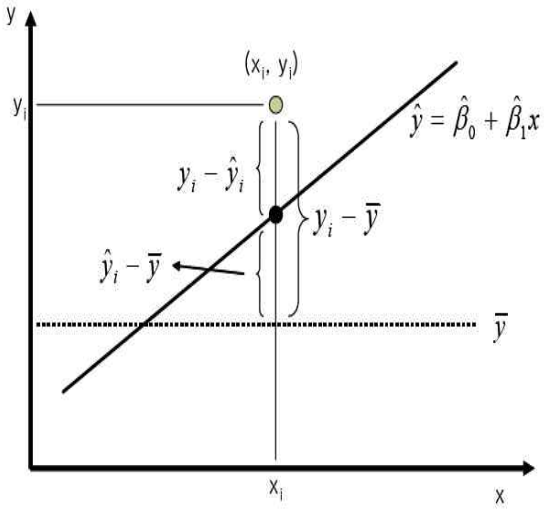
$$\hat{e}_i = Y_i - \hat{Y}_i$$

(1) 잔차의 관측값 : $\hat{e}_i = y_i - \hat{y}_i \ (i = 1, 2, \cdots, n)$

- 오차항 e_i의 관측값

(2) 잔차의 의미 : 회귀직선에 설명 안되는 편차





(3) 잔차가 **<u>작으면 작을수록 최소제곱회귀직선이 실제 관측 결과를 잘 설명해 준다 할 수 있음</u>**

오차항의 분산 σ^2이 작을수록 잔차는 0에 가깝게 나타날 것이고, 오차항의 분산 σ^2이 클수록 큰 값으로 나타날 수 있음.

(4) 잔차제곱합 SSE(residual sum of squares) 또는 잔차변동

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

- 오차항의 분산 σ^2에 대한 정보를 줌

- 잔차제곱합의 계산은 다음 식을 이용하면 편리함 : $SSE = S_{yy} - \hat{\beta}^2 S_{xx} = S_{yy} - \dfrac{S_{xy}^2}{S_{xx}}$

5) 오차분산 σ^2의 추정 : 평균제곱오차 (mean squared error)

$$\hat{\sigma^2} = MSE = \frac{SSE}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y_i})^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = \frac{1}{n-2}\sum_{i=1}^{n}\{y_i - \bar{y} - \hat{\beta}(x_i - \bar{x})\}^2$$

- SSE를 SSE자유도 $n-2$로 나누면 평균제곱오차를 얻음. 자유도가 $n-2$인 것은 회귀모수 α, β를 고려하기 때문.

- $\hat{\sigma^2} = MSE$: $E(MSE) = \sigma^2$, σ^2의 불편추정량

- 예제 8.3 : 예8.2에서 단순선형회귀모형을 적용할 때, 오차분산 σ^2의 추정값을 구하여라.

  $S_{yy} = 18.649$

  $$SSE = S_{yy} - \hat{\beta}^2 S_{xx} = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 18.649 - (-21.649)^2/24.649 = 0.280$$

  $$\hat{\sigma^2} = MSE = \frac{SSE}{n-2} = 0.289/10 = 0.0289$$

6) 총편차(총제곱합)의 분해

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

총편차(total deviation) = (오차항에 기인한 부분 : 잔차)+(회귀직선에 의해 설명되는 부분)

양변제곱 : $\left[\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y_i})^2 + \sum_{i=1}^{n}(\hat{y_i} - \bar{y})^2\right]$

[총제곱합 = 오차제곱합 + 회귀제곱합]

$SST = SSE + SSR$

자유도 : $n-1 = n-2 + 1$

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

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

$$\left(SSR = \sum_{i=1}^{n}(\hat{y_i}-\bar{y})^2 = \sum_{i=1}^{n}(\hat{\alpha}+\hat{\beta}x_i-\bar{y})^2 = \sum_{i=1}^{n}(\bar{y}-\hat{\beta}\bar{x}+\hat{\beta}x_i-\bar{y})^2 = \sum_{i=1}^{n}\hat{\beta}^2(x_i-\bar{x})^2 \right.$$
$$\left. = \sum_{i=1}^{n}\hat{\beta}^2 S_{xx} = [\frac{S_{xy}}{S_{xx}}]^2 S_{xx} = \frac{S_{xy}^2}{S_{xx}} = (\sum_{i=1}^{n} c_i y_i)^2, \quad where \quad c_i = (x_i-\bar{x})/\sqrt{S_{xx}} \right)$$

$$SSE = \sum_{i=1}^{n}(y_i-\hat{y_i})^2 = SST - SSR = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

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

7) 결정계수 R^2 : 총변동 가운데 회귀직선으로 설명되는 비율

- 즉 과연 이 회귀직선이 데이터를 얼마나 설명할까 ? ⇒ 결정계수 R^2의 값으로 설명

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \frac{S_{xy}^2}{S_{xx}S_{yy}}$$

- 총제곱합(총변동) *SST*중에서 회귀제곱합(회귀직선) *SSR*이 차지하는 비중이 크면, 즉, 결정계수가 1에 가까울수록 회귀직선에 의해 설명이 잘됨을 의미함

- 단순회귀분석에서는 R^2은 표본상관계수의 제곱.

- 예제8.4 : 예8.2에서 단순선형회귀모형을 적용할 때, 결정계수를 구하고 이를 해석하여라.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \frac{(-21.169)^2}{(24.649)(18.469)} = 0.9844$$

(회귀직선이 전체 반응변수값의 산포 중 98%를 설명한다는 뜻임)

※ 잔차(residual)와 관련된 몇가지 성질

① $\sum_{i=1}^{n}\hat{e_i} = 0$ => pf) $\sum_{i=1}^{n}\hat{e_i} = \sum_{i=1}^{n}(y_i-\hat{y_i}) = \sum_{i=1}^{n}y_i - n\hat{\alpha} - \hat{\beta}\sum_{i=1}^{n}x_i = 0 (\because$ 정규방정식)

② $\sum_{i=1}^{n}x_i\hat{e_i} = 0$ =>

pf) $\sum_{i=1}^{n}x_i\hat{e_i} = \sum_{i=1}^{n}x_i(y_i-\hat{y_i}) = \sum_{i=1}^{n}x_i y_i - \hat{\alpha}\sum_{i=1}^{n}x_i - \hat{\beta}\sum_{i=1}^{n}x_i^2 = 0 (\because$ 정규방정식)

③ $\sum_{i=1}^{n}\hat{y_i}\hat{e_i} = 0$ => pf) $\sum_{i=1}^{n}\hat{y_i}\hat{e_i} = \hat{\alpha}\sum_{i=1}^{n}\hat{e_i} + \hat{\beta}\sum_{i=1}^{n}x_i\hat{e_i} = 0 \quad (\because \sum_{i=1}^{n}\hat{e_i} = 0, \sum_{i=1}^{n}x_i\hat{e_i} = 0)$

④ $2\sum_{i=1}^{n}(y_i-\hat{y_i})(\hat{y_i}-\bar{y})=0$

pf) $2\sum_{i=1}^{n}(y_i-\hat{y_i})(\hat{y_i}-\bar{y})=2\sum_{i=1}^{n}\hat{e_i}\hat{y_i}-2\bar{y}\sum_{i=1}^{n}\hat{e_i}=0$ [$\because$ by (1) (3)]

4. 단순회귀분석에서의 추론

1) 회귀모수의 타당성과 추정을 어떻게 하는 것일까 ?

$$Y_i=\alpha+\beta x_i+e_{i,} \quad e_i\sim N(0,\sigma^2), i=1,2,...,n\text{(오차가 정규분포를 따른다는 가정)}$$

① 단순회귀모형에서 설명변수가 반응변수의 변화를 설명하는데 의미가 있는가를 어떻게 설명할 것인가? $\Rightarrow$ $H_0:\beta=0 \quad v.s \quad H_1:\beta\neq 0$

("회귀직선이 유의한가 ?"를 의미한다.)

② 기울기 β에 관한 추론, 평균반응 $E(Y|x)=\alpha+\beta x$에 관한 추론, 절편α에 관한추론

2) 회귀직선의 유의성 검정

(1) 회귀직선의 유의성에 대한 가설

$H_0:\beta=0 \quad v.s \quad H_1:\beta\neq 0$

(2) 이론적 배경

귀무가설 H_0이 사실일 때, $F=\frac{MSR}{MSE}=\frac{SSR/1}{SSE/(n-2)}\sim F(1,n-2)$ 가 성립한다.

※ β가 0으로부터 멀어질수록 반응변수 값의 변동을 나타내는 총제곱합 SST중 회귀제곱합 SSR이 차지하는 비중이 커지고 잔차제곱합 SSE가 차지하는 비중이 작아진다.

즉, $\frac{SSR}{SSE}$의 값이 커질수록 회귀직선이 유의하다는 증거임.

$MSE=\frac{SSE}{(n-2)}$: 잔차평균제곱, $MSR=\frac{SSR}{1}$: 회귀평균제곱을 이용하여 $F=\frac{MSR}{MSE}$을 정의.

두 통계량 MSE, MSR 은 모두 카이제곱분포를 따르면서 서로 독립이므로 검정통계량 $F=\frac{MSR}{MSE}$은 $F(1,n-2)$를 따른다.

(3) 가설검정절차

$H_0 : \beta = 0 \quad v.s \quad H_1 : \beta \neq 0$

유의수준 α

검정통계량 : $F = \dfrac{MSR}{MSE} = \dfrac{SSR/1}{SSE/(n-2)} \sim F(1, n-2)$

검정통계량의 관측값 : $f = \dfrac{MSR}{MSE} = \dfrac{SSR/1}{SSE/(n-2)}$

유의확률	유의수준 α의 기각역
$p = P(F \geq f), \ F \sim F(1, n-2)$	$f \geq F_{\alpha}(1, n-2)$

- 검정통계량의 관측값 f가 주어진 유의수준 α에서의 기각역에 포함되거나, 유의확률(P-값)이 유의수준 α보다 작으면 귀무가설 H_0를 기각한다.(즉, 대립가설 H_1을 채택한다.)

(4) 회귀직선의 유의성 검정을 위한 분산분석표(Analysis of Variance Table)

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

- 예제 8.5 : 예 8.2에서 단순선형회귀모형을 적용할 때, 회귀직선의 유의성 검정을 하여라.

$H_0 : \beta = 0 \quad v.s \quad H_1 : \beta \neq 0$

$SST = S_{yy} = 18.469$

$SSR = (S_{xy})^2 / S_{xx} = (-21.169)^2 / 24.649 = 18.180$

$SSE = SST - SSR = 18.469 - 18.180 = 0.289$

요인	제곱합	자유도	평균제곱	F 값	유의확률
회귀	18.180	1	18.180	629.76	0.0001
잔차	0.289	10	0.0289		
계	18.469	11			

$MSR = SSR/1 = 18.180/1 = 18.180$

$MSE = SSE/(n-2) = 0.289/10 = 0.0289$

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

$p = P(F \geq f) = p(F > 629.76) = 0.0001$

그러므로 회귀직선은 유의하다고 할 수 있다.

- Advance Course-

● **회귀직선의 유의성 검정과 상관계수 ρ의 검정과의 관계**

1) 단순회귀분석에서 회귀직선 유의성검정 통계량

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

2) 상관계수 ρ의 검정에서의 검정통계량

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

1)과 2)의 관계 : $T^2\backsim F$이므로

$$\because\ T^2=(n-2)\frac{r^2}{1-r^2}=(n-2)(SSR/SST)/(1-SSR/SST)=(n-2)SSR/SSE$$

※ 모집단 회귀직선을 $y=\alpha+\beta x$이라고 하면 여기서 α, β는 모수이다. 실제에 있어서 모집단에 속해 있는 모든 관찰점들을 전부 관찰할 수가 없으므로 모집단으로부터 n개의 관찰점 (x_i,y_i)를 추출하여, 이 표본으로부터 표본 회귀직선($\hat{y}=\hat{\alpha}+\hat{\beta}x$)을 추정하는 것이다.

여기서, $\hat{y}$는 $E(y)$, $\hat{\alpha}$은 α, 그리고 $\hat{\beta}$은 β의 점추정량이다. 이 추정량들을 물론 통계량이고 분산을 가지고 있으며, 분포를 가지고 있다. 따라서, 구간추정과 가설검정을 할 수 있다.

3) 모회귀계수 β에 관한 추론 σ

(1) 이론적 배경

- 모집단 회귀직선의 기울기 β을 추정하기 위하여 표본으로부터 얻어지는 $\hat{\beta}=\frac{S_{xy}}{S_{xx}}$.

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

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

$\hat{\beta}$ 는 Y_i의 선형결합으로 나타낼 수 있음

① $E(\hat{\beta})=\beta$, ② $Var(\hat{\beta})=\frac{\sigma^2}{S_{xx}}$

증명) $\hat{\beta}= \dfrac{S_{xy}}{S_{xx}}= \dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}= \sum_{i=1}^{n}\dfrac{(x_i-\bar{x})}{S_{xx}}y_i = \sum_{i=1}^{n}k_iY_i, \quad where\ \ k_i = \dfrac{(x_i-\bar{x})}{S_{xx}}$

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

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

i) $\sum k_i = 0$

$\sum k_i = \sum\dfrac{(x_i-\bar{x})}{S_{xx}} = 0, \quad where\ \ \sum(x_i-\bar{x})=0, \quad \sum x_i - n\bar{x}=0$

ii) $\sum k_ix_i = 1$

$\sum k_ix_i = \sum\dfrac{(x_i-\bar{x})}{S_{xx}}x_i = \sum\dfrac{(x_i^2-\bar{x}x_i)}{S_{xx}} = \dfrac{\sum x_i^2-\bar{x}\sum x_i}{\sum(x_i-\bar{x})^2}$

$= \dfrac{\sum x_i^2 - n\bar{x}^2}{\sum(x_i-\bar{x})^2} = 1$

$,where\ \ \sum(x_i-\bar{x})^2 = \sum x_i^2 + n\bar{x}^2 - 2\bar{x}\sum x_i = \sum x_i^2 + n\bar{x}^2 - 2n\bar{x}^2 = \sum x_i^2 - n\bar{x}^2$

iii) $\sum k_i^2 = \dfrac{1}{\sum(x_i-\bar{x})^2}$

$\sum k_i^2 = \sum\left[\dfrac{(x_i-\bar{x})}{S_{xx}}\right]^2 = \dfrac{\sum(x_i-\bar{x})^2}{S_{xx}^2} = \dfrac{\sum(x_i-\bar{x})^2}{\left[\sum(x_i-\bar{x})^2\right]^2} = \dfrac{1}{\sum(x_i-\bar{x})^2}$

① $E(\hat{\beta}) = E\left(\sum k_iY_i\right) = \sum k_iE(Y_i)$

$= \sum k_iE(\alpha+\beta x_i+e_i) = \sum k_i\left[\alpha+\beta x_i+E(e_i)\right]$

$= \sum k_i(\alpha+\beta x_i) = \alpha\sum k_i + \beta\sum k_ix_i = \beta$

② $Var(\hat{\beta}) = Var(\sum k_iY_i) = \sum k_i^2\,Var(Y_i) = \sum k_i^2\,Var(\alpha+\beta x_i+e_i) = \sum k_i^2\,Var(e_i)$

$= \dfrac{\sigma^2}{\sum(x_i-\bar{x})^2} = \dfrac{\sigma^2}{S_{xx}}$

● $Y_i \sim N(\alpha+\beta x_i,\ \sigma^2)$이고 $\hat{\beta}$이 $\sum_{i=1}^{n}k_iY_i$로 Y_i들의 일차결합이므로 $\hat{\beta}$도 정규분포를 따른다.

(2) 추정량 $\hat{\beta}$에 대한 정리

① $\hat{\beta}$는 정규분포를 따른다 : $\hat{\beta} \sim N\left(\beta,\ \frac{\sigma^2}{S_{xx}}\right)$

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

② 평균제곱오차 $MSE = \frac{SSE}{n-2}$: σ^2의 불편추정량, 즉 $E(MSE) = \sigma^2$

③ $Var(\hat{\beta})$의 불편추정량 : $\frac{\hat{\sigma^2}}{S_{xx}} = \frac{MSE}{S_{xx}}$

④ $\sigma \Rightarrow \hat{\sigma} = \sqrt{MSE}$이므로 $T = \frac{\hat{\beta}-\beta}{\hat{\sigma}/\sqrt{S_{xx}}} = \frac{\hat{\beta}-\beta}{\sqrt{\frac{MSE}{S_{xx}}}} \sim t(n-2)$

(3) β에 관한 $100(1-\alpha)\%$ 신뢰구간 : $\hat{\beta} \pm t_{\alpha/2}(n-2)\sqrt{\frac{MSE}{S_{xx}}}$

(4) $H_0 : \beta = \beta_0$의 가설검정

- 검정통계량 : $T = \frac{\hat{\beta}-\beta_0}{\sqrt{\frac{MSE}{S_{xx}}}}$ $(\sim t(n-2))$
- 검정통계량의 관측값 : t
- 가설검정

대립가설	유의수준 α에서 기각역	유의확률
$H_1 : \beta > \beta_0$ 일때	$t \geqq t_\alpha(n-2)$	$p = P(T \geqq t)$
$H_1 : \beta < \beta_0$ 일때	$t \leqq -t_\alpha(n-2)$	$p = P(T \leqq t)$
$H_1 : \beta \neq \beta_0$ 일때	$\lvert t \rvert \geqq t_{\alpha/2}(n-2)$	$p = P(\lvert T \rvert \geqq \lvert t \rvert)$

- 예제8.6

4) 평균반응 $E(Y|x)=\alpha+\beta x$에 관한 추론

(1) 이론적 배경

- 어떤 주어진 x값에서 y의 기댓값

● $E(Y|x)=E(Y)=\alpha+\beta x$의 추정량

: $\hat{Y}=\hat{\alpha}+\hat{\beta}x=\overline{Y}-\hat{\beta}\overline{x}+\hat{\beta}x=\overline{Y}+\hat{\beta}(x-\overline{x})=\sum_{i=1}^{n}\left(\frac{1}{n}+(x-\overline{x})\frac{(x_i-\overline{x})}{S_{xx}}\right)Y_i$

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

$where,\ S_{xx}=\sum_{i=1}^{n}(x_i-\overline{x})^2$

증명) $\hat{Y}=\hat{\alpha}+\hat{\beta}x$ or $\hat{Y}=\overline{Y}+\hat{\beta}(x-\overline{x})$

① $E(\hat{Y})=E(\hat{\alpha}+\hat{\beta}x)=\alpha+\beta x$

② $Var(\hat{Y})=Var[\overline{Y}+\hat{\beta}(x-\overline{x})]=Var(\overline{Y})+(x-\overline{x})^2Var(\hat{\beta})+2(x-\overline{x})Cov(\overline{Y},\hat{\beta})$

$=\frac{\sigma^2}{n}+(x-\overline{x})^2\frac{\sigma^2}{S_{xx}}=\sigma^2\left[\frac{1}{n}+\frac{(x-\overline{x})^2}{S_{xx}}\right]$

($where\ \ Cov(\overline{Y},\hat{\beta})=0, \quad Var(\hat{\beta})=\frac{\sigma^2}{S_{xx}}$)

③ $Put,\ \overline{Y}=\sum a_iY_i,\ a_i=\frac{1}{n}, \quad \hat{\beta}=\sum k_iY_i, \quad k_i=\frac{(x_i-\overline{x})}{S_{xx}}$

$Cov(\overline{Y},\ \hat{\beta})=Cov(\sum a_iY_i,\ \sum k_iY_i)=\sigma^2\sum a_ik_i=\frac{\sigma^2\sum(x_i-\overline{x})}{(nS_{xx})}=0$

$(where\ \sum(x_i-\overline{x})=0, \quad \sum x_i-n\overline{x}=0)$

● $\hat{\alpha}+\hat{\beta}x$는 서로 독립이며 $Y_i\sim N(\alpha+\beta x_i,\ \sigma^2)$ 정규분포를 따르는 확률변수 $Y_1, Y_2,\cdots, Y_n$의 일차결합이므로 $\hat{\alpha}+\hat{\beta}x$는 정규분포를 따른다.

(2) $E(Y)=\alpha+\beta x$의 추정량 $\hat{\alpha}+\hat{\beta}x$에 대한 정리

① $\hat{\alpha}+\hat{\beta}x$는 정규분포를 따른다. : $\hat{\alpha}+\hat{\beta}x\sim N\left(\alpha+\beta x,\ \sigma^2\left[\frac{1}{n}+\frac{(x-\overline{x})^2}{S_{xx}}\right]\right)$

$\therefore\ Z=\frac{(\hat{\alpha}+\hat{\beta}x)-(\alpha+\beta x)}{\sqrt{\sigma^2(\frac{1}{n}+\frac{(x-\overline{x})^2}{S_{xx}})}}\sim N(0,1)$

② $\sigma \Rightarrow \hat{\sigma}=\sqrt{MSE}$이므로

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

(3) $\alpha+\beta x$에 관한 $100(1-\alpha)\%$ 신뢰구간 : $(\hat{\alpha}+\hat{\beta}x)\pm t_{\alpha/2}(n-2)\sqrt{MSE\left(\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}\right)}$

(4) $H_0:\alpha+\beta x=\mu_0$에 대한 가설검정

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

- 검정통계량의 관측값 : t

- 가설검정

대립가설	유의수준 α에서 기각역	유의확률
$H_1:\alpha+\beta x>\mu_0$일때	$t\geqq t_\alpha(n-2)$	$p=P(T\geqq t)$
$H_1:\alpha+\beta x<\mu_0$일때	$t\leqq -t_\alpha(n-2)$	$p=P(T\leqq t)$
$H_1:\alpha+\beta x\neq\mu_0$일때	$\lvert t\rvert\geqq t_{\alpha/2}(n-2)$	$p=P(\lvert T\rvert\geqq\lvert t\rvert)$

- 예제 8.7, 그림8.8

5) 모회귀계수 α에 관한 추론

(1) 이론적 배경

- 모집단 회귀직선의 절편 α는 표본으로부터 얻어지는 $\hat{\alpha}=\bar{y}-\hat{\beta}\bar{x}$으로 추정됨.

● $\hat{\alpha}=\bar{Y}-\hat{\beta}\bar{x}, \quad E(\hat{\alpha})=\alpha, \quad Var(\hat{\alpha})=\sigma^2(\frac{1}{n}+\frac{(\bar{x})^2}{S_{xx}})$

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

$$\hat{\alpha}=\bar{Y}-\hat{\beta}\bar{x}=\frac{1}{n}\sum Y_i-\sum\frac{(x_i-\bar{x})}{S_{xx}}Y_i\bar{x}=\sum\left(\frac{1}{n}-\bar{x}\frac{(x_i-\bar{x})}{S_{xx}}\right)Y_i$$

증명)

① $E(\hat{\alpha})=E(\bar{Y})-\bar{x}E(\hat{\beta})=\alpha+\beta\bar{x}-\bar{x}\beta=\alpha$

② $Var(\hat{\alpha})=Var(\bar{Y}-\hat{\beta}\bar{x})$

$$=Var(\bar{Y})+(\bar{x})^2Var(\hat{\beta})-2\bar{x}Cov(\bar{Y},\hat{\beta})$$

$$=\frac{\sigma^2}{n}+(\bar{x})^2\frac{\sigma^2}{S_{xx}}=\sigma^2\left(\frac{1}{n}+\frac{(\bar{x})^2}{S_{xx}}\right)$$

$(\ where\ \ Cov(\bar{Y},\hat{\beta})=0, \quad Var(\hat{\beta})=\frac{\sigma^2}{S_{xx}}\ )$

③ $Put,\ \bar{Y}=\sum a_iY_i,\ a_i=\frac{1}{n}, \quad \hat{\beta}=\sum k_iY_i, \quad k_i=\frac{(x_i-\bar{x})}{S_{xx}}$

$$Cov(\bar{Y},\ \hat{\beta})=Cov(\sum a_iY_i,\ \sum k_iY_i)=\sigma^2\sum a_ik_i=\frac{\sigma^2\sum(x_i-\bar{x})}{(nS_{xx})}=0$$

$(where\ \sum(x_i-\bar{x})=0, \quad \sum x_i-n\bar{x}=0)$

● $\hat{\alpha}$는 $Y_i\sim N(\alpha+\beta x_i,\ \sigma^2)$정규분포를 따르는 확률변수 $Y_1, Y_2,\cdots, Y_n$의 일차결합이므로 $\hat{\alpha}$는 정규분포를 따른다.

(2) 추정량 $\hat{\alpha}$에 대한 정리

① $\hat{\alpha}$는 정규분포를 따른다. : $\hat{\alpha}=\bar{y}-\hat{\beta}\bar{x}\sim N\left(\alpha,\ \sigma^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]\right)$

$$\therefore\ Z=\frac{\hat{\alpha}-\alpha_0}{\sqrt{\sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}}\sim N(0,1)$$

② $\sigma \Rightarrow \hat{\sigma} = \sqrt{MSE}$이므로

$$T = \frac{\hat{\alpha} - \alpha_0}{\hat{\sigma}\sqrt{(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}})}} = \frac{\hat{\alpha} - \alpha_0}{\sqrt{MSE\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)}} \sim t(n-2)$$

(2) α에 관한 $100(1-\alpha)\%$ 신뢰구간 : $\hat{\alpha} \pm t_{\alpha/2}(n-2)\sqrt{MSE\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)}$

(3) $H_0 : \alpha = \alpha_0$의 가설검정

- 검정통계량 : $T = \frac{\hat{\alpha} - \alpha_0}{\hat{\sigma}\sqrt{(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}})}} = \frac{\hat{\alpha} - \alpha_0}{\sqrt{MSE\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)}} \sim t(n-2)$
- 검정통계량의 관측값 : t
- 가설검정

대립가설	**유의수준 α에서 기각역**	**유의확률**
$H_1 : \alpha > \alpha_0$일때	$t \geqq t_\alpha(n-2)$	$p = P(T \geqq t)$
$H_1 : \alpha < \alpha_0$일때	$t \leqq -t_\alpha(n-2)$	$p = P(T \leqq t)$
$H_1 : \alpha \neq \alpha_0$일때	$\lvert t \rvert \geqq t_{\alpha/2}(n-2)$	$p = P(\lvert T \rvert \geqq \lvert t \rvert)$

- 단순선형회귀모형에서 회귀직선의 절편인 α에 관한 추론은 평균반응 $E(Y|x) = \alpha + \beta x$에서 $x = 0$인 경우에 해당됨

- 예제 8.8

5. 단순회귀에서의 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) 회귀모형($Y_i = \alpha + \beta x_i + e_i$)과 오차항(e_i)에 대하여 다음과 같은 가정이 내포됨

① 선형성 : $E(Y_i|x_i) = \alpha + \beta x_i$

② 등분산성 : $Var(e_1) = ... = Var(e_n) = \sigma^2$

③ 독립성 : $e_1, e_2, ..., e_n$ 은 서로 독립

④ 정규성 : $e_i \sim N(0, \sigma^2)$

위의 가정 4가지에 대한 타당성을 검토 : 잔차분석(Analysis of residual)

3) 스튜던트화 잔차(studentized residual) 및 잔차도

(1) 표준화된 오차항

$$\frac{e_i}{sd(e_i)} = \left(\frac{Y_i - (\alpha + \beta x_i)}{\sigma}\right) \sim N(0,1), \quad i = 1,2,...,n \text{ 이고 서로 독립}$$

(2) 잔차 $\hat{e_i} = y_i - \hat{y_i}$

(3) 잔차의 분산 $Var(\hat{e_i}) = Var(y_i - \hat{y_i}) = \sigma^2 \left\{ 1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}} \right\}$

(4) 스튜던트화 잔차(studentized residual)

- 표준화된 통계량을 이용. 따라서 잔차분석에서는 표준화된 잔차를 이용하는데 표준화 잔차는 잔차 $\hat{e_i}$ 를 표준편차로 나눈 것

$$\widehat{e_{st,i}} = \frac{\hat{e_i}}{\hat{sd}(\hat{e_i})} = \frac{y_i - \hat{y_i}}{\sqrt{\hat{\sigma}\left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}\right)}} \ , \ \hat{\sigma} = \sqrt{MSE}$$

※ 스튜던트화 잔차를 표준화된 오차항의 관측값으로 간주

(5) 잔차도(residual plot) : $(x_1, \widehat{e_{st,1}}), (x_2, \widehat{e_{st,2}}), ..., (x_n, \widehat{e_{st,n}})$ 의 plotting

4) 잔차분석

- 스튜던트화 잔차($\widehat{e_{st,i}},\ \ i=1,\cdots,n$)가 표준정규분포의 서로 독립인 n개의 관측값과 유사하게 나타나는가를 검토하여 단순선형회귀모형의 적용타당성을 알아봄.

- $(x_1,\widehat{e_{st,1}}),(x_2,\widehat{e_{st,2}}),...,(x_n,\widehat{e_{st,n}})$의 위치를 표시한 잔차도(residual plot)가 다음과 같은 성질을 만족하면 단순선형회귀모형 적용이 타당하다고 간주함.
  ① 선형성 : 대략 0에 관하여 대칭.
  ② 등분산성 :설명변수 값에 따른 잔차의 산포가 크게 다르지 않음.
  ③ 독립성 : 점들이 산재된 모양에 특별한 경향이 없음.
  ④ 정규성 : 대부분의 점들이 ± 2의 범위 내에 있음.

● 단순회귀분석의 예
- 가정이 잘 맞은 경우 ⇒ 그림 8.9 참조
- 위의 ①,②,③,④의 가정에 위배되는 경우의 잔차 plot의 그림 ⇒ 그림 8.10참조

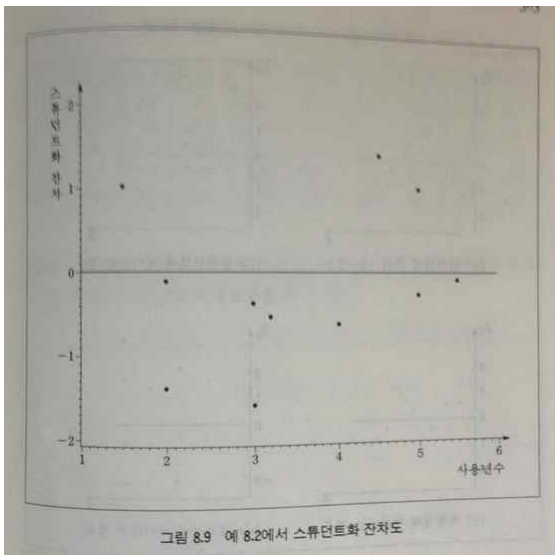


그림 8.9 예 8.2에서 스튜던트화 잔차도

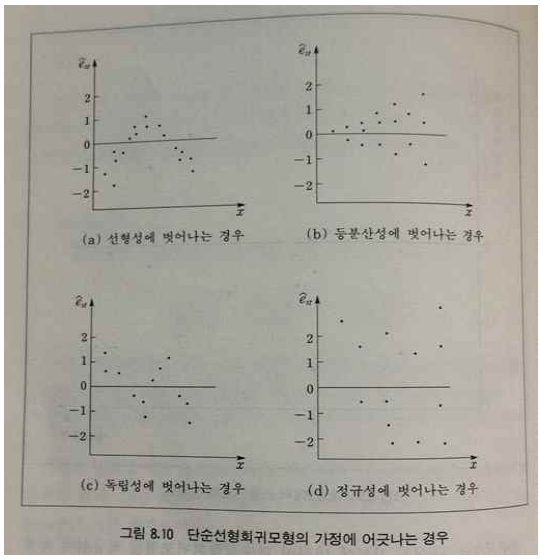
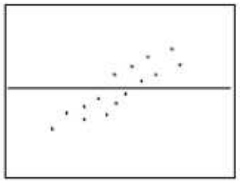
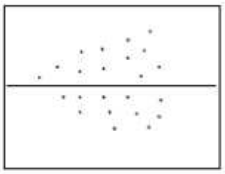


그림 8.10 단순선형회귀모형의 가정에 어긋나는 경우

※ 가정이 어긋난 경우

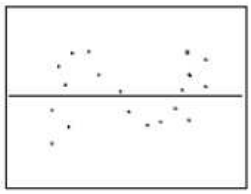


선형항을 추가하는 것이 타당

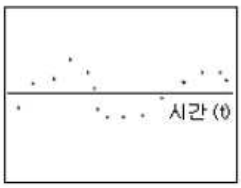


오차항의 정규성, 등분산성 위배
- 가중회귀모형 타당

선형성 위배
이차선형모형 타당



선형성 위배
3차 회귀모형이 타당



독립성 위배
시계열 분석이 타당

6. 중회귀분석

단순회귀분석을 확장한 개념으로 **두 개 이상의 설명변수에 의하여 영향을 받는 경우가 많이 있는데,** 이러한 경우에 여러 개의 설명변수를 잘 선택하여 이들의 함수로서 반응변수의 변화를 설명하면 하나의 설명변수를 가정한 단순회귀보다 나은 분석을 할 수 있다.

1) 중회귀모형 : 설명변수가 k개 있는 중회귀모형

$$Y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + e_i, \quad i = 1, \cdots, n$$

$\beta_0, \beta_1, \cdots . \beta_k$: 회귀모수(모회귀계수)

$x_{1i}, \cdots, x_{ki}$: 설명변수(독립변수)

$y_1, \cdots, y_n$: 반응변수(종속변수)

$e_1, \cdots, e_n$: 서로 독립인 $N(0, \sigma^2)$ 확률변수(오차항)- 선형성, 독립성, 등분산성

- 행렬형식으로 표현 : $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$

  단, $\boldsymbol{X}$는 i번째 행이 $(1, x_{1i}, \cdots, x_{ki})$인 $n \times (k+1)$ 행렬

  $\boldsymbol{\beta}$: $(\beta_0, \beta_1, \cdots, \beta_k)'$로 된 $(k+1)$ 열벡터,

  $\boldsymbol{e}$: $(e_1, \cdots, e_n)'$으로 된 n 열벡터, $\boldsymbol{Y}$: $(Y_1, \cdots, Y_n)'$으로 된 n 열벡터

2) 회귀모수 추정방법 : 단순회귀경우와 동일 By LSE(least Square Estimator)

- $Q(\beta_0, \beta_1, ..., \beta_k) = \sum_{i=1}^{n} \left\{ Y_i - (\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki})^2 \right\}$를 최소로 하는 $\hat{\beta_0}, \hat{\beta_1}, ..., \hat{\beta_k}$을 각각 $\beta_0, \beta_1, ..., \beta_k$의 최소제곱추정량(least squares estimator) 이라 함.

  행렬형식을 이용하여 다음과 같이 최소제곱추정량을 구할 수 있음

- 최소제곱추정량과 적합된 모형

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \qquad (\boldsymbol{X}' : \boldsymbol{X}\text{의 전치행렬})$$

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

● 교재 372p ~ 373p : k=2인 경우

$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$

$e_i \sim N(0, \sigma^2)$이고 서로 독립 $(i = 1, 2, \cdots, n)$

$$\sum_{i=1}^{n} (e_i)^2 = \sum_{i=1}^{n} \{ Y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}) \}^2$$

- 예제. 8.10

사출온도(x_1)	195	179	205	204	201	184	210	209
사출시간(x_2)	57	61	60	62	61	54	58	61
강도(y)	81.4	122.2	101.7	145.2	135.9	64.8	92.1	113.8

- R 실행결과

```
lm(formula = y ~ x1 + x2)

Residuals:
       1        2        3        4        5        6        7        8
-4.41018 -6.66819 -11.26091  12.34641  12.17458   5.90680  -0.03319  -8.05532

Coefficients:
             Estimate  Std. Error t value Pr(>|t|)
(Intercept) -428.8851    98.9870  -4.333  0.00748 **
x1            -0.2338     0.3832  -0.610  0.56854
x2             9.8295     1.6232   6.056  0.00177 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.87 on 5 degrees of freedom
Multiple R-squared:  0.8876,    Adjusted R-squared:  0.8426
F-statistic: 19.74 on 2 and 5 DF,  p-value: 0.004239
```

$$\hat{y} = -428.8850 - 0.2338x_1 + 9.8295x_2$$

3) 회귀직선 유의성 검정($H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$, $H_1 : \beta_1, \ldots, \beta_k$가 모두 0은 아니다.)

(1) 총편차제곱합의 분해

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

$$SST = SSE + SSR$$

$$(n-1) = (n-k-1) + (k)$$

※ 단, 자유도의 개수에 유의

(2) 분산분석표

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

4) 결정계수와 수정결정계수(adjusted R^2) :

$$R^2 = \frac{SSR}{SST}; \qquad adjusted R^2 = 1 - \frac{n-1}{n-k-1}\frac{SSE}{SST}$$

- 예제 8.11 : 예제 8.10에서 중선형회귀모형을 적용할 때,
  이 계산 결과를 이용하여 모회귀함수의 유의성검정을 유의수준 5%에서 하고 결정계수 R^2의 값을 구하여라.

요인	제곱합	자유도	평균제곱	F 값	유의확률
회귀	4,666.7274	2	2,333.3637	19.736	0.0042
잔차	591.1514	5	118.2303		
계	5,257.8788	7			

Analysis of Variance Table

```
Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1  331.2   331.2  2.8016 0.155029
x2         1 4335.5  4335.5 36.6696 0.001771 **
Residuals  5  591.2   118.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5) 잔차분석

- 중회귀분석에서도 단순회귀분석에서와 마찬가지로 추론을 하려면 중선형회귀모형의 타당성에 대한 검토가 이루어져야 함
- 스튜던트화 잔차(studentized residual) 이용

$$\widehat{e_{st,i}} = \frac{\hat{e_i}}{\widehat{sd(e_i)}}, \quad i = 1,2,...,n \quad where \quad \hat{e_i} = y_i - \hat{y_i}$$

- **중회귀모형의 잔차분석에서는 설명변수가 여러개이므로 설명변수의 값을 가로축으로 할 수 없고 추측값 $\hat{y_i}$을 가로축, 잔차를 세로축으로 하여 $(\hat{y_1}, \widehat{e_{st,1}}), (\hat{y_2}, \widehat{e_{st,2}}), ..., (\hat{y_n}, \widehat{e_{st,n}})$의 위치를 표시한 잔차도(residual plot)을 이용 (그림 8.13 참고)**