9. 상관분석과 회귀분석

단순선형회귀모형

반응변수 y와 설명변수 x 사이에 존재하는 관계를 가장 잘 나타내는 직선을 찾고자 한다. x와 y에 대한 데이터 (x_1, y_1), $\cdots$, (x_n, y_n)을 단순선형회귀모형

$$y = \beta_0 + \beta_1 x + \epsilon$$

을 가정하여 표현하면

$$y_j = \beta_0 + \beta_1 x_j + \epsilon_j, \quad j = 1, \cdots, n$$

이 된다. 여기서 $\epsilon_j \sim IIDN(0, \sigma^2)$라고 가정한다. β_0와 β_1의 추정치를 각각 b_0와 b_1이라 하면 $x = x_j$에서의 y의 추측값은 $\hat{y}_j = b_0 + b_1 x_j$이므로

$$e_j \equiv y_j - \hat{y}_j$$

은 실제값과 추측값의 차이를 나타내는데 이를 잔차(residual)라 한다. 최소제곱법은 잔차의 제곱합이 최소가 되도록 b_0와 b_1의 값을 정해주는 방법이다. 잔차의 제곱합

$$Q \equiv \sum_{j=1}^{n} e_j^2 = \sum_{j=1}^{n} (y_j - b_0 - b_1 x_j)^2$$

을 최소화하는 b_0와 b_1의 값은 두 방정식 (정규방정식이라 함)

$$\frac{\partial Q}{\partial b_0} = 0 \Rightarrow -2\sum_{j=1}^{n}(y_j - b_0 - b_1 x_j) = 0 \quad (\sum_{j=1}^{n} e_j = 0)$$

$$\frac{\partial Q}{\partial b_1} = 0 \Rightarrow -2\sum_{j=1}^{n} x_j (y_j - b_0 - b_1 x_j) = 0 \quad (\sum_{j=1}^{n} x_j e_j = 0)$$

을 연립시켜 풀면 되는데, 그 해는 다음과 같다.

$$b_0 = \overline{y} - b_1 \overline{x}, \quad b_1 = \frac{S_{xy}}{S_{xx}} \quad (\text{단, } S_{xx} = \sum_{j=1}^{n}(x_j - \overline{x})^2,\ S_{xy} = \sum_{j=1}^{n}(x_j - \overline{x})(y_j - \overline{y}))$$

b_0와 b_1을 각각 β_0와 β_1에 대한 최소제곱추정량이라 부른다. 이들은 모두 불편추정량임을 보일 수 있다. 한편 σ^2의 불편추정량은

$$\hat{\sigma}^2 = \frac{SS_E}{n-2} \equiv MS_E \quad (\text{단, } SS_E = \sum_{j=1}^{n} e_j^2)$$

임을 보일 수 있다.

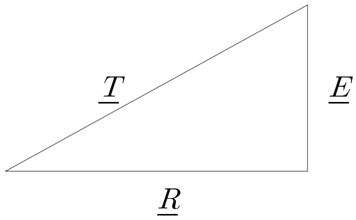
- 유의성 검증

두 변수 x와 y간에 회귀관계가 존재하지 않으면 $\beta_1 = 0$이 되고, 따라서 축소모형 $y_j = \beta_0 + \epsilon_j$ (즉, $H_0 : \beta_1 = 0$) 하에서의 β_0의 최소제곱추정량은 $b_0 = \overline{y}$가 된다. 이때 $\tilde{y}_j = \overline{y}$, $j = 1, \cdots, n$ 이므로 다음 관계가 성립한다.

$$y_j - \overline{y} = (\hat{y}_j - \overline{y}) + (y_j - \hat{y}_j)$$

$\uparrow$ $\quad$ $\uparrow$ $\quad$ $\uparrow$

H_0하에서의 두 잔차의 회귀모형하에서의
잔차 차이 잔차



$\underline{T} \;=\; \underline{R} \;+\; \underline{E}$ $\leftarrow$ n차원 벡터

$|\underline{T}|^2 \;=\; |\underline{R}|^2 \;+\; |\underline{E}|^2$

$$(\because\ \underline{R}'\underline{E} = \sum_j (\hat{y}_j - \overline{y})(y_j - \hat{y}_j) = \sum_j \hat{y}_j e_j - \overline{y}\sum_j e_j = b_0 \sum_j e_j + b_1 \sum_j x_j e_j = 0)$$

$SS_T \;=\; SS_R \;+\; SS_E$

$n-1 \qquad 1 \qquad n-2$ $\leftarrow$ 자유도

SS_T에 비하여 SS_E가 작을수록, 즉 SS_E에 비하여 SS_R이 클수록 $H_0: \beta_1 = 0$가 틀릴 가능성이 높다. $SS_E/\sigma^2 \sim \chi^2(n-2)$이고, H_0하에서 $SS_R/\sigma^2 \sim \chi^2(1)$이며 SS_E와 SS_R이 서로 독립인 사실을 이용하면 H_0가 참일 때

$$F_0 = \frac{(SS_R/\sigma^2)/1}{(SS_E/\sigma^2)/(n-2)} = \frac{SS_R/1}{SS_E/(n-2)} \equiv \frac{MS_R}{MS_E} \sim F(1,\, n-2)$$

이므로 유의수준 α에서의 기각역은 다음과 같다.

$$F_0 > F_{1,\, n-2,\, \alpha}$$

<u>다중선형회귀모형</u>

일반적 형태: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$

예) $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \epsilon$ (다항회귀모형)

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon$ (2차반응표면모형)

- 모형 적합

$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_k x_{kj} + \epsilon_j,\ j = 1, \cdots, n$ $(\epsilon_j \sim IIDN(0, \sigma^2))$(표 9.2)

$\Leftrightarrow\ \underline{y} = X\underline{\beta} + \underline{\epsilon}$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

최소제곱법: $\underline{y} = X\underline{b} + \underline{e}$에서 잔차제곱합 $\sum_{j=1}^{n} e_j^2 = \underline{e}'\underline{e}$가 최소가 되도록 $\underline{b}$를 결정

$$Q \equiv \underline{e}'\underline{e} = (\underline{y} - X\underline{b})'(\underline{y} - X\underline{b}) = \underline{y}'\underline{y} - 2\underline{b}'X'\underline{y} + \underline{b}'X'X\underline{b}$$

$$\frac{\partial Q}{\partial \underline{b}} = -2X'\underline{y} + 2X'X\underline{b} = \underline{0}\ \ (\text{참고}:\ \frac{\partial(\underline{a}'\underline{x})}{\partial \underline{x}} = \underline{a},\ \ \frac{\partial(\underline{x}'A\underline{x})}{\partial \underline{x}} = 2A\underline{x},\ A\text{는 대칭})$$

$\Rightarrow X'X\underline{b} = X'\underline{y}$ (정규방정식)

$$\underline{b} = (X'X)^{-1}X'\underline{y} \quad (Cov(\underline{b}) = \sigma^2(X'X)^{-1})$$

$$\hat{\sigma}^2 = \frac{\underline{e}'\underline{e}}{n-k-1} = \frac{SS_E}{n-k-1}$$

- **유의성 검증**

1) $H_0 : \beta_1 = \cdots = \beta_k = 0$ 대 H_1 :적어도 하나의 i에 대하여 $\beta_i \neq 0$

$$\sum_j (y_j - \overline{y})^2 = \sum_j (\hat{y}_j - \overline{y})^2 + \sum_j (y_j - \hat{y}_j)^2$$

$$SS_T \quad = \quad SS_R \quad + \quad SS_E$$

자유도 $n-1$ $\qquad k \qquad n-k-1$

$$SS_T = \underline{y}'\underline{y} - n\overline{y}^2, \; SS_R = \underline{b}'X'\underline{y} - n\overline{y}^2, \; SS_E = \underline{y}'\underline{y} - \underline{b}'X'\underline{y}$$

(여기서 $SS_E = \underline{e}'\underline{e} = \underline{y}'\underline{y} - 2\underline{b}'X'\underline{y} + \underline{b}'X'X\underline{b} = \underline{y}'\underline{y} - \underline{b}'X'\underline{y}$)

특히 단순선형회귀모형의 경우

$$SS_E = \underline{y}'\underline{y} - \underline{b}'X'\underline{y} = \sum_j y_j^2 - (b_0, b_1)(\sum_j y_j, \sum_j x_j y_j)'$$

$$= \sum_j y_j^2 - b_0 \sum_j y_j - b_1 \sum_j x_j y_j = \sum_j y_j^2 - (\overline{y} - b_1 \overline{x})(n\overline{y}) - b_1 \sum_j x_j y_j$$

$$= \sum_j y_j^2 - n\overline{y}^2 + b_1 n \overline{x}\,\overline{y} - b_1 \sum_j x_j y_j$$

$$= \sum_j (y_j - \overline{y})^2 - b_1 \sum_j (x_j - \overline{x})(y_j - \overline{y}) = S_{yy} - b_1 S_{xy} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

(단, $S_{yy} \equiv \sum_j (y_j - \overline{y})^2$)

$$SS_R = SS_T - SS_E = \sum_j (y_j - \overline{y})^2 - (S_{yy} - \frac{S_{xy}^2}{S_{xx}}) = \frac{S_{xy}^2}{S_{xx}}$$

$SS_E/\sigma^2 \sim \chi^2(n-k-1)$이고, H_0하에서 $SS_R/\sigma^2 \sim \chi^2(k)$이며 SS_E와 SS_R이 서로 독립인 사실을 이용하면 H_0가 참일 때

$$F_0 = \frac{(SS_R/\sigma^2)/k}{(SS_E/\sigma^2)/(n-k-1)} = \frac{SS_R/k}{SS_E/(n-k-1)} \equiv \frac{MS_R}{MS_E} \sim F(k, n-k-1)$$

이므로 유의수준 α에서의 기각역은 다음과 같다.

$$F_0 > F_{k,\, n-k-1,\, \alpha}$$

$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$ (설명변수의 수가 증가함에 따라 R^2는 항상 증가함)

$$R_{adj}^2 = 1 - \frac{SS_E/(n-k-1)}{SS_T/(n-1)}$$

2) $H_0 : \beta_i = 0$ 대 $H_1 : \beta_i \neq 0$

$$t = \frac{b_i}{\sqrt{C_{ii} \cdot MS_E}} \quad |t| > t_{n-k-1,\,\alpha/2}\text{이면 } H_0\text{를 기각}$$

$\nwarrow (X'X)^{-1}$의 $(i+1)$번째 대각원소

3) $\underline{\beta} = \begin{pmatrix} \underline{\beta}_1 \\ \underline{\beta}_2 \end{pmatrix} \begin{matrix} r \times 1 \\ (p-r) \times 1 \end{matrix}$ (절편이 있을 경우 $p = k+1$, 없을 경우 $p = k$)

$X = (X_1 \ \vdots \ X_2)$

$H_0 : \underline{\beta}_1 = \underline{0}$ 대 $H_1 : \underline{\beta}_1 \neq \underline{0}$

$\underline{y} = X\underline{\beta} + \epsilon = X_1 \underline{\beta}_1 + X_2 \underline{\beta}_2 + \underline{\epsilon}$ (완전모형)

$\underline{y} = X_2 \underline{\beta}_2 + \underline{\epsilon}$ (축소모형)

완전모형 하에서

$$SS_R(\underline{\beta}) = \underline{b}' X' \underline{y} \ \left(= \sum_j \hat{y}_j^2 \neq \sum_j (\hat{y}_j - \overline{y})^2 = \sum_j \hat{y}_j^2 - n\overline{y}^2\right)$$

$$MS_E = \frac{\underline{y}'\underline{y} - \underline{b}'X'\underline{y}}{n-p}$$

축소모형 하에서

$SS_R(\underline{\beta}_2) = \underline{b}_2' X_2' \underline{y}$ (단, $\underline{b}_2 = (X_2'X_2)^{-1} X_2' \underline{y}$)

이므로

$$SS_R(\underline{\beta}_1 | \underline{\beta}_2) = SS_R(\underline{\beta}) - SS_R(\underline{\beta}_2) = \underline{b}'X'\underline{y} - \underline{b}_2' X_2' \underline{y}$$

추가회귀제곱합 $= (\underline{y}'\underline{y} - \underline{b}_2'X_2'\underline{y}) - (\underline{y}'\underline{y} - \underline{b}'X'\underline{y}) = SS_E(\underline{\beta_2}) - SS_E(\underline{\beta})$

$$F_0 = \frac{SS_R(\underline{\beta}_1 | \underline{\beta}_2)/r}{MS_E} > F_{r,\,n-p,\,\alpha} \Rightarrow H_0\text{를 기각}$$

예) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon$

$H_0 : \beta_{11} = \beta_{22} = \beta_{12} = 0$

$$\underline{\beta}_1 = \begin{pmatrix} \beta_{11} \\ \beta_{22} \\ \beta_{12} \end{pmatrix} \quad \underline{\beta}_2 = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

($H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$의 경우는 $\underline{\beta}_1 = (\beta_1, \cdots, \beta_k)'$, $\underline{\beta}_2 = \beta_0$이고

$SS_R = SS_R(\beta_1, \cdots, \beta_k | \beta_0) = SS_R(\beta_0, \beta_1, \cdots, \beta_k) - SS_R(\beta_0) = \underline{b}'X'\underline{y} - n\overline{y}^2$)

(예 9.2), (예 9.3)

H.W. #1

1. A 건설사는 24번의 과거 계약에서 계약규모(x:억원)와 단위계약금액당 순이익(y:백만원)간에 어떤 관계가 있는지 알고 싶다. 다음의 데이터를 사용하여 y와 x간에 1차 모형이 적합한지 아니면 2차 또는 그 이상의 모형이 적합한지를 설명하시오.

x	1	1	2	2	2	3	3	4	4	4	5	5
y	5	6	6	7	8	7	8	7	8	9	8	9

x	5	6	6	6	7	7	8	8	8	9	9	10
y	10	8	9	10	8	9	7	8	9	7	8	7

2. 다음 자료는 소나무의 평균 직경(y)과 세 개의 설명변수, 즉 나이(x_1), 높이(x_2), 1 에이커 당 그루수(x_3) 값이다. 이 자료를 사용하여 가장 적합한 선형회귀모형을 찾으시오.

x_1	x_2	x_3	y
19	51.5	500	7.0
14	41.3	900	5.0
11	36.7	650	6.2
13	32.2	480	5.2
13	39.0	520	6.2
12	29.8	610	5.2
18	51.2	700	6.2
14	46.8	760	6.4
20	61.8	930	6.4
17	55.8	690	6.4
13	37.3	800	5.4
21	54.2	650	6.4
11	32.5	530	5.4
19	56.3	680	6.7
17	52.8	620	6.7
15	47.0	900	5.9
16	53.0	620	6.9
16	50.3	730	6.9
14	50.5	680	6.9
22	57.7	480	7.9

10. 공분산분석

반응변수의 값이 실험자가 통제할 수 없는 다른 설명변수와 선형관계를 가지며 변화할 때가 있는데, 이때 이 설명변수의 영향을 보정해 주지 않으면 오차제곱합이 커져서 처리효과를 찾아내기가 힘들어진다. (표 10.1 ~ 표 10.3)

<u>일원배치법 공분산분석모형</u>

$$y_{ij} = \mu' + \tau_i + \beta x_{ij} + \epsilon_{ij}, \quad i = 1, \cdots, t; j = 1, \cdots, n \ (\sum_i \tau_i = 0)$$

처리효과 회귀계수 공변수(covariate): 조절불가능한 연속형 변수(cf: 블록요인)

$$\epsilon_{ij} \sim IIDN(0, \sigma^2)$$

1) $\beta = 0 \Rightarrow y_{ij} = \mu' + \tau_i + \epsilon_{ij}$: 일원배치법모형

2) $\tau_i = 0 \ \forall i \Rightarrow y_{ij} = \mu' + \beta x_{ij} + \epsilon_{ij}$: 단순선형회귀모형

$\mu' = \mu - \beta \overline{x}$로 놓으면

$$y_{ij} = \mu + \tau_i + \beta(x_{ij} - \overline{x}) + \epsilon_{ij}, \quad i = 1, \cdots, t; j = 1, \cdots, n \ (\sum_i \tau_i = 0) \leftarrow \text{완전모형}$$

<u>최소제곱추정</u>

1) 완전모형

$$\hat{\mu} = \overline{y}$$

$$\hat{\tau}_i = \overline{y}_{i.} - \overline{y} - \hat{\beta}_{full}(\overline{x}_{i.} - \overline{x})$$

$$\hat{\beta}_{full} = \frac{E_{xy}}{E_{xx}} \ (\text{단, } E_{xx} = \sum_i \sum_j (x_{ij} - \overline{x}_{i.})^2,\ E_{xy} = \sum_i \sum_j (x_{ij} - \overline{x}_{i.})(y_{ij} - \overline{y}_{i.}))$$

$$SS_E = E_{yy} - \frac{(E_{xy})^2}{E_{xx}} \ (\text{단, } E_{yy} = \sum_i \sum_j (y_{ij} - \overline{y}_{i.})^2)$$

$$SS_R = \frac{(E_{xy})^2}{E_{xx}} \ (SS_R + SS_E = E_{yy} \leftarrow \text{공변수가 없을 때의 잔차제곱합})$$

$$\hat{\sigma}^2 = MS_E = \frac{SS_E}{t(n-1)-1} \quad (\text{자유도: } tn - \underset{\mu}{\underset{\uparrow}{1}} - \underset{\beta}{\underset{\uparrow}{1}} - \underset{\tau_i}{\underset{\uparrow}{(t-1)}})$$

2) 축소모형

귀무가설 $H_0 : \tau_1 = \cdots = \tau_t = 0$이 참일 때의 모형은 다음과 같다.

$$y_{ij} = \mu + \beta(x_{ij} - \overline{x}) + \epsilon_{ij}, \ i = 1, \cdots, t; j = 1, \cdots, n \leftarrow \text{축소모형}$$

$$\hat{\mu} = \overline{y}$$

$$\hat{\beta}_{reduced} = \frac{S_{xy}}{S_{xx}} \text{ (단, } S_{xx} = \sum_i \sum_j (x_{ij} - \overline{x})^2,\ S_{xy} = \sum_i \sum_j (x_{ij} - \overline{x})(y_{ij} - \overline{y}))$$

$$SS_E{}' = S_{yy} - \frac{(S_{xy})^2}{S_{xx}} \text{ (단, } S_{yy} = \sum_i \sum_j (y_{ij} - \overline{y})^2)$$

<u>가설 검증</u>

$$SS_{Trt} = SS_E{}' - SS_E$$

$$F = \frac{SS_{Trt}/(t-1)}{MS_E} > F_{t-1,\, t(n-1)-1,\, \alpha} \Rightarrow H_0\text{를 기각 (표 10.4)}$$

($SS_T \neq SS_R + SS_{Trt} + SS_E$임에 유의할 것!)

Note: 1) 공분산분석모형은 처리수준에 관계없이 회귀직선의 기울기가 β로 같고 다만 절편이 $\mu' + \tau_i,\ i = 1, \cdots, t$로 다르다고 가정한다. 즉, 처리효과와 공변수간에 상호작용이 존재하지 않음을 가정한다. 따라서 공분산분석을 행하기에 앞서 모형

$$y_{ij} = \mu + \tau_i + \beta(x_{ij} - \overline{x}) + (\tau\beta)_i(x_{ij} - \overline{x}) + \epsilon_{ij},\ i = 1, \cdots, t;\ j = 1, \cdots, n$$

에서 $H_0 : (\tau\beta)_1 = \cdots = (\tau\beta)_t = 0$을 채택할 수 있어야 한다. 이를 기울기의 동질성검증이라 한다.

2) $H_0 : \beta = 0$이 기각되지 않으면 공변수의 필요성이 없어지므로 회귀의 유의성검증도 필요하다.

3) 보정된 처리평균: $\overline{y}_{i.(adj)} = \overline{y} + \hat{\tau}_i = \overline{y}_{i.} - \hat{\beta}_{full}(\overline{x}_{i.} - \overline{x}) \Rightarrow$ 다중비교

4) $T_{xx} = \sum_i \sum_j (\overline{x}_{i.} - \overline{x})^2,\ T_{yy} = \sum_i \sum_j (\overline{y}_{i.} - \overline{y})^2,$

$T_{xy} = \sum_i \sum_j (\overline{x}_{i.} - \overline{x})(\overline{y}_{i.} - \overline{y})$라 하면 다음 관계가 성립한다.

$$S_{yy} = \sum_i \sum_j (y_{ij} - \overline{y})^2 = \sum_i \sum_j (y_{ij} - \overline{y}_{i.})^2 + \sum_i \sum_j (\overline{y}_{i.} - \overline{y})^2 = E_{yy} + T_{yy}$$

마찬가지 방법으로 다음 관계가 성립함을 보일 수 있다.

$$S_{xx} = E_{xx} + T_{xx},\quad S_{xy} = E_{xy} + T_{xy}$$

(이 관계식을 이용하면 E_{xx}, E_{yy}, E_{xy}를 간접 계산할 수 있음.)

(예 10.1) 분석 결과: 표 10.5 (표 10.2와 비교!)

$H_0 : (\tau\beta)_1 = (\tau\beta)_2 = (\tau\beta)_3 = 0$은 채택 ($F = 0.49$) ← 출력 10.1

공분산분석모형:

$$y_{ij} = \mu + \tau_i + \beta(x_{ij} - \overline{x}) + \epsilon_{ij},\quad i = 1, 2, 3;\ j = 1, \cdots, 5\ \left(\sum_i \tau_i = 0\right)$$

$H_0: \beta = 0$은 기각. ($F=70.08$) $\hat{\beta}_{full} = 0.954$ ← 출력 10.2

$H_0: \tau_1 = \tau_2 = \tau_3 = 0$은 기각 못함. ($F=2.61$)

그림 설명: 기계별 회귀직선과 전체 회귀직선

주의: 보정된 처리평균의 비교에서 $H_0: \tau_1 = \tau_2 = \tau_3 = 0$이 채택되었음에도 $\bar{y}_{2.(adj)}$와 $\bar{y}_{3.(adj)}$가 유의하게 다르다고 결론이 난 이유는?

<u>기타 공분산분석모형</u>

공변수의 도입은 위에서 고려한 일원배치법 뿐만 아니라 어떤 종류의 분산분석모형에도 가능하다. 예를 들어 랜덤화블록설계에 공변수를 도입하면 다음의 모형을 얻는다.

$$y_{ij} = \mu + \underbrace{\tau_i}_{\text{처리효과}} + \underbrace{\beta_j}_{\text{블록효과}} + \beta(x_{ij} - \bar{x}) + \underbrace{\epsilon_{ij}}_{\sim IIDN(0,\sigma^2)}, \quad i = 1, \cdots, a;\ j = 1, \cdots, b$$

이때

$$\sum_i \sum_j (y_{ij} - \bar{y})^2 = \sum_i \sum_j (\bar{y}_{i.} - \bar{y})^2 + \sum_i \sum_j (\bar{y}_{.j} - \bar{y})^2 + \sum_i \sum_j (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y})^2$$

$$S_{yy} = T_{yy} + R_{yy} + E_{yy}$$

인 관계가 성립하고 마찬가지로

$$\sum_i \sum_j (x_{ij} - \bar{x})^2 = \sum_i \sum_j (\bar{x}_{i.} - \bar{x})^2 + \sum_i \sum_j (\bar{x}_{.j} - \bar{x})^2 + \sum_i \sum_j (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x})^2$$

$$S_{xx} = T_{xx} + R_{xx} + E_{xx}$$

$$\sum_i \sum_j (x_{ij} - \bar{x})(y_{ij} - \bar{y}) = \sum_i \sum_j (\bar{x}_{i.} - \bar{x})(\bar{y}_{i.} - \bar{y}) + \sum_i \sum_j (\bar{x}_{.j} - \bar{x})(\bar{y}_{.j} - \bar{y})$$
$$+ \sum_i \sum_j (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x})(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y})$$

$$S_{xy} = T_{xy} + R_{xy} + E_{xy}$$

이 성립한다. 위의 완전모형 하에서의 추정치는

$$\hat{\beta}_{full} = \frac{E_{xy}}{E_{xx}}$$

$$SS_E = E_{yy} - \frac{E_{xy}^2}{E_{xx}}$$

$$\hat{\sigma}^2 = MS_E = \frac{SS_E}{ab - a - b} \quad \text{(자유도: } ab - 1 - 1 - (a-1) - (b-1)\text{)}$$

또한 축소모형 하에서의 추정치는

$$\hat{\beta}_{reduced} = \frac{S_{xy}'}{S_{xx}'} \quad \left(\text{단, } S_{xx}' = \sum_i \sum_j (x_{ij} - \bar{x}_{.j})^2 = T_{xx} + E_{xx},\ S_{xy}' = \sum_i \sum_j (x_{ij} - \bar{x}_{.j})(y_{ij} - \bar{y}_{.j}) = T_{xy} + E_{xy}\right)$$

$$SS_E{}' = S_{yy}{}' - \frac{(S_{xy}{}')^2}{S_{xx}{}'} \quad (\text{단, } S_{yy}{}' = \sum_i \sum_j (y_{ij} - \overline{y}_{.j})^2 = T_{yy} + E_{yy})$$

이 된다. 따라서 $H_0 : \tau_1 = \cdots = \tau_a = 0$ 은

$$SS_{Trt} = SS_E{}' - SS_E$$

라 정의할 때

$$F = \frac{SS_{Trt}/(a-1)}{MS_E} > F_{a-1,\, ab-a-b,\, \alpha}$$

이면 유의수준 α에서 기각된다. 이때에도 처리효과의 검증에 앞서 기울기의 동질성 검증과 회귀의 유의성 검증이 필요한 것은 물론이다.

공분산분석모형의 처리효과 검증을 위한 통계량을 보다 쉽고 체계적으로 구할 수 있도록 도표화하면 다음과 같다.

1) 일원배치법

source	SS_y	SS_x	SS_{xy}
처리	T_{yy}	T_{xx}	T_{xy}
오차	E_{yy}	E_{xx}	E_{xy}
총	S_{yy}	S_{xx}	S_{xy}

$$SS_E{}' = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

$$SS_E = E_{yy} - \frac{E_{xy}^2}{E_{xx}}$$

$$SS_{Trt} = SS_E{}' - SS_E$$

2) 랜덤화블록설계

source	SS_y	SS_x	SS_{xy}
블록	R_{yy}	R_{xx}	R_{xy}
처리	T_{yy}	T_{xx}	T_{xy}
오차	E_{yy}	E_{xx}	E_{xy}
총	S_{yy}	S_{xx}	S_{xy}

$$SS_E{}' = T_{yy} + E_{yy} - \frac{(T_{xy} + E_{xy})^2}{T_{xx} + E_{xx}}$$

$$SS_E = E_{yy} - \frac{E_{xy}^2}{E_{xx}}$$

$$SS_{Trt} = SS_E{}' - SS_E$$

3) 이원배치법

모형: $y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$

분산분석: $y_{ijk} - \overline{y} = (\overline{y}_{i..} - \overline{y}) + (\overline{y}_{.j.} - \overline{y}) + (\overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}) + (y_{ijk} - \overline{y}_{ij.})$

$$S_{yy} \quad = \quad A_{yy} \quad + \quad B_{yy} \quad + \quad (A\times B)_{yy} \quad + \quad E_{yy}$$

source	SS_y	SS_x	SS_{xy}
A	A_{yy}	A_{xx}	A_{xy}
B	B_{yy}	B_{xx}	B_{xy}
$A\times B$	$(A\times B)_{yy}$	$(A\times B)_{xx}$	$(A\times B)_{xy}$
오차	E_{yy}	E_{xx}	E_{xy}
총	S_{yy}	S_{xx}	S_{xy}

예를 들어 $H_0 : \alpha_1 = \cdots = \alpha_a = 0$을 검정하려면

$$SS_E{}' = A_{yy} + E_{yy} - \frac{(A_{xy} + E_{xy})^2}{A_{xx} + E_{xx}}$$

$$SS_E = E_{yy} - \frac{E_{xy}^2}{E_{xx}}$$

$$SS_A = SS_E{}' - SS_E$$

를 계산하면 되고 마찬가지 방법으로 $H_0 : \beta_1 = \cdots = \beta_b = 0$과 $H_0 : (\alpha\beta)_{ij} = 0$ $\forall i, j$도 검정할 수 있다.

(예 10.1a) 다음은 품종(A)과 급수량(B)에 따른 꽃의 수확량(y) 자료이다. 재배면적의 크기(x)에 따라 수확량에 차이가 있을 것이므로 공분산분석모형을 사용하여 두 요인과 상호작용에 대한 검정을 실시해 보자.

	B_1		B_2			B_1		B_2	
	y	x	y	x		y	x	y	x
A_1	98	15	71	10	A_2	55	4	76	11
	60	4	80	12		60	5	68	10
	77	7	86	14		75	8	43	2
	80	9	82	13		65	7	47	3
	95	14	46	2		87	13	62	7
	64	5	55	3		78	11	70	9

	B_1		B_2		합계	
	y	x	y	x	y	x
A_1	474	54	420	54	894	108
A_2	420	48	366	42	786	90
합계	894	102	786	96	1680	198

위의 부분합 표를 이용하여 y에 관한 요인별 제곱합을 구하면 다음과

같다.

$$S_{yy} = \sum_i \sum_j \sum_k y_{ijk}^2 - CT_y = 122686 - 117600 = 5086$$
$$A_{yy} = \sum_i y_{i..}^2 / br - CT_y = [(894)^2 + (786)^2]/12 - 117600 = 486$$
$$B_{yy} = \sum_j y_{.j.}^2 / ar - CT_y = [(894)^2 + (786)^2]/12 - 117600 = 486$$
$$AB_{yy} = \sum_i \sum_j y_{ij.}^2 / r - CT_y = [(474)^2 + (420)^2 + (420)^2 + (366)^2]/6 - 117600 = 972$$
$$(A \times B)_{yy} = AB_{yy} - A_{yy} - B_{yy} = 972 - 486 - 486 = 0$$
$$E_{yy} = S_{yy} - AB_{yy} = 5086 - 972 = 4114$$
$$(S_{xy} = \sum_i \sum_j \sum_k x_{ijk} y_{ijk} - CT_{xy} \quad (CT_{xy} = (GT_x \cdot GT_y)/abr)$$
$$A_{xy} = \sum_i x_{i..} y_{i..} / br - CT_{xy} \qquad B_{xy} = \sum_j x_{.j.} y_{.j.} / ar - CT_{xy}$$
$$AB_{xy} = \sum_i \sum_j x_{ij.} y_{ij.} / r - CT_{xy} \qquad (A \times B)_{xy} = AB_{xy} - A_{xy} - B_{xy}$$
$$E_{xy} = S_{xy} - AB_{xy})$$

마찬가지 방법으로 x와 xy에 관한 요인별 제곱합을 구하여 요약하면 다음과 같다.

	SS_y	SS_x	SS_{xy}
A	486	13.5	81
B	486	1.5	27
$A \times B$	0	1.5	0
E	4114	372	1219
S	5086	388.5	1327

1) 요인 A에 대한 검정

$$SS_E = E_{yy} - \frac{E_{xy}^2}{E_{xx}} = 4114 - \frac{(1219)^2}{372} = 119.481$$
$$SS_E' = A_{yy} + E_{yy} - \frac{(A_{xy} + E_{xy})^2}{A_{xx} + E_{xx}} = 486 + 4114 - \frac{(81 + 1219)^2}{13.5 + 372} = 216.083$$
$$SS_A = SS_E' - SS_E = 96.602$$
$$F_0 = \frac{96.602}{6.2885} = 15.36 \ (MS_E = 119.481/19 = 6.2885)$$

2) 요인 B에 대한 검정

$$SS_E' = 486 + 4114 - \frac{(27 + 1219)^2}{1.5 + 372} = 443.331$$

$$SS_B = SS_E' - SS_E = 323.850$$

$$F_0 = \frac{323.850}{6.2885} = 51.50$$

3) 상호작용 A×B에 대한 검정

$$SS_E' = 0 + 4114 - \frac{(0+1219)^2}{1.5+372} = 135.523$$

$$SS_{A\times B} = SS_E' - SS_E = 16.042$$

$$F_0 = \frac{16.042}{6.2885} = 2.55$$

H.W. #2

p.259 1 (랜덤화블록설계로 바꿔서 분석하시오.), 2