정보통계학과 **백재욱** 교수

제6강 (6장)

회귀분석과 공분산분석

6.1 회귀분석

6.2 공분산분석

제6강 회귀분석과 공분산분석

6.1 회귀분석

6.1 회귀분석

- 독립변수(들)와 종속변수 간의 관계를 함수식으로 표현하여 살펴보는 분석방법
- 독립변수와 종속변수는 연속적인 값을 취함
- **단순회귀** : 독립변수의 수가 1개인 경우
- **다중회귀** : 독립변수의 수가 2개 이상인 경우

6.1 회귀분석

◆ 다중선형회귀

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon, \boldsymbol{k \geq 2} \quad \text{........................}(6.1)$$

◆ 선형회귀모형의 또 다른 예: 다항회귀모형(polynomial regression model)

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \varepsilon \quad \text{..................................}(6.2)$$

◆ 독립변수가 불연속적인 값을 취하며

- 요인의 수가 1개인 경우 **일원배치**
- 요인의 수가 2개인 경우 **이원배치**

비선형회귀 예

$$y = \beta_0 (1 - e^{-\beta_1 x}) + \epsilon$$

$$y = \frac{e^{\beta_1 x}}{1 + e^{\beta_1 x}} + \epsilon$$

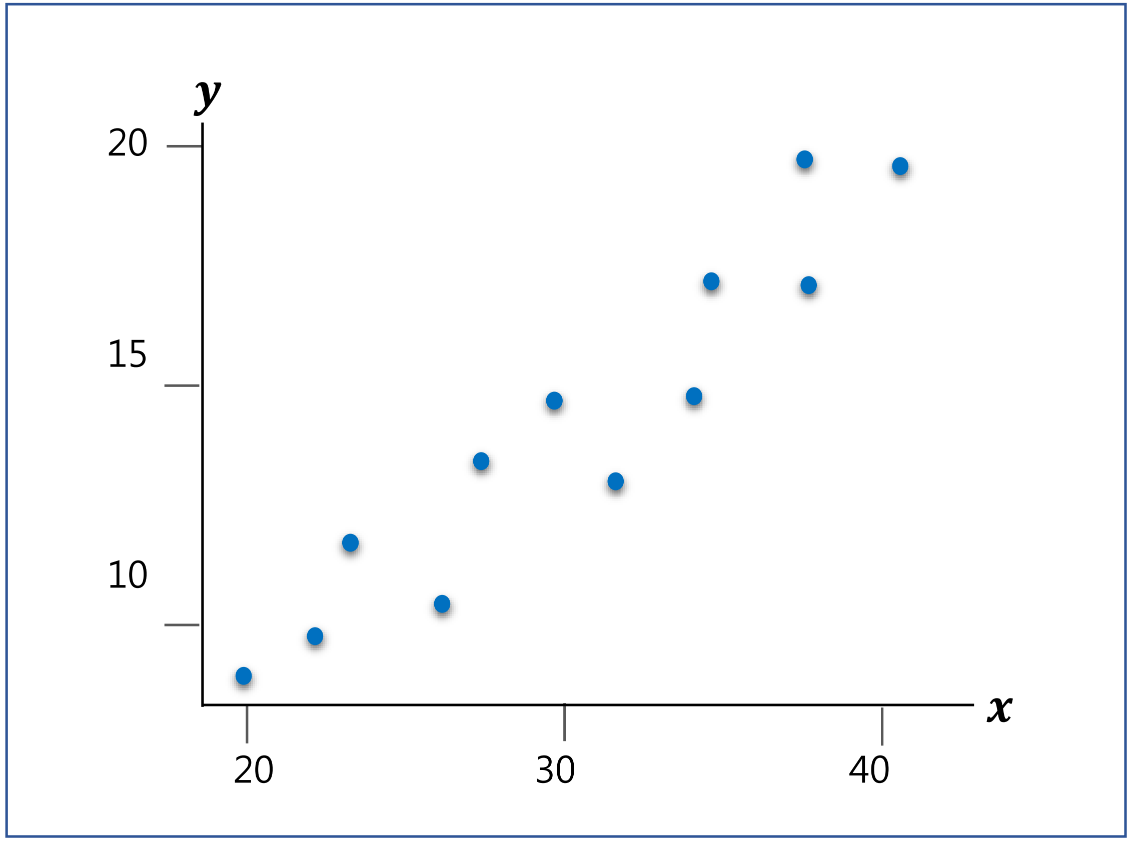
6.1 회귀분석

◆ 산점도 (scatter plot)

(예) 장치의 회전률(x)과 그때 야기되는 페인트의 불순 퍼센티지(y) 측정 데이터

<표 6-1> 페인트 자료

x (rpm)	20	22	24	26	28	30
y (%)	8.4	9.5	11.8	10.4	13.3	14.8
x (rpm)	32	34	36	38	40	42
y (%)	13.2	14.7	16.4	16.5	18.9	18.5



[그림 6-1] 페인트 자료에 대한 산점도

6.1 회귀분석

◆ 상관계수

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad , \quad -1 \le r \le 1$$

- 1에 가까운 r => 양(+)의 상관관계 높음
- -1에 가까운 r => 음(-)의 상관관계 높음
- 0에 가까운 r => 두 변수간 상관관계 없음
- 주의 : r 은 두 변수(변량) 간 <u>직선적</u> 상관관계를 측정함

6.1 회귀분석

◆ 단순선형회귀분석

- 모형

$$y = \beta_0 + \beta_1 x + \varepsilon \quad \cdots\cdots\cdots \quad (6.3)$$

ε는 오차로서 확률변수로 $N(0, \sigma^2)$을 따름

- 두 변수 (x,y) 에 대한 n개의 관측 (x_1, y_1), (x_2, y_2), $\cdots$, (x_n, y_n)

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1,2,\cdots,n \quad \cdots\cdots\cdots \quad (6.4)$$

추정식 : $\hat{y}_i = b_0 + b_1 x_i, \; i = 1,2,\cdots,n$ ⋯⋯⋯ (6.6)

잔차 (residual) : $e_i = y_i - \hat{y}_i, \; i = 1,2,\cdots,n$ ⋯⋯⋯ (6.7)

6.1 회귀분석

◆ 최소제곱법 (least squares method)

$$\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$$

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 = Q \quad \cdots\cdots\cdots\cdots\cdots (6.8)$$

$$\frac{\partial Q}{\partial \beta_0} = 0 \Leftrightarrow \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)] = 0$$

$$\cdots\cdots\cdots\cdots\cdots (6.9)$$

$$\frac{\partial Q}{\partial \beta_1} = 0 \Leftrightarrow \sum_{i=1}^{n} x_i [y_i - (\beta_0 + \beta_1 x_i)] = 0$$

$$b_1 = \frac{S_{xy}}{S_{xx}}, \ b_0 = \bar{y} - b_1 \bar{x} \quad \cdots\cdots\cdots\cdots\cdots (6.10)$$

6.1 회귀분석

◆ 최소제곱법 (least squares method)

$$\bar{y} = \sum_{i=1}^{n} y_i \, / \, n \, , \qquad \bar{x} = \sum_{i=1}^{n} x_i \, / \, n$$

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

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

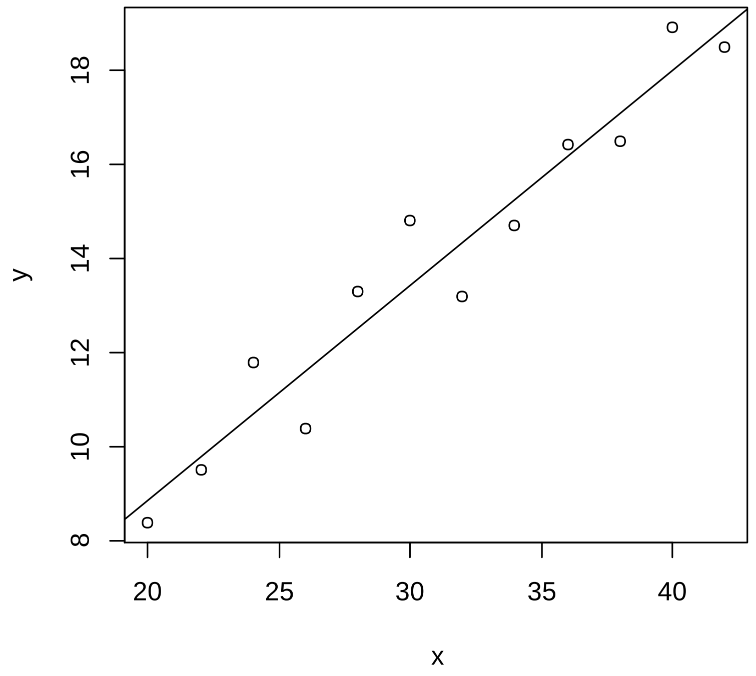
6.1 회귀분석

예 6.1 페인트 자료에 대해 단순선형회귀분석을 실시하라.

<표 6-1> 페인트 자료

x (rpm)	20	22	24	26	28	30
y (%)	8.4	9.5	11.8	10.4	13.3	14.8
x (rpm)	32	34	36	38	40	142
y (%)	13.2	14.7	16.4	16.5	18.9	18.5

풀이



[그림 6-1] 페인트 자료에 대한 산점도 및 회귀직선

6.1 회귀분석

풀이 (계속)

$$n = 12, \qquad \bar{x} = 31, \qquad \bar{y} = 13.86667$$

$$S_{xx} = 572, \qquad S_{xy} = 261.2$$

$$\widehat{\beta_1} = b_1 = \frac{S_{xy}}{S_{xx}} = \frac{261.2}{572} = 0.45664$$

$$\widehat{\beta_0} = b_0 = \bar{y} - b_1\bar{x} = 13.86667 - 0.45664 \times 31 = -0.28928$$

$$\hat{y} = -0.28928 + 0.45664x$$

x=30일 때 y=14.8

$\hat{y}$= −0.2879 + 0.4566(30) = 13.41

잔차 $e = y - \hat{y} = 14.8 - 13.41 = 1.39$

6.1 회귀분석

R 실습

```
x = c(20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42)
y = c(8.4, 9.5, 11.8, 10.4, 13.3, 14.8, 13.2, 14.7, 16.4, 16.5, 18.9, 18.5)
lm.out = lm(y~x)
plot(x, y)
abline(lm.out)
with(paint, cor.test(y, x))
```

Pearson's product-moment correlation

data: y and x

t = 11.8798, df = 10, p-value = 3.211e-07

alternative hypothesis: true correlation is not equal to 0

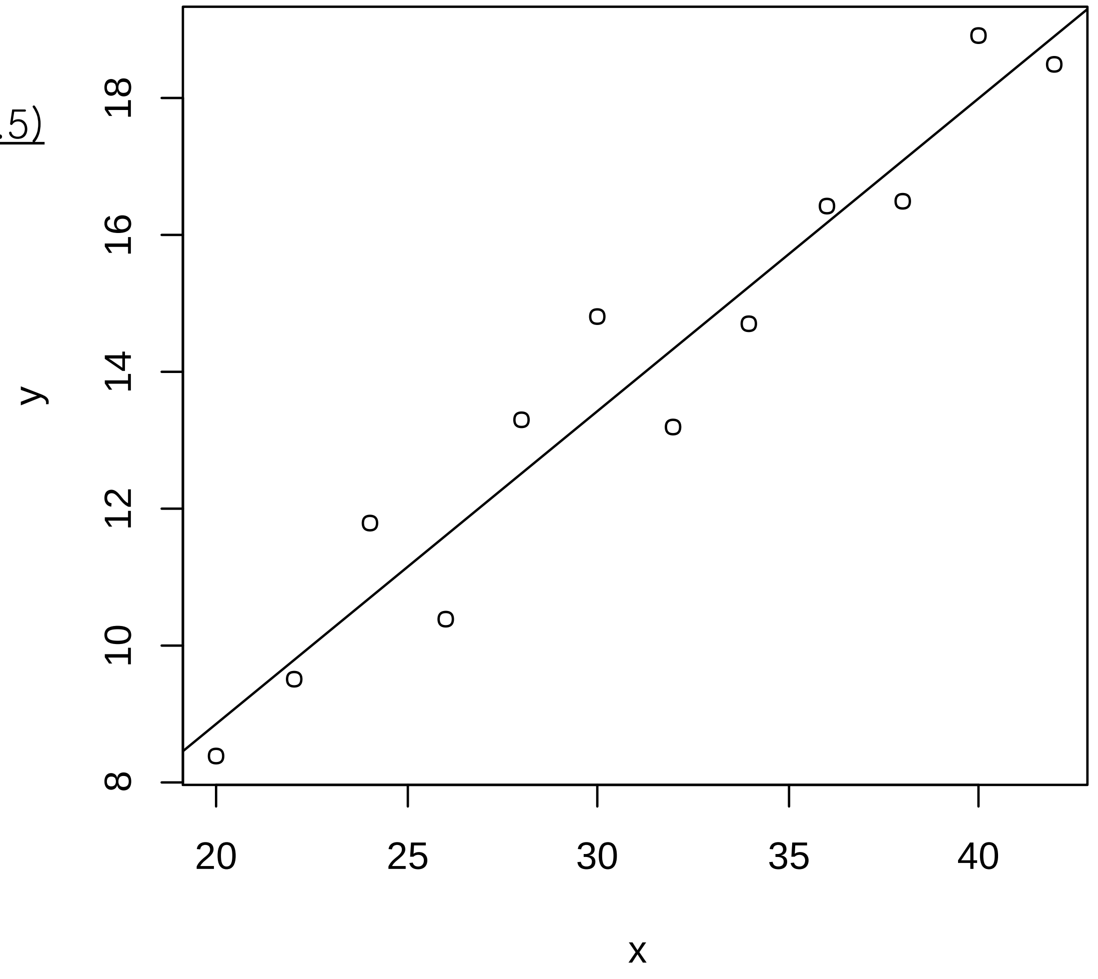
95 percent confidence interval:

0.8810937 0.9907768

sample estimates:

cor

0.9663498



6.1 회귀분석

R 실습

summary(lm.out)

```
Call:
lm(formula = y ~ x)
Residuals:
    Min       1Q   Median     3Q     Max
-1.1834 -0.5432 -0.3233 0.8333 1.3900
Coefficients:
             Estimate  Std. Error  t value   Pr(>|t|)
(Intercept) -0.28928  1.22079  -0.237    0.817
x             0.45664  0.03844  11.880    3.21e-07
(Intercept)
x ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.9193 on 10 degrees of freedom
Multiple R-squared: 0.9338, Adjusted R-squared: 0.9272
F-statistic: 141.1 on 1 and 10 DF, p-value: 3.211e-07
```

6.1 회귀분석

◆ 회귀선의 유의성 검정

- 두 변수 사이에 회귀관계가 없다면 β_1의 값은 0이 되어 다음 관계식이 적절함

$$y = \beta_0 + \varepsilon \cdots\cdots (6.11)$$

- 단순선형회귀에서 **총편차의 분해**

$$y - \bar{y} = (y - \hat{y}) + (\hat{y} - \bar{y}) \cdots\cdots (6.12)$$

실제값과 평균의 차이 = 실제값과 적합치의 차이 + 적합치와 평균과의 차이

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

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

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

⇔ 총제곱합 = 잔차제곱합 + 회귀제곱합

자유도: $n - 1 = (n-2) + 1$

$\phi_T = \phi_E + \phi_R$

- 제곱합

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

$$SS_R = \sum_{i-1}^{n}(\hat{y} - \bar{y})^2 = b_1 S_{xy} (= b_1^2 S_{xx})$$

$$SS_E = \sum_{i-1}^{n}(y_i - \hat{y})^2 = S_{yy} - b_1 S_{xy}$$

6.1 회귀분석

- **결정계수(coefficient of determination) $\boldsymbol{R^2}$**

$$R^2 = \frac{SS_R}{SS_T}, \quad 0 \le R^2 \le 1 \quad (\text{상관계수 } r\text{의 제곱})$$

- 회귀계수 $\boldsymbol{\beta_1}$의 유의성

H_0: 회귀 관계가 없다. ⟷ $H_0: \beta_1 = 0$

H_1: 회귀 관계가 있다. ⟷ $H_1: \beta_1 \neq 0$

6.1 회귀분석

〈표 6-2〉 분산분석표: 회귀의 유의성 검정

변인	제곱합	자유도	평균제곱	EMS	F_0
회귀	$SS_R = b_1 S_{xy}$	1	MS_R	$\sigma^2 + \beta_1^2 S_{xx}$	$MS_R \;/\; MS_E$
오차	$SS_E = S_{yy} - b_1 S_x$	$n-2$	MS_E	σ^2	
총	$SS_T = S_{yy}$	$n-1$			

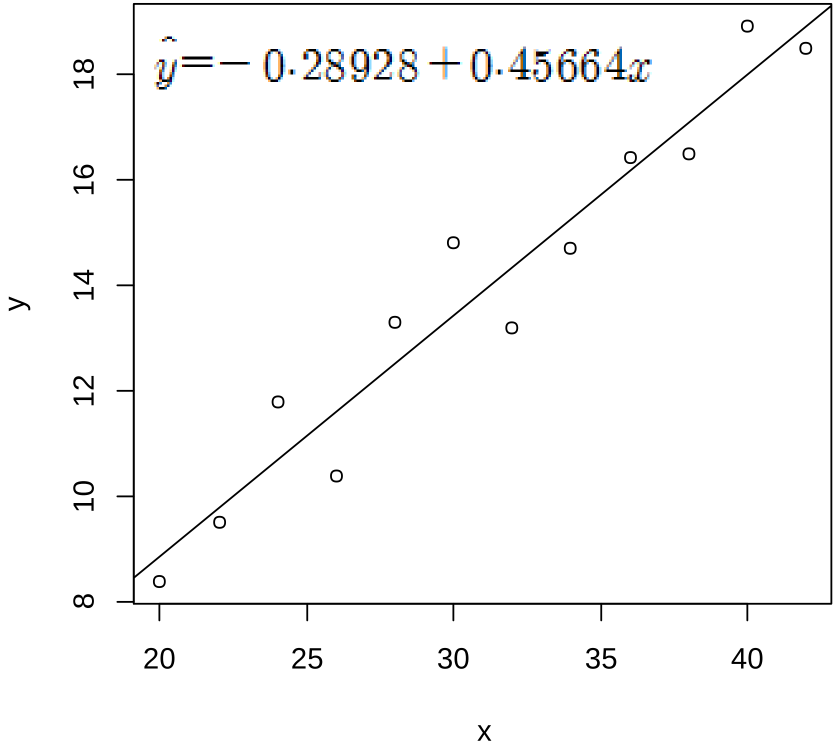
검정통계량 $F_0 = \dfrac{SS_R/1}{SS_E/(n-2)} = \dfrac{MS_R}{MS_E}$

$F_0 > F(1, n-2; \alpha)$이면 귀무가설 기각

6.1 회귀분석

예 6.2 페인트 자료에 대해 회귀선의 유의성을 검정해보라.

풀이



[그림 6-1] 페인트 자료에 대한 산점도 및 회귀직선

<표 6-3> 페인트 자료에 대한 분산분석표: 회귀의 유의성 검정

변인	제곱합	자유도	평균제곱	F_0
회귀	119.274	1	119.274	140.99
오차	8.459	10	0.846	
총	127.733	11		

6.1 회귀분석

풀이 R실습

anova(lm.out)

```
Analysis of Variance Table
Response: y
Df Sum Sq Mean Sq F value Pr(>F)
x 1 119.275 119.275 141.13 3.211e-07 ***
Residuals 10 8.451 0.845
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

6.1 회귀분석

신뢰구간

β_1 에 대한 100(1−α)% 신뢰구간:

$$\left[b_1 \pm t(n-2; a/2) \sqrt{\frac{MS_E}{S_{xx}}} \right]$$

$$\left[0.4566 \pm (2.228) \sqrt{\frac{0.846}{572.0}} \right] = [0.4566 \pm 0.0857] = (0.3709,\ 0.5423)$$

$x = x_1$ 에서 E_y에 대한 100(1−α)% 신뢰구간:

$$\left[(b_0 + b_1 x_1) \pm t(n-2; a/2) \sqrt{MS_E \left(\frac{1}{n} + \frac{(x_1 - \bar{x})^2}{S_{xx}} \right)} \right]$$

6.1 회귀분석

R 실습

```
confint(lm.out, level=0.95)

                 2.5 %      97.5 %
(Intercept)  -3.009365   2.4308107
x             0.370997   0.5422898

predict(lm.out, newdata, interval="confidence")
       fit      lwr      upr
1 13.41002 12.81254 14.00751
```

6.1 회귀분석

다중회귀분석

〈표 6-4〉 다중선형회귀에 대한 자료 구조

회귀변수				반응변수
x_1	x_2	$\cdots$	x_k	y
x_{11}	x_{21}	$\cdots$	x_{k1}	y_1
x_{12}	x_{22}	$\cdots$	x_{k2}	y_2
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_{1n}	x_{2n}	$\cdots$	x_{kn}	y_n

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2) \quad \cdots\cdots \quad (6.14)$$

정보통계학과 **백재욱** 교수

제6강 회귀분석과 공분산분석

6.2 공분산분석

6.2 공분산분석

Q. 공분산분석이란?

A. 분산분석 + 회귀분석 -〉 공분산분석

예 X_1: 기계(3대) → Y: 섬유제품의 강도

➡ 기존의 일원배치

예 X_1: 기계(3대), X_2: 원사의 두께 (얇음, 두꺼움)

→ Y: 섬유제품의 강도

➡ 기존의 이원배치

공변수(covariate)

예제 6.2 X_1: 기계(3대), X_2: 원사의 두께 (연속적인 값)

→ Y: 섬유제품의 강도

➡ 공분산분석

<표 6-5> 섬유제품의 강도자료 (y : 강도, x : 두께)

기계 1		기계 2		기계 3	
y	x	y	x	y	x
36	20	40	22	35	21
41	25	48	28	37	23
39	24	39	22	42	26
42	25	45	30	34	21
49	32	44	28	32	15

6.2 공분산분석

◆ 기초적인 공분산분석

- **공분산분석 모형 예(일원배치법 모형에 공변수가 하나 추가된 형태)**

$y_{ij} = \mu' + a_i + \beta x_{ij} + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim N(0, \sigma^2)$ ……………… (6.15)

$(i = 1,2,\cdots,a;\ \ j = 1,2,\cdots,n)$

$<\ =\ >\ y_{ij} = \mu + a_i + \beta(x_{ij} - \bar{\bar{x}}) + \varepsilon_{ij}$ ……………………… (6.16)

앞의 **모형(확대모형)**으로부터 **오차제곱합 SS_E**을 구한다.

모형 (6.15)를 사용하기 전에 다음과 같은 귀무가설이 적절한지 살펴보아야 한다.

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_a$$

$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$

H_1: 적어도 하나의 α_i에 대하여 $\alpha_i \neq 0$

- **이때의 모형(축소모형)은 다음과 같다.**

$y_{ij} = \mu + \beta(x_{ij} - \bar{\bar{x}}) + \epsilon'_{ij}$ ……………………………… (6.18)

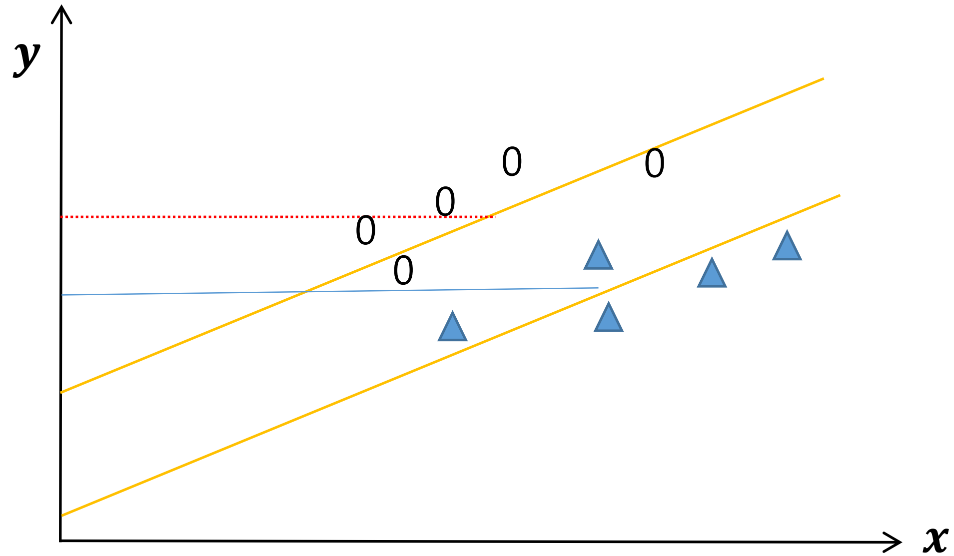
앞의 **모형(축소모형)** 으로부터 **오차제곱합** $SS_E{}'$를 구한다.

검정통계량: $F_0 = \dfrac{(SS'_E - SS_E)/(a-1)}{MS_E}$

$F_0 > F(a-1, a(n-1)-1; \alpha)$이면 **귀무가설 기각**

6.2 공분산분석

예 1) 두 개의 사료(o,▲) =〉 사료 섭취 후 체중 y ➡ **일원배치**

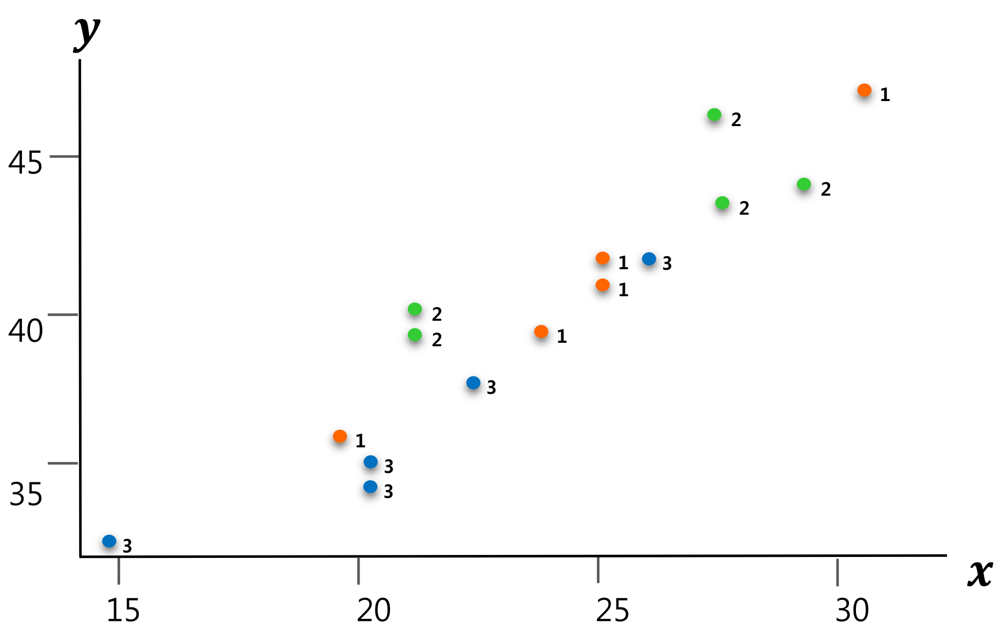


예 2) 두 개의 사료(o, ▲), 초기체중 x(연속적인 값)
=〉 사료 섭취 후 체중 y ➡ **공분산분석**

6.2 공분산분석

〈표 6-5〉 섬유제품의 강도자료(y: 강도, x: 두께)

기계 1		기계 2		기계 3	
y	x	y	x	y	x
36	20	40	22	35	21
41	25	48	28	37	23
39	24	39	22	42	26
42	25	45	30	34	21
49	32	44	28	32	15



$$y_{ij} = \mu + \alpha_i + \beta\left(x_{ij} - \bar{\bar{x}}\right) + \varepsilon_{ij}$$

$$i = 1,2,3; \quad j = 1,2,3,4,5$$

$$F_0 = \frac{(41.27 - 27.99)/2}{27.99 \ / \ 11} = 2.61 < F(2, 11; 0.1)$$

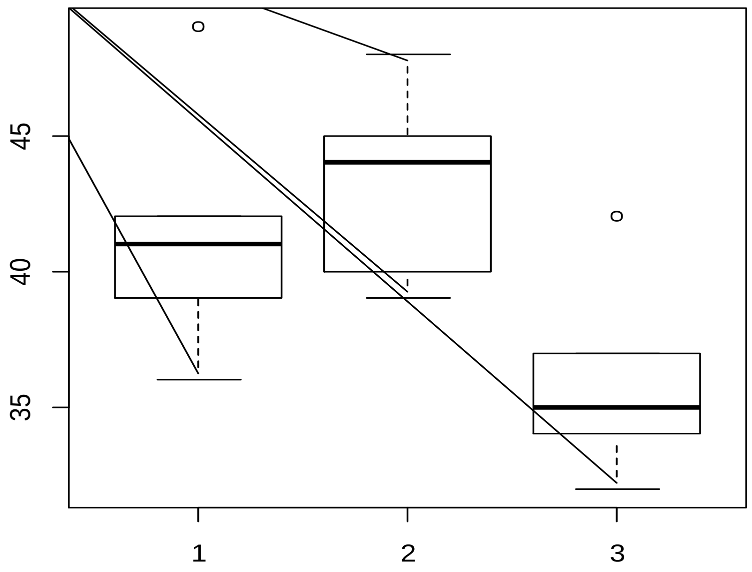
$$= 2.86$$

➡ 10% 유의수준에서 기계 간 섬유제품의 강도에 있어서 차이가 없다.

6.2 공분산분석

R 실습

```
du <- c(20, 25, 24, 25, 32, 22, 28, 22, 30, 28, 21, 23, 26, 21, 15)
gang <- c(36, 41, 39, 42, 49, 40, 48, 39, 45, 44, 35, 37, 42, 34, 32)
machine <- c("1", "2", "3")
machine <- rep(machine, c(5, 5, 5))
textile.data <- data.frame(du, gang, machine)
boxplot(gang ~ machine)
```




```
oneway.out = aov(gang ~ machine) # du(원사의 두께)의 영향력을 배제한 상태임
summary(oneway.out)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
machine	2	140.4	70.20	4.089	**0.0442 ***
Residuals	12	206.0	17.17		

Signif. codes:

*0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1*

6.2 공분산분석

R 실습 (계속)

```
anova <- aov(gang ~ du+machine, data=textile.data) # du(원사의 두께)의 영향력도 감안함
summary(anova)
            Df  Sum Sq  Mean Sq  F value   Pr(>F)
du           1  305.13   305.13  119.933  2.96e-07 ***
machine      2   13.28     6.64    2.611  0.118
Residuals   11   27.99     2.54
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

다음 시간 안내

제7강 (7장)

요인배치법