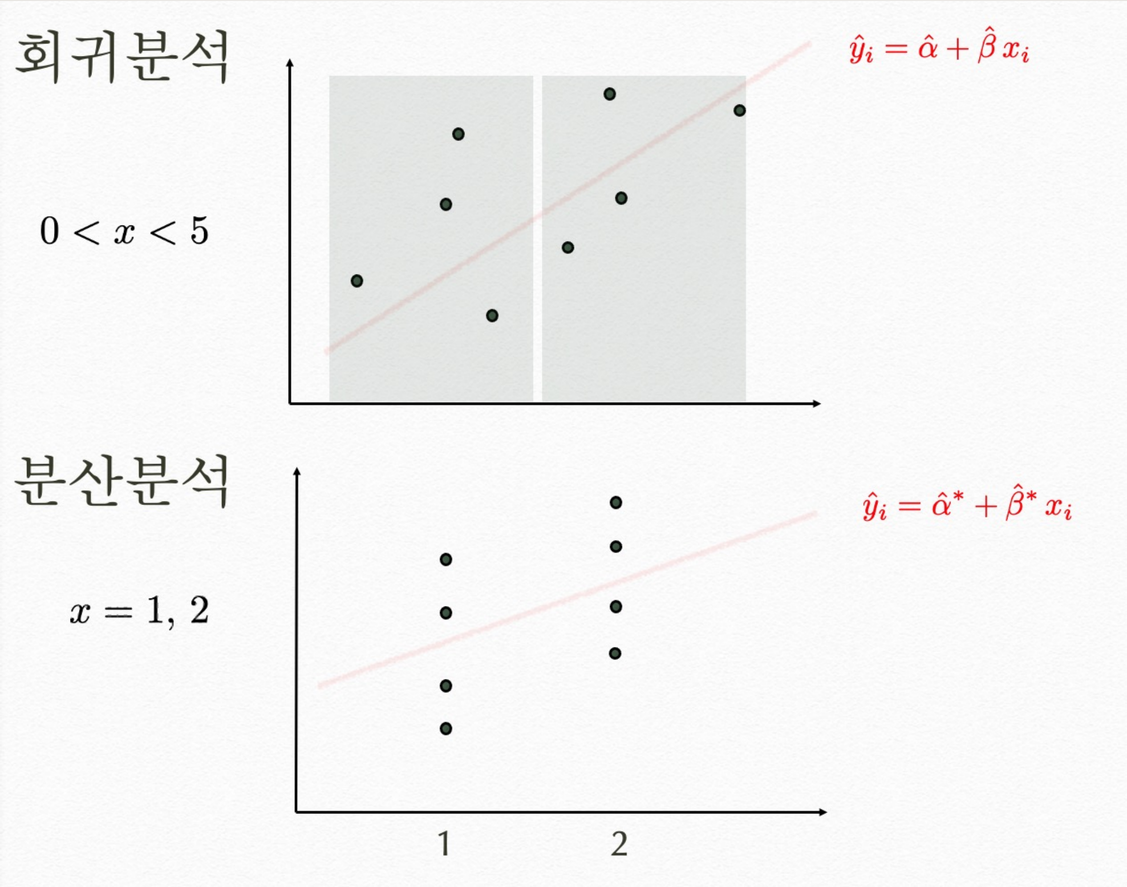
3장 | 회귀 분석과 실험계획의 비교

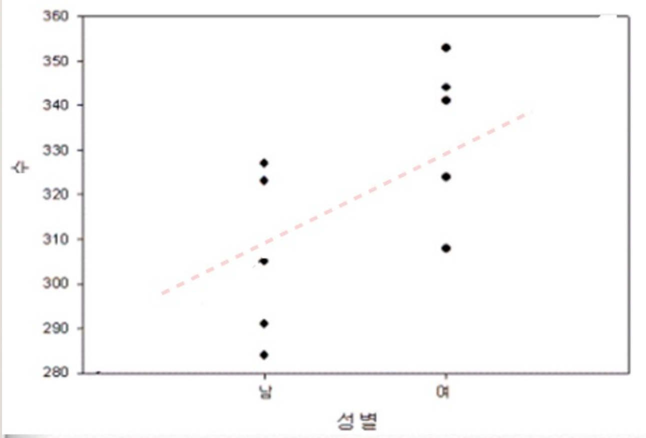
SAS를 이용한 실험 계획과 분산 분석 (자유아카데미)

회귀 분석 VS 분산 분석 모형



예)

성적 vs. 성별



✔ 성별에 따른 평균 점수 차이는 유의한가?

$H_0: \mu_1 = \mu_2$

✔ 모평균의 차이 t-검정

✔ 회귀선의 기울기 t- 검정

회귀분석의 ANOVA F-검정

✔ 회귀분석의 ANOVA table

✔ 실험계획의 ANOVA table

$y_{ij} =$ i-번째 그룹의 j-번째 학생의 성적

$y_{ij} = \alpha + \beta x_{ij} + \epsilon_{ij}$ ⟶ $x_{ij} = ??$

명목변수 (Norminal Variable)

$x_{ij} = 0 \quad \text{or} \quad 1$

$x_{ij} = -1 \quad \text{or} \quad 1$

회귀 분석 VS 분산 분석 모형

$$\boxed{y_{ij} = \alpha + \beta x_{ij} + \epsilon_{ij}, \quad x_{ij} = -1 \ (male) \quad or \quad 1 \ (female)}$$

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{15} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \\ y_{25} \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & -1 \\ 1 & -1 \\ 1 & -1 \\ 1 & -1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{14} \\ \epsilon_{15} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{24} \\ \epsilon_{25} \end{pmatrix} \blacktriangleright \quad \mathbf{y} = X\mathbf{b} + \epsilon$$

$$\hat{\mathbf{b}} = \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = (X^T X)^{-1} X^T \mathbf{y} = \begin{pmatrix} \bar{y}_{..} \\ (\bar{y}_{2.} - \bar{y}_{1.})/2 \end{pmatrix}$$

$$\hat{y}_{ij} = \hat{\alpha} + \hat{\beta} x_{ij} = \bar{y}_{..} + \frac{\bar{y}_{2.} - \bar{y}_{1.}}{2} x_{ij} \quad \blacktriangleright \quad \boxed{\hat{y}_{ij} = \begin{cases} \bar{y}_{1.} & \text{if } x = -1, \\ \bar{y}_{2.} & \text{if } x = 1 \end{cases}}$$

회귀 분석 VS 분산 분석 모형

$$\boxed{\text{t-test for } H_0 : \beta = 0}$$

$$S_{xx} = \sum_i \sum_j (x_{ij} - \bar{x}_{..})^2 = \sum_j (1)^2 + \sum_j (-1)^2 = 10 \quad , \quad MSE = \frac{SSE}{10-2} = \frac{\sum_i \sum_j (y_{ij} - \bar{y}_{i.})}{10-2}$$

$$t_0 = \frac{\hat{\beta}}{\sqrt{MSE/S_{xx}}} = \frac{(\bar{y}_{2.} - \bar{y}_{1.})/2}{\sqrt{MSE/10}} = \frac{\bar{y}_{2\cdot} - \bar{y}_{2\cdot}}{\sqrt{MSE(2/5)}}$$

$$\boxed{\text{t-test for } H_0 : \mu_1 = \mu_2}$$

$$S_p^2 = \frac{4S_1^2 + 4S_2^2}{5+5-2} = \frac{\sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2}{10-2} = MSE$$

$$t_0 = \frac{\bar{y}_{2\cdot} - \bar{y}_{2\cdot}}{\sqrt{S_p^2(2/5)}} = \frac{\bar{y}_{2\cdot} - \bar{y}_{2\cdot}}{\sqrt{MSE(2/5)}}$$

회귀 분석 VS 분산 분석 모형

(Recall : 두 그룹에 대한 ANOVA table)

$$\sum_{i=1}^{2}\sum_{j=1}^{5}(y_{ij}-\bar{y}_{..})^2=\sum_{i=1}^{2}\sum_{j=1}^{5}(\bar{y}_{i.}-\bar{y}_{..})^2+\sum_{i=1}^{2}\sum_{j=1}^{5}(y_{ij}-\bar{y}_{i.})^2$$

SST = SStreat + SSE

(Recall : 회귀분석의 분산분석)

$$\sum_{i=1}^{2}\sum_{j=1}^{5}(y_{ij}-\bar{y}_{i\cdot})^2$$

$$\sum_{i=1}^{2}\sum_{j=1}^{5}(y_{ij}-\bar{y}_{..})^2=\sum_{i=1}^{2}\sum_{j=1}^{5}(\hat{y}_{ij}-\bar{y}_{..})^2+\sum_{i=1}^{2}\sum_{j=1}^{5}(y_{ij}-\hat{y}_{ij})^2$$

SST = SSR + SSE

$$\sum_{i=1}^{2}\sum_{j=1}^{5}(\hat{y}_{ij}-\bar{y}_{..})^2=\sum_{i=1}^{2}\sum_{j=1}^{5}(\hat{\beta}x_{ij})^2=\sum_{i=1}^{2}\sum_{j=1}^{5}\frac{(\bar{y}_{2\cdot}-\bar{y}_{1\cdot})^2}{4}$$

$$=\frac{5}{2}(\bar{y}_{2\cdot}-\bar{y}_{1\cdot})^2=\sum_{i=1}^{2}\sum_{j=1}^{5}(\bar{y}_{i\cdot}-\bar{y}_{..})^2$$

회귀 분석 VS 분산 분석 모형 (분산분석표)

$$y_i = \alpha + \beta x_i + \epsilon_i$$

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

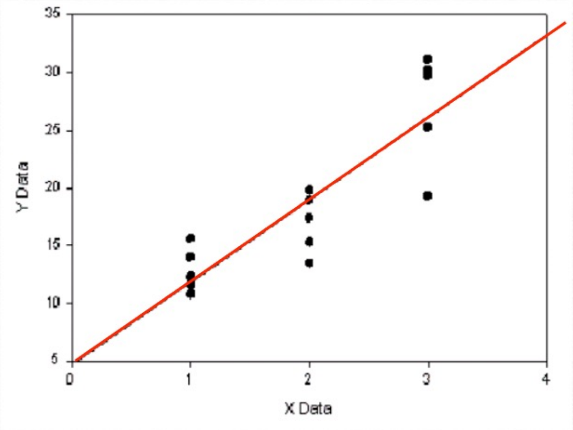
요인	자유도	S.S.	M.S.	Fo
Regression	1	SSR	MSR	MSR/MSE
Error	N-2	SSE	MSE	
Total	N-1	SST	SST = SSR + SSE	

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

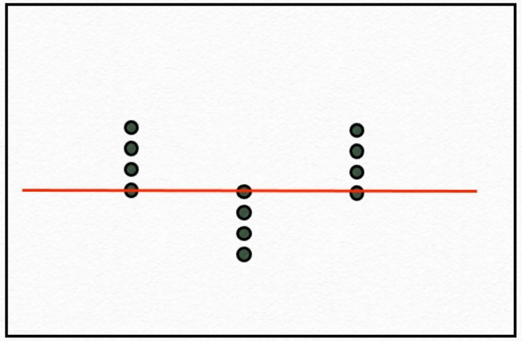
$$\boxed{SST = \sum_{i}\sum_{j}(y_{ij} - \bar{y}_{..})^2}$$

요인	d.f.	S.S.	M.S.	Fo
Treatment	1	SStreat	MStreat	MStreat/MSE
Error	N-2	SSE	MSE	
Total	N-1	SST	SST = SStreat + SSE	

세 그룹 모평균 비교



만약



분산분석 :

$$\boxed{H_0 : \mu_1 = \mu_2 = \mu_3}$$

회귀분석 :

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \epsilon_{ij}$$

$$x_{ij} = 1,\ 2,\ 3$$

$$\boxed{H_0 : \beta_1 = 0 \quad ?}$$

세 그룹 모평균 비교

독립변수의 개수

요인	자유도	SS	MS	F_0
회귀(Reg)	1	511.225	511.225	39.076
잔차(Error)	13	170.075	13.082	
전체(Total)	14	681.300		

그룹개수 - 1

요인	자유도	SS	MS	F_0
처리(Treat)	2	541.828	270.914	23.31
오차(Error)	12	139.472	11.623	
전체(Total)	14	681.300		

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \epsilon_{ij}$$

```
proc reg;
  model y= x;
run;
```

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

```
proc glm;
  class x;
  model y= x;
run;
```

GENERALIZATION

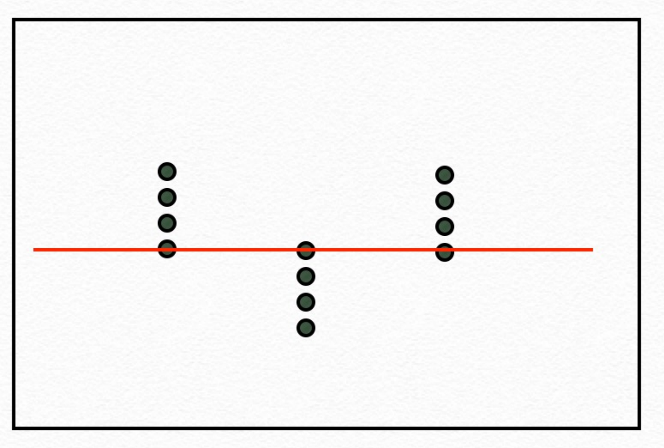
$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 x_{ij}^2 + \epsilon_{ij}$$

$$H_0 : \beta_1 = \beta_2 = 0$$

요인	자유도	S.S.	MS	F
회귀(Reg)	2	541.828	270.914	23.31
잔차(Error)	12	139.472	11.623	
전체(Total)	14	681.300		

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

요인	자유도	SS	MS	F_0
처리(Treat)	2	541.828	270.914	23.31
오차(Error)	12	139.472	11.623	
전체(Total)	14	681.300		



- ❖3개 그룹 비교>> 2차곡선
- ❖4개 그룹비교>> 3차곡선
- ❖a개 그룹비교>> a-1 차 곡선

SSR = SStreat.

회귀분석의 모형식

$$y_{ij} = \alpha + \beta x_{ij} + \epsilon_{ij} \qquad i = 1,\ 2,$$

$$x_{ij} = -1 \text{ or } 1 \qquad j = 1,\ 2,\ 3,\ 4,\ 5$$

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{15} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \\ y_{25} \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & -1 \\ 1 & -1 \\ 1 & -1 \\ 1 & -1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{14} \\ \epsilon_{15} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{24} \\ \epsilon_{25} \end{pmatrix}$$

$$\mathbf{y} = X\mathbf{b} + \epsilon$$

➡️Estimation ?

$$\boxed{\hat{\alpha},\ \hat{\beta}}$$

➡️Test ?

$$\boxed{H_0 : \quad \beta = 0}$$

분산 분석의 모형식 (행렬로 표현)

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad i = 1, 2, \quad j = 1, 2, 3, 4, 5$$

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{15} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \\ y_{25} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{14} \\ \epsilon_{15} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{24} \\ \epsilon_{25} \end{pmatrix}$$

$$\mathbf{y} = X\mathbf{b} + \epsilon$$

➡️Estimation ?

$$\boxed{\hat{\mu}, \ \hat{\tau}_1, \ \hat{\tau}_2}$$

➡️Test ?

$$\boxed{H_0 : \mu_1 = \mu_2 = \mu_3}$$

$$\boxed{H_0 : \tau_1 = \tau_2 = \tau_3}$$

$$\boxed{H_0 : \tau_1 = \tau_2 = \tau_3 = 0}$$

분산 분석의 모형식 (최종)

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad i = 1, 2, \cdots, a, \quad j = 1, 2, \cdots, n_i$$

$$\sum_{i=1}^{a} \tau_i = 0$$

$$\epsilon_{ij} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

i.i.d. =independent and identically distributed