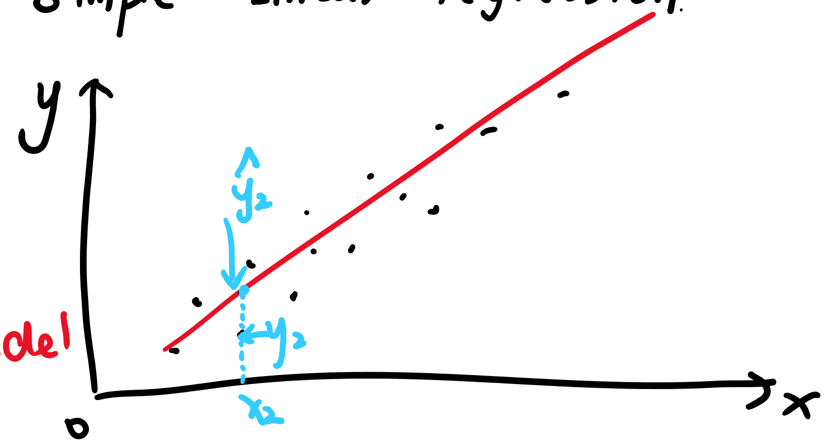
Lecture 11

June 4, 2018 6:55 PM

Chapter 11: Simple Linear Regression

x, y



true model

$$\boxed{y = \beta_0 + \beta_1 x}$$, r correlation coefficient.

Suppose $(x_1, y_1), (x_2, y_2) \ldots\ldots (x_n, y_n)$

$$\boxed{\hat{y} = b_0 + b_1 x}$$

$\hat{y} \Rightarrow y$

$b_0 \Rightarrow \beta_0$

$b_1 \Rightarrow \beta_1$

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

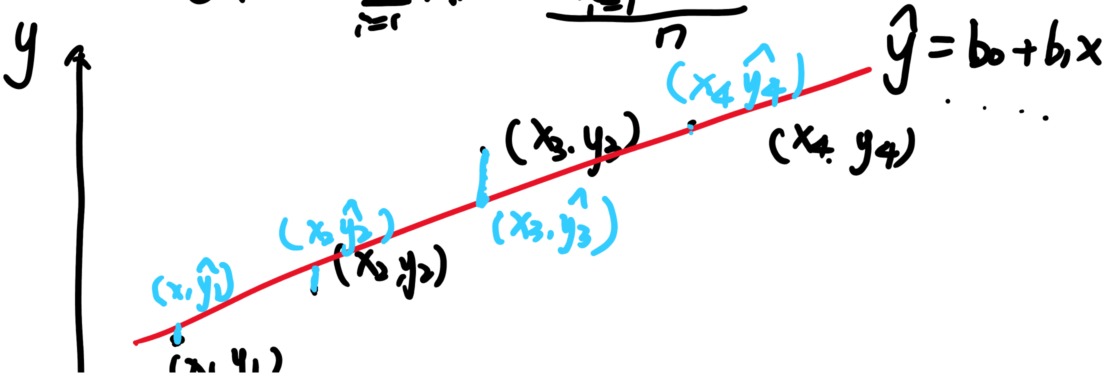
Least $\boxed{\text{square}}$ method. $(\bar{x} = \hat{\mu})$

$$L = \sum_{i=1}^{n} \hat{e}_i^2 = \sum (y_i - \hat{y}_i)^2 \Leftarrow \text{minimize } L.$$

$$\frac{\partial L}{\partial \beta_0} = 0, \quad \frac{\partial L}{\partial \beta_1} = 0 \Longrightarrow \begin{cases} b_1 = \dfrac{S_{xy}}{S_{xx}} = \hat{\beta}_1 \\ b_0 = \bar{y} - b_1 \bar{x} = \hat{\beta}_0 \end{cases}$$

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$$

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



$$\hat{e}_1 = y_1 - \hat{y}_1 , \quad \ldots\ldots$$

$$\left(\sum_{i=1}^{n} \hat{e}_i = 0 \right)$$

Example:

x	1	2	4	7
y	5	3	2	1

Find regression line.

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

$$= 70 - \frac{14^2}{4} = 21$$

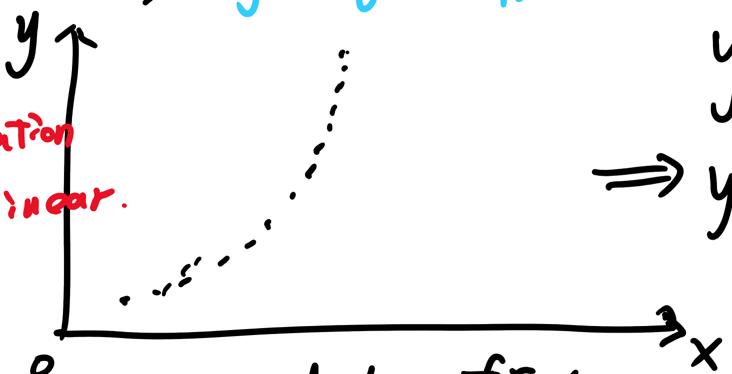
$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n} = 26 - \frac{14 \times 11}{4} = -\frac{25}{2}.$$

$$b_1 = \hat{\beta}_1 = -\frac{25}{42}$$

$$b_0 = \hat{\beta}_0 = \frac{29}{6}$$

$$\Longrightarrow \hat{y} = \frac{29}{6} - \frac{25}{42} x$$

$$Z = X^2$$



transformation for non-linear.

$$y = ax^2 + b$$

$$\Longrightarrow y = aZ + b$$

whats the residual for (x_1, y_1)

$$x_1 = 1 , \quad y_1 = 5$$

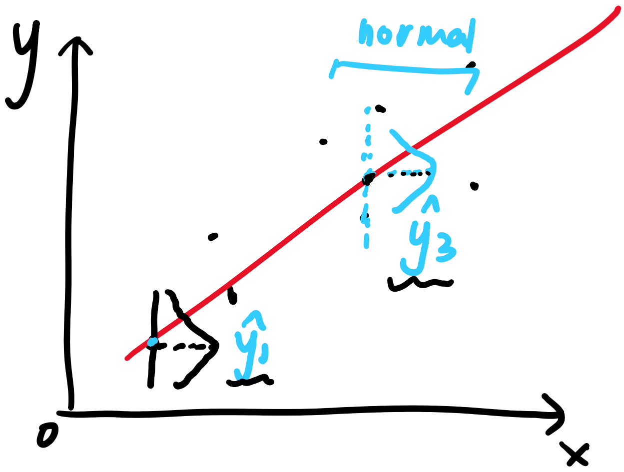
$$\hat{y}_1 = \frac{29}{6} - \frac{25}{42} = \frac{29 \cdot 7 - 25}{42} = 4.238$$

$$y_1 = 5$$

$$\hat{e}_1 = y_1 - \hat{y}_1 = 5 - 4.238 = 0.762$$

$e_1 = y_1 - \hat{y}_1 = 4.350 - 0.762$

$\varepsilon_i \leftarrow$ real error.



Assumption:

1. $E(\varepsilon_i) = 0$
2. $V(\varepsilon_i) = \sigma^2 \leftarrow$ SSE
3. $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0$
4. Normal for ε_i

Estimate σ^2.

The $\boxed{\text{sum of square error}}$

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

$$\hat{\sigma}^2 = \frac{SSE}{n-2} \Rightarrow \left[E(\hat{\sigma}^2) = \sigma^2 \right] \text{ unbiased}$$

The total sum of squares.

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

The regression sum of squares:

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

$$SST = SSE + SSR$$

Properties of Least squares estimator:

(1) $E(\hat{\beta}_1) = \beta_1$ (slope) or $E(b_1) = \beta_1$

(2) $V(\hat{\beta}_1) = V(b_1) = \dfrac{\sigma^2}{S_{xx}} \leftarrow$ variance for residuals or errors

(3) $\hat{\beta}_1$ is normal R.V

as in our assumption, $\varepsilon \sim N(0, \sigma^2)$

as in our assumption, $\varepsilon \sim N(0, \sigma^2)$

$\underbrace{\hat{y} = b_0 + b_1 x}_{\text{Normal}} + \varepsilon \leftarrow$ Normal

Note: If $b_1 = 0$ $\quad \hat{y} = b_0 + \underset{\text{normal}}{\varepsilon}$

then, y has no linear relation with x

Hypothesis test for simple linear regression.

Test for $\beta_1 = 0$.

1. Hypothesis: $H_0: \beta_1 = 0$

$H_A: \beta_1 \neq 0$ (no linear relationship)

2. $T = \dfrac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{S_{xx}}} \sim t_{n-2}$ where $\hat{\sigma} = \sqrt{\dfrac{SSE}{n-2}}$

Degree of freedom $= n-2$.

3. Critical value.

4. Conclusion.

Example:

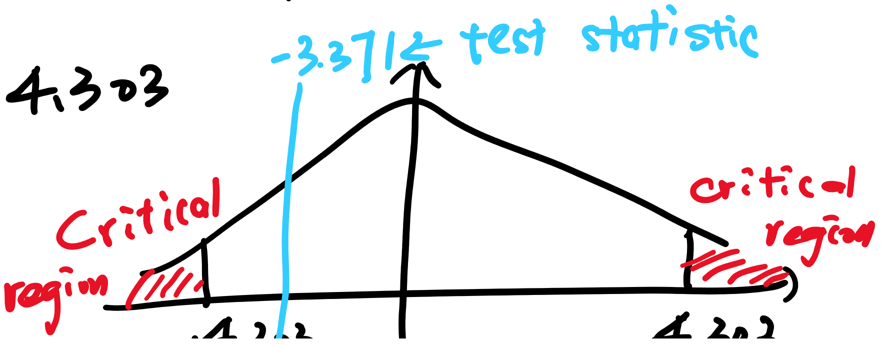
$H_0: \beta_1 = 0$ $\quad (n=4)$

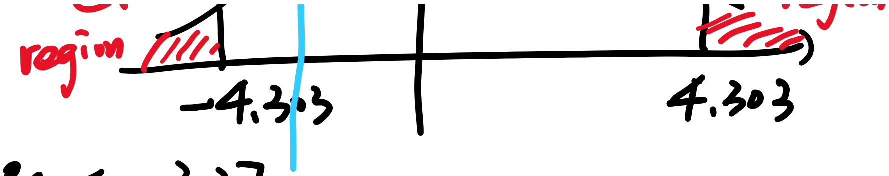
$H_A: \beta_1 \neq 0$

$$T = \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{S_{xx}}} = \frac{-25/42}{\sqrt{\frac{55/42}{4-2}}\Big/\sqrt{21}} = -3.371$$

$\alpha = 5\%$

$t_{2, \frac{5}{2}\%} = 4.303$





$\therefore$ As $t_{2, \frac{5}{2}\%} < -3.371$

do not reject H_0

ANOVA: Analysis of Variance.

Recall:

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

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

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

$SST = SSR + SSE$.

$$F = \frac{SSR}{SSE/(n-2)}$$ $\sim$ F distribution with d.f. $(1, n-2)$

Hypothesis Test (ANOVA)

1. Hypothesis: $H_0: \beta_1 = 0$

   $H_A: \beta_1 \neq 0$

$$\boxed{\begin{gathered} y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \\ H_0: \beta_1 = \beta_2 = 0 \\ H_A: \text{at least one of } \beta_1, \beta_2 \text{ are not zero.} \end{gathered}}$$

2. Test statistic

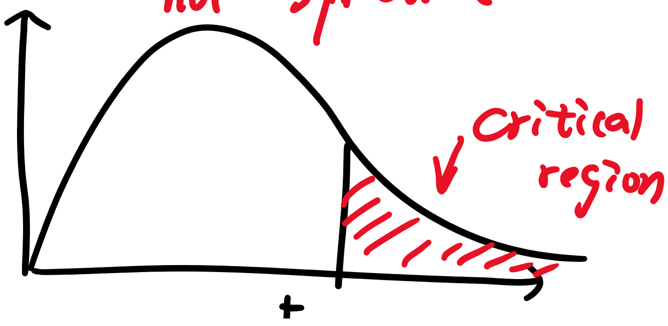
$$F = \frac{SSR}{SSE/(n-2)} \sim F(1, n-2)$$

always one tail as F is a positive distribution.

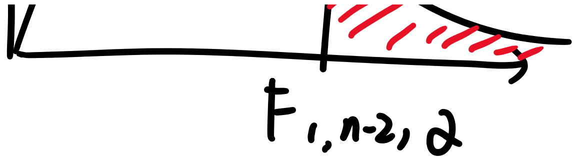
not symmetric

3. Critical value

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



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



4. Conclusion

If $F > F_{1,n-2,\alpha}$, then reject H_0

otherwise, do not reject H_0.

Example: $n=4$, $\alpha=5\%$

critical value. $\to F_{1,2,5\%} = 18.513$

$$\therefore F = \frac{(-25/42)(-25/2)}{55/42/2} = 11.36$$

$\therefore$ do not reject H_0

ANOVA Table:

Source of Variation	Sum of squares	Degree of Freedom	Mean Square	F
Regression	SSR	1	$MSR=\frac{SSR}{1}$	$\frac{MSR}{MSE}$
Error	SSE	$n-2$	$MSE=\frac{SSE}{n-2}$	
Total	SST	$n-1$		

Example

	SS	D.F	MS	F
Regression	625/84	1	625/84	625/55=11.36
Error	55/42	4-2=2	55/84	
Total	25/4			

Error	55/42	[illegible]	[illegible]/84	
Total	25/4			

Confidence Intervals:

1. Slope β_1:

$$\underbrace{\hat{\beta}_1}_{b_1} \pm \underbrace{t_{n-2,\frac{\alpha}{2}}} \frac{\hat{\sigma}}{\sqrt{S_{XX}}}$$

$\leftarrow SD(\hat{\beta}_1)$

2. $\boxed{\text{mean}}$ of y $(\bar{Y})$ at $X = X_0$

$$\hat{y}_0 \pm t_{n-2,\frac{\alpha}{2}} \cdot \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{XX}}}$$

where $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0$

3. Predting new values. (prediction interval) for a $\boxed{\text{single}}$ y_0 at $X = X_0$

$$\hat{y}_0 \pm t_{n-2,\frac{\alpha}{2}} \cdot \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{XX}}}$$

where $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0$

Example:

1. A 95% C.I for β_1:

$$-\frac{25}{42} \pm t_{2,0.025} \cdot \frac{\sqrt{55/84}}{\sqrt{21}}$$

$\leftarrow \hat{\sigma}$ $\leftarrow S_{XX}$

$= (-1.355, \ 0.646)$

2. A 95% CI for $\bar{Y}$ at $X_0 = 5$

X_0 $\bar{X}$

2. A 95% CI for Y at $x_0=5$

$$\underbrace{\left(\frac{29}{6}-\frac{25}{42}\cdot 5\right)}_{\uparrow\ \hat{y}_0} \pm t_{2,0.025}\cdot\underbrace{\sqrt{\frac{55}{84}}}_{\hat{\sigma}}\underbrace{\sqrt{\frac{1}{4}+\frac{(5-\frac{14}{4})^2}{21}}}_{S_{xx}}$$

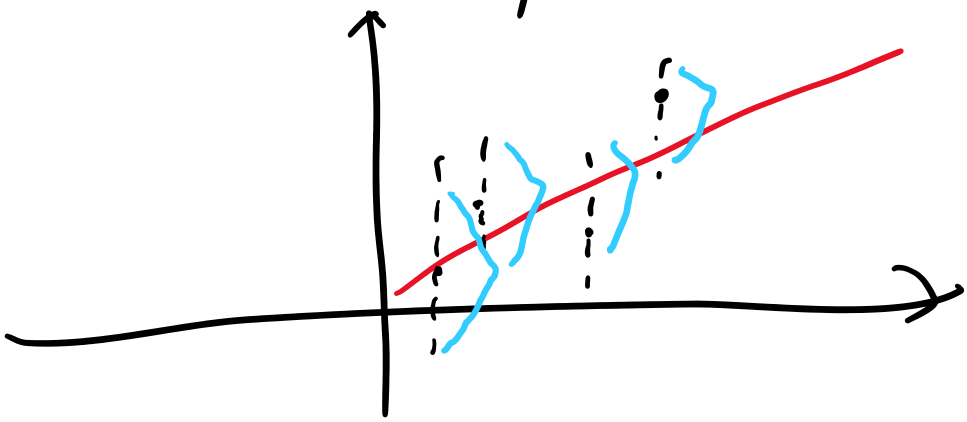
3. A 95% CI for y_0 at $x_0=5$

$(-2.196,\ 6.307)$ $\qquad \sqrt{1+\qquad}$

Assumptions of Regression Model:

Example:

x	y	$\hat{y}_i$	$\hat{e}_i$
1	5	4.238	0.762
2	3	3.6429	-0.6429
4	2	2.4524	-0.4524
7	1	0.6667	0.3333

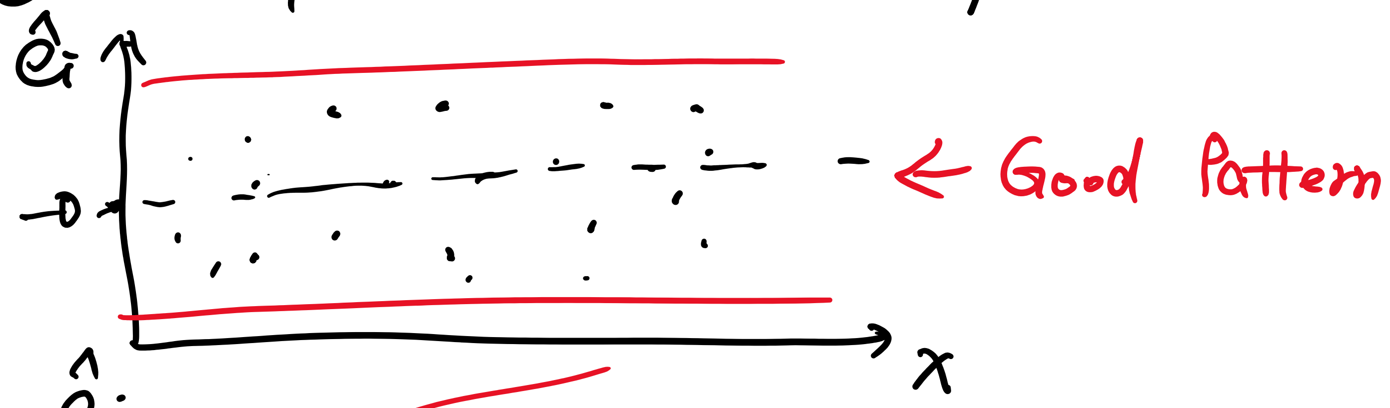


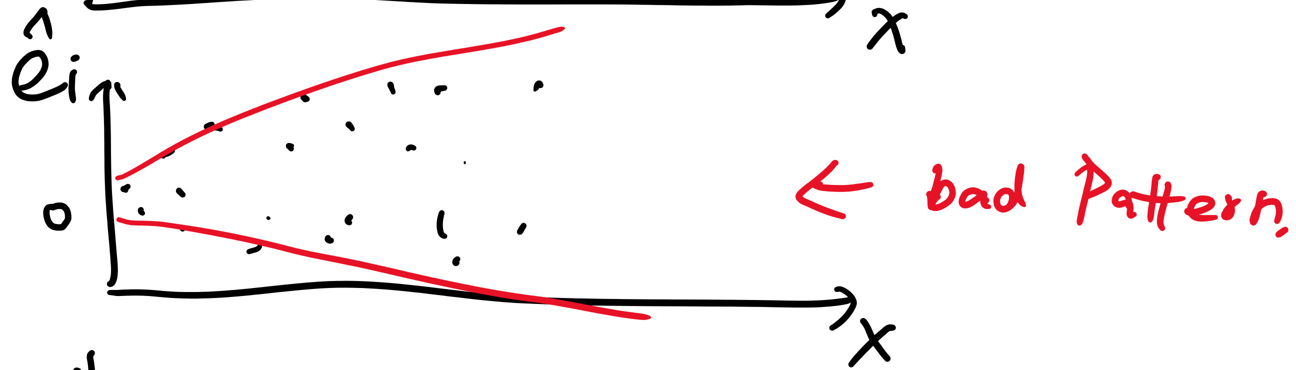
$V(\varepsilon)=\sigma^2 \Leftarrow$ Equal Variances

1. Check Normal Assumptions
   to check if $\varepsilon_i \sim N(0,\sigma^2)$, do a normal probability plot of the observed residuals

2. Check Equal variance Assumptions.

2. Check Equal variance Assumptions.





Recall. R ← correlation coefficient.

R^2 ← coefficient of determination.

$$r = R = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{(n-1)S_x S_y} = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$= \frac{S_{xy}}{\sqrt{S_{xx}\,SST}}$$

$$= \sqrt{\frac{SSR}{SST}}$$

$$\therefore R^2 = \frac{SSR}{SST}$$

R is a estimation of the population coefficent of ρ.

$$R = r = \hat{\rho}$$

$R = r = \hat{\rho}$

Chapter 13: Analysis of Variance.

Overall F-Test

Example:

7.5	5.8	5.9	6.2
6.2	7.3	6.2	6.8
6.9	8.2	5.8	5.7
7.4	7.1	4.7	4.9
9.2	7.8	8.3	6.2
8.5		7.2	7.1
7.6		6.2	5.8
			5.4
$n_1 = 7$	$n_2 = 5$	$n_3 = 7$	$n_4 = 8$

ANOVA to test mean of groups are equal.

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$

H_A: At least one of them not equal.

Notation:

a = # of groups / treatment $\quad a = 4$

$y_{i.}$ = total of the ith group. $\quad y_{1.} = 7.5 + \cdots 7.6$

$\bar{y}_{i.}$ = mean of the ith group. $\quad \bar{y}_{1.} = \frac{y_{1.}}{7} = 7.59$

$y_{..}$ = grand total

$y_{..} =$

$\bar{y}_{..}$ = grand mean

N = total number observations.

n_i = # of observations in the ith group.

y_{ij} = ith group's jth observations.

S_i = sample standard deviation for the ith group.

Example:

$y_{1.} = 53.1$ $\quad$ $y_{2.} = 36.2$ $\quad$ $y_{3.} = 44.3$ $\quad$ $y_{4.} = 48.1$

$\bar{y}_{1.} = 7.59$ $\quad$ $\bar{y}_{2.} = 7.24$ $\quad$ $\bar{y}_{3.} = 6.33$ $\quad$ $\bar{y}_{4.} = 6.01$

$n_1 = 7$ $\quad$ $n_2 = 5$ $\quad$ $n_3 = 7$ $\quad$ $n_4 = 8$

$y_{..} = 181.7$

$\bar{y}_{..} = 6.73$

$N = 27$

Hypothesis: $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$

H_A: At least one of them not equal

Test statistic:

SSR

$$F_{a-1, N-a} = \frac{SST_r/(a-1)}{SSE/(N-a)}$$

where: $SST = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{a} \sum_{j=1}^{n_i} y_{ij}^2 - \frac{y_{..}^2}{N}$

variance of

d.f. = N-1

$$d.f = N-1$$

variance of total

$$SST_r = \sum_{i=1}^{a} n_i(\bar{y}_{i.} - \bar{y}_{..})^2 = \sum_{i=1}^{a} \frac{y_{i.}^2}{n_i} - \frac{y_{..}^2}{N}$$

similar to SSR

variance among groups

$$d.f = a-1$$

$$SSE = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 = \sum_{i=1}^{a} (n_i - 1) S_i^2$$

variance within groups

$$d.f = N-a$$

ANOVA TABLE:

Source of variance	SS	D.F	M.S.	F
Treatment	SST_r	$a-1$	$\frac{SST_r}{a-1} = MST_r$	$\frac{MST_r}{MSE}$
Error	SSE	$N-a$	$\frac{SSE}{N-a} = MSE$	
Total	SST	$N-1$		

Example:

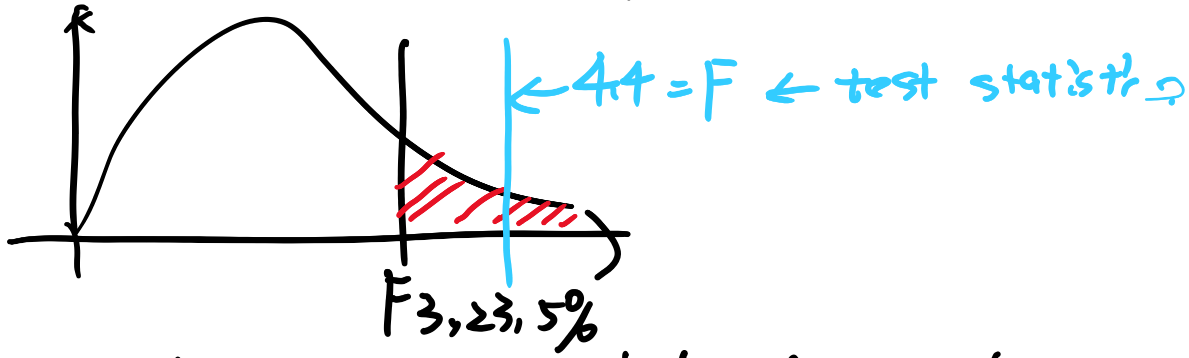
	SS	D.F	MS	F
Treatments	11.673 SST_r	3 $a-1$	3.891	$\frac{3.891}{0.8828} = 4.4$
Error	20.3043 SSE	23 $N-a$	0.8828	
Total	31.9773 SST	26 $N-1$		

Critical value

$$F_{a-1, N-a, 5\%} = F_{3,23,5\%}$$

$$F_{a-1,\,N-a,\,5\%} = F_{3,23,5\%}$$
$$= 3.03$$

$$F = 4.4 > F_{3,23,5\%} = 3.03$$



∴ we reject H_0. conclude that at least one of the mean is not equal with others.

Test for individual pairs of means.

Fisher's LSD (Least Significant Difference)

1. Hypothesis: $H_0: \mu_i = \mu_j$

$H_A: \mu_i \neq \mu_j$, $i \neq j$ any pair.

$$\binom{4}{2} = 6$$

2. Test statistic.

① $\bar{y}_{i.} - \bar{y}_{j.}$

② $LSD = t_{n-a,\frac{\alpha}{2}} \cdot \sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$

↑ ANOVA Table.

3. Critical Region:

Reject H_0 if

$$|\bar{y}_{i.} - \bar{y}_{j.}| > LSD$$

otherwise do not reject H_0.

4. Conclusion.

Example: Test μ_1 and μ_2 whether have a difference at $\alpha = 5\%$.

Hypothesis: $H_0: \mu_1 = \mu_2$

$H_A: \mu_1 \neq \mu_2$.

Test statistic:

$$|\bar{y}_{1.} - \bar{y}_{2.}| = |7.59 - 7.24| = 0.35$$

$$LSD = t_{23,\ 0.025}\sqrt{MSE\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$= 2.069 \cdot \sqrt{0.8828\left(\frac{1}{7} + \frac{1}{5}\right)}$$

$$= 1.138$$

$\because$ $0.35 < LSD$

$\therefore$ do not reject H_0.

Hypothesis	LSD	$\lvert\bar{y}_{i.} - \bar{y}_{j.}\rvert$	Conclusion
$H_0: \mu_1 = \mu_2$	1.138	0.35	
$\mu_1 = \mu_3$	1.039	1.26	$\mu_1 \neq \mu_3$
$\mu_1 = \mu_4$	1.006	1.58	$\mu_1 \neq \mu_4$

$\mu_1=\mu_4$	1.006	1.38	$\mu_1 \neq \mu_4$
$\mu_2=\mu_3$	1.138	0.91	
$\mu_2=\mu_4$	1.108	1.23	$\mu_2 \neq \mu_4$
$\mu_3=\mu_4$	1.006	0.32	

Assumptions for ANOVA:

1. Each population must be normal.
2. The population have equal variance.
3. The samples must be independent.
4. All other factors must be equal other than groups.

Residuals for ANOVA:

$\hat{e}_{ij} = y_{ij} - \overline{y_{i.}}$ (observation minus sample average)

Assumption: 1. residuals follow a normal distribution.

2. For equal group, the variance of residuals should be similar.