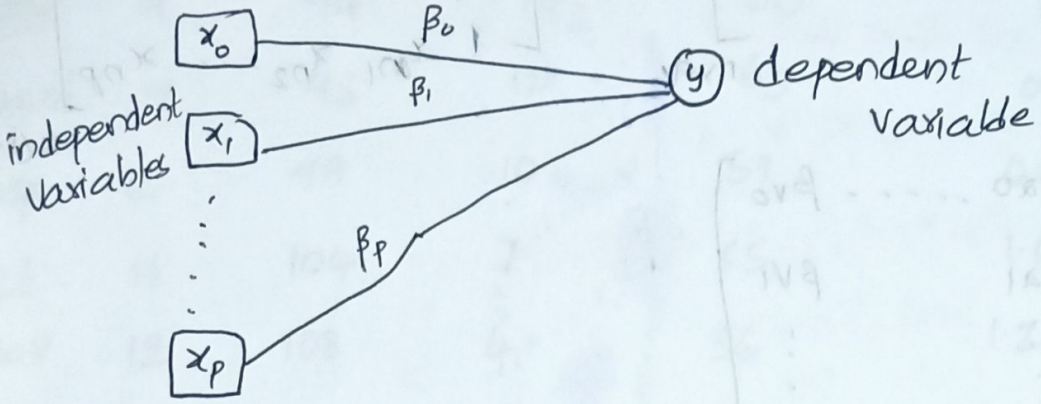
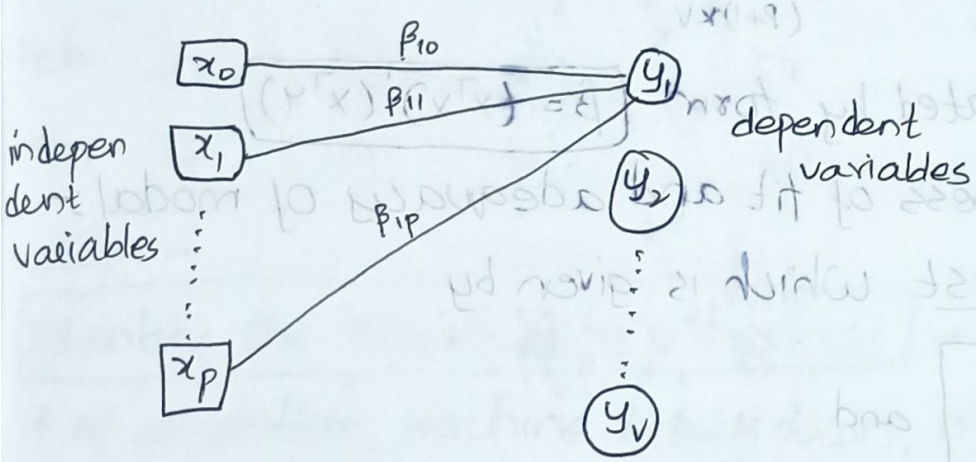
UNIT-3

Multi-variate Regression:—

Multiple linear regression:- (MLR)



Multivariate linear regression:- (MvLR)



- Multivariate linear regression model is given by

$$y_1 = \beta_{10} + \beta_{11}x_1 + \beta_{12}x_2 + \cdots\cdots + \beta_{1p}x_p + \epsilon_1$$

$$y_2 = \beta_{20} + \beta_{21}x_1 + \beta_{22}x_2 + \cdots\cdots + \beta_{2p}x_p + \epsilon_2$$

$$\vdots$$

$$y_v = \beta_{v0} + \beta_{v1}x_1 + \beta_{v2}x_2 + \cdots + \beta_{vp}x_p + \epsilon_v$$

Let the multivariate linear regression is of the form

$$\boxed{Y = X\beta + \epsilon}$$

$$\Rightarrow Y_{n\times q} = X_{n\times(p+1)} \cdot \beta_{(p+1)\times q} + \epsilon_{n\times q}$$

where

$$Y = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1q} \\ y_{21} & y_{22} & \cdots & y_{2q} \\ \vdots & & & \\ y_{n1} & y_{n2} & \cdots & y_{nq} \end{bmatrix}_{n \times q}$$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}_{n \times (p+1)}$$

$$\beta = \begin{bmatrix} \beta_{10} & \beta_{20} & \cdots & \beta_{v0} \\ \beta_{11} & \beta_{21} & & \beta_{v1} \\ \vdots & \vdots & & \vdots \\ \beta_{1p} & \beta_{2p} & & \beta_{vp} \end{bmatrix}_{(p+1) \times v}$$

where β is calculated by formula $\hat{\beta} = (X^T X)^{-1} (X^T Y)$

To test the Goodness of fit and adequacy of modal :- we apply Wilk's-test which is given by

$$\Lambda^* = \frac{|W|}{|T|} \quad \text{and}$$

$$F = \frac{n-K-1}{K-1} * \frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}} \sim F_{P(K-1),\, P(n-K-1)}$$

which is obtained by the multivariate analysis of varience (MANOVA)

Eg:- Fit a multivariate linear regression model with profit & sales depending on percentage of absentism, machine breakdown and M ratio.

Month	Profit	sales	% of absentis	Machine Breaks	M-Ratio
April	10	100	9	62	1.0
May	12	110	8	58	1.3
June	11	105	7	64	1.2
July	9	94	14	60	0.8
August	9	95	12	63	0.8
Sep	10	99	10	57	0.9
oct	11	104	7	55	1.0
Nov	12	108	4	56	1.2
Dec	11	105	6	59	1.1
Jan	10	98	5	61	1.0
Feb	11	103	7	57	1.2
Mar	12	110	6	60	1.2

step 1:-

Identify the variables of Interest → profit, sales, % of absentism, machine breakdown, M-ratio

step 2:- Identify the response variables →

y_1 → profit
y_2 → sales
} dependent variables

step 3:- Identify the exclamatory variables →

x_1 → % of absentism
x_2 → machine breakdowns
x_3 → M-ratio
} independent variables

step 4:- Find the dependent relationships →

$$y = f(x)$$

Step 5:- Solve the above system of relations by applying the formula.

$$\hat{\beta} = (X^T X)^{-1} (X^T Y)$$

- Let the multivariate linear regression modal in this problem the of the form

$$Y_1 = \beta_{10} + \beta_{11} X_1 + \beta_{12} X_2 + \beta_{13} X_3$$

$$Y_2 = \beta_{20} + \beta_{21} X_1 + \beta_{22} X_2 + \beta_{23} X_3$$

where

$$Y = \begin{matrix} \text{Profit} & \text{Sales} \end{matrix} \begin{bmatrix} 10 & 100 \\ 12 & 110 \\ 11 & 105 \\ 9 & 94 \\ 9 & 95 \\ 10 & 99 \\ 11 & 104 \\ 12 & 108 \\ 11 & 105 \\ 10 & 98 \\ 11 & 103 \\ 12 & 110 \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & 9 & 62 & 1.0 \\ 1 & 8 & 58 & 1.3 \\ 1 & 7 & 64 & 1.2 \\ 1 & 14 & 60 & 0.8 \\ 1 & 12 & 63 & 0.8 \\ 1 & 10 & 57 & 0.9 \\ 1 & 7 & 55 & 1.0 \\ 1 & 4 & 56 & 1.2 \\ 1 & 6 & 59 & 1.1 \\ 1 & 5 & 61 & 1.0 \\ 1 & 7 & 57 & 1.2 \\ 1 & 6 & 60 & 1.2 \end{bmatrix}$$

where $\hat{\beta} = (X^T X)^{-1} (X^T Y)$

$$\Rightarrow \hat{\beta} = \begin{bmatrix} 10.897 & 91.097 \\ -0.045 & -0.064 \\ -0.087 & -0.294 \\ 5.035 & 27.835 \end{bmatrix}$$

Y_1 (Profit) $= 10.2917 - 0.045X_1 - 0.087X_2 + 5.035X_3$

Y_2 (sales) $= 91.097 - 0.064X_1 - 0.294X_2 + 27.835X_3$

Eg: Construct one way MANOVA table to the following data.

			$\bar{y}_i$	$\bar{\bar{y}}$
Treatment 1:	$\begin{bmatrix}9\\3\end{bmatrix}$	$\begin{bmatrix}6\\2\end{bmatrix}$ $\begin{bmatrix}9\\7\end{bmatrix}$	$\begin{bmatrix}8\\4\end{bmatrix}$	
Treatment 2:	$\begin{bmatrix}0\\4\end{bmatrix}$	$\begin{bmatrix}2\\0\end{bmatrix}$	$\begin{bmatrix}1\\2\end{bmatrix}$	$\begin{bmatrix}4\\5\end{bmatrix}$
Treatment 3:	$\begin{bmatrix}3\\8\end{bmatrix}$	$\begin{bmatrix}1\\9\end{bmatrix}$ $\begin{bmatrix}2\\7\end{bmatrix}$	$\begin{bmatrix}2\\8\end{bmatrix}$	

Sol:

$H_0: (\mu_{11} = \mu_{12} = \mu_{13}) = (\mu_{21} = \mu_{22}) = (\mu_{31} = \mu_{32} = \mu_{33})$

$H_1: (\mu_{11} \neq \mu_{12} \neq \mu_{13}) \neq (\mu_{21} \neq \mu_{22}) \neq (\mu_{31} = \mu_{32} = \mu_{33})$

Level of significance = 5% (choosen)

To Test the above hypothesis, the procedure is as follows:-

For y_1:

① Sum of squares due to error

$$SSe = \sum (y_{ij} - \bar{y}_{i\cdot})^T (y_{ij} - \bar{y}_{i\cdot}) = \sum (y_{ij} - \bar{y}_{i\cdot})^2$$

$$= (9-8)^2 + (6-8)^2 + (9-8)^2 + (0-1)^2 + (2-1)^2 + (3-2)^2 + (1-2)^2 + (2-2)^2 = 10$$

② Sum of squares due to total

$$SST = \sum (y_{ij} - \bar{\bar{y}})^T (y_{ij} - \bar{\bar{y}}) = \sum (y_{ij} - \bar{\bar{y}})^2$$

$$= (9-4)^2 + (6-4)^2 + (9-4)^2 + (0-4)^2 + (2-4)^2 + (3-4)^2 + (1-4)^2 + (2-4)^2 = 88$$

③ Sum of squares due to Regression

$$SSR = \sum n_i (\bar{y}_i - \bar{\bar{y}})^T (\bar{y}_i - \bar{\bar{y}}) = SST - SSe = 88 - 10 = 78$$

For y_2:

① $SSe = \sum (y_{ij} - \bar{y}_i)^T (y_{ij} - \bar{y}_i) = \sum (y_{ij} - \bar{y}_i)^2$

$= (3-4)^2 + (2-4)^2 + (7-4)^2 + (4-2)^2 + (0-2)^2 + (8-8)^2 +$

$(9-8)^2 + (7-8)^2 = 24$

② $SST = \sum (y_{ij} - \bar{\bar{y}})^T (y_{ij} - \bar{\bar{y}}) = \sum (y_{ij} - \bar{\bar{y}})^2$

$= (3-5)^2 + (2-5)^2 + (7-5)^2 + (4-5)^2 + (0-5)^2 + (8-5)^2 + (9-5)^2 +$

$(7-5)^2 = 72$

③ $SSy = \sum n_i (\bar{y}_i - \bar{\bar{y}})(\bar{y}_i - \bar{\bar{y}}) = SST - SSe = 72 - 24 = 48$

· Cross product values of y_1 & y_2:-

① $SSe = (9\times3 - 8\times4) + (6\times2 - 8\times4) + (9\times7 - 8\times4)$

$+ (0\times4 - 1\times2) + (2\times0 - 1\times2) + (3\times8 - 2\times8) + (1\times9 - 2\times8)$

$+ (2\times7 - 2\times8) = 1$

② $SST = (9\times3 - 4\times5) + (6\times2 - 4\times5) + (9\times7 - 4\times5) + (0\times4 - 4\times5)$

$+ (2\times0 - 4\times5) + (3\times8 - 4\times5) + (1\times9 - 4\times5) + (2\times7 -$

$4\times5) = 7 + (-8) + 43 - 20 - 20 + 4 - 11 - 6 = -11$

③ $SSy = SST - SSe = -11 - 1 = -12$

MANOVA One Way Classification Table:-

Source of variation	sum of squares	degrees of freedom	Wilks value	F-Test
Regression	$\sum n_i (\bar{y}_i - \bar{\bar{y}})^T (\bar{y}_i - \bar{\bar{y}}) = B$	$K-1$	$\Lambda^* = \frac{\lvert W \rvert}{\lvert T \rvert}$	$F = \frac{n-K-1}{K-1} *$
error	$\sum (y_{ij} - \bar{y}_i)^T (y_{ij} - \bar{y}_i) = W$	$n-k$		$\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}} \sim F_{P(K-1),\ P(n-K-1)}$
Total	$\sum (y_{ij} - \bar{\bar{y}})^T (y_{ij} - \bar{\bar{y}}) = T$	$n-1$		

Inference:-

- If $F_{cal} > F_{table}$ value, we reject H_0

Hence we conclude that otherwise we accept H_0

MANOVA One way classification Table:-

Source of variation	Sum of squares	degrees of freedom	Will's value	F-Test
Regression	$\begin{pmatrix} 78 & -12 \\ -12 & 48 \end{pmatrix} = B$	$3-1=2$	$\Lambda^* = \frac{\|W\|}{\|T\|} = \frac{939}{6215} = 0.0384$	$F = \frac{n-k-1}{k-1} * \frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}$
Error	$\begin{pmatrix} 10 & 1 \\ 1 & 24 \end{pmatrix} = W$	$8-3=5$		$= \frac{8-3-1}{3-1} * \frac{1-\sqrt{0.038}}{\sqrt{0.038}}$
Total	$\begin{pmatrix} 88 & -11 \\ -11 & 72 \end{pmatrix} = T$	$8-1=7$		$= 8.2$

$\sim F_{2(3-1), 2(8-3-1)}$

$\sim F_{4,8}$

Inference:-

The table value at 5% L.O.S at 4,8 degrees of freedom = 3.84

$\therefore F_{cal} > F_{table}$, we reject H_0

Hence we conclude that there is no homogenity among the treatments.

② Construct one-way MANOVA table to the following data.

Treatment A: $\begin{bmatrix} 2 \\ 3 \end{bmatrix} \begin{bmatrix} 3 \\ 4 \end{bmatrix} \begin{bmatrix} 5 \\ 4 \end{bmatrix} \begin{bmatrix} 2 \\ 5 \end{bmatrix}$

Treatment B: $\begin{bmatrix} 4 \\ 8 \end{bmatrix} \begin{bmatrix} 5 \\ 6 \end{bmatrix} \begin{bmatrix} 6 \\ 7 \end{bmatrix}$

Treatments :- $\begin{bmatrix}7\\6\end{bmatrix}\begin{bmatrix}8\\7\end{bmatrix}\begin{bmatrix}10\\8\end{bmatrix}\begin{bmatrix}9\\5\end{bmatrix}\begin{bmatrix}7\\6\end{bmatrix}$

Sol:- $H_0: (\mu_{11}=\mu_{12}=\mu_{13}=\mu_{14}) = (\mu_{21}=\mu_{22}=\mu_{23}) = (\mu_{31}=\mu_{32}=\mu_{33}=\mu_{34}=\mu_{35})$

$H_1: (\mu_{11}=\mu_{12}=\mu_{13}=\mu_{14}) \neq (\mu_{21}=\mu_{22}=\mu_{23}) \neq (\mu_{31}=\mu_{32}=\mu_{33}=\mu_{34}=\mu_{35})$

Level of significance = 5% (choosen)

To test the above hypothesis, the procedure is as follows:-

		$\bar{y}_i$	$\bar{\bar{y}}$
A	$\begin{bmatrix}2\\3\end{bmatrix}\begin{bmatrix}3\\4\end{bmatrix}\begin{bmatrix}5\\4\end{bmatrix}\begin{bmatrix}2\\5\end{bmatrix}$	$\begin{bmatrix}3\\4\end{bmatrix}$	$\begin{bmatrix}5.67\\5.75\end{bmatrix}$
B	$\begin{bmatrix}4\\8\end{bmatrix}\begin{bmatrix}5\\6\end{bmatrix}\begin{bmatrix}6\\7\end{bmatrix}$	$\begin{bmatrix}5\\7\end{bmatrix}$	
C	$\begin{bmatrix}7\\6\end{bmatrix}\begin{bmatrix}8\\7\end{bmatrix}\begin{bmatrix}10\\8\end{bmatrix}\begin{bmatrix}9\\5\end{bmatrix}\begin{bmatrix}7\\6\end{bmatrix}$	$\begin{bmatrix}8\\6.4\end{bmatrix}$	

For y1:-

① $SSE = \sum (y_{ij}-\bar{y}_i)^T (y_{ij}-\bar{y}_i) = \sum (y_{ij}-\bar{y}_i)^2$

$= (2-3)^2+(3-3)^2+(5-3)^2+(2-3)^2+(4-5)^2+(5-5)^2+(6-5)^2+(7-8)^2+(8-8.2)^2+(10-8.2)^2+(9-8.2)^2+(7-8.2)^2 = 14.8$

② $SST = \sum (y_{ij}-\bar{\bar{y}})^T (y_{ij}-\bar{\bar{y}}) = \sum (y_{ij}-\bar{\bar{y}})^2$

$= (2-5.67)^2+(3-5.67)^2+(5-5.67)^2+(2-5.67)^2+(4-5.67)^2+(5-5.67)^2+(6-5.67)^2+\dots+(7-5.67)^2$

$= 76.67$

③ $SSR = \sum n_i(\bar{y}_i - \bar{\bar{y}})^T(\bar{y}_i - \bar{\bar{y}}) = SST - SSE = 61.87$

For y_2:

① $SSE = \sum(y_{ij} - \bar{y}_i)^T(y_{ij} - \bar{y}_i) = \sum(y_{ij} - \bar{y}_i)^2$

$= (3-4)^2 + (4-4)^2 + (4-4)^2 + (5-4)^2 + (8-7)^2 + (6-7)^2 +$
$(7-7)^2 + (6-6.4)^2 + (7-6.4)^2 + (8-6.4)^2 + (5-6.4)^2 + (6-6.4)^2$

$= 9.2$

② $SST = \sum(y_{ij} - \bar{\bar{y}})^T(y_{ij} - \bar{\bar{y}}) = \sum(y_{ij} - \bar{\bar{y}})^2$

$= (3-5.75)^2 + (4-5.75)^2 + (5-5.75)^2 + \dots + (6-5.75)^2$

$= 28.25$

③ $SSR = \sum n_i(\bar{y}_i - \bar{\bar{y}})(\bar{y}_i - \bar{\bar{y}}) = SST - SSE = 19.05$

For both y_1 & y_2:-

① $SSE = (2\times3 - 3\times4) + (3\times4 - 3\times4) + (5\times4 - 3\times4) + (2\times5 - 3\times4)$
$+ (4\times8 - 5\times7) + (5\times6 - 5\times7) + (6\times7 - 5\times7) +$
$(7\times6 - 8.2\times6.4) + (8\times7 - 8.2\times6.4) \dots = 1.6$

② $SST = (2\times3 - 5.67\times5.75) + (3\times4 - 5.67\times5.75) + \dots$
$\dots + (7\times6 - 5.67\times5.75) = 26$

③ $SSR = SST - SSE = 26 - 1.6 = 24.4$

MANOVA ONE WAY CLASSIFICATION TABLE

Source of variation	Sum of squares	degrees of freedom	Will's value	F-Test
Regression	$\begin{pmatrix} 61.87 & 24.4 \\ 24.4 & 19.05 \end{pmatrix} = B$	$3-1=2$	$\Lambda^* = \frac{\lvert W\rvert}{\lvert T\rvert}$	$F = \frac{12-3-1}{3-1} \times \frac{1-\sqrt{0.0896}}{\sqrt{0.0896}}$
error	$\begin{pmatrix} 14.8 & 1.6 \\ 1.6 & 9.2 \end{pmatrix} = W$	$12-3=9$	$= \frac{133.6}{1489}$	
Total	$\begin{pmatrix} 76.67 & 26 \\ 26 & 28.25 \end{pmatrix} = T$	$12-1=11$	$= 0.0896$	$F = 9.36$
				$\sim F_{4,16}$

Inference:-

The table value of F at 5% L.O.S for 4,16 degrees of freedom is $F_{tab} = 3.01$

$F_{cal} > F_{tab}$, we reject H_0

hence we conclude that there is no homogenity between the constraints.

③ Carry out MANOVA one way classification at 1% level of significance and comment:

		$\bar{y}_i$	$\bar{\bar{y}}$
Treatment 1:	$\begin{bmatrix}4\\6\end{bmatrix}\begin{bmatrix}3\\7\end{bmatrix}\begin{bmatrix}7\\5\end{bmatrix}\begin{bmatrix}4\\8\end{bmatrix}\begin{bmatrix}5\\8\end{bmatrix}$	$\begin{bmatrix}4.6\\6.8\end{bmatrix}$	
Treatment 2:	$\begin{bmatrix}4\\5\end{bmatrix}\begin{bmatrix}2\\7\end{bmatrix}\begin{bmatrix}2\\5\end{bmatrix}$	$\begin{bmatrix}2.667\\5.667\end{bmatrix}$ $\begin{bmatrix}3.75\\5.33\end{bmatrix}$	
Treatment 3:	$\begin{bmatrix}3\\6\end{bmatrix}\begin{bmatrix}6\\2\end{bmatrix}\begin{bmatrix}2\\1\end{bmatrix}\begin{bmatrix}3\\4\end{bmatrix}$	$\begin{bmatrix}3.5\\3.25\end{bmatrix}$	

Sol:- $H_0 : (\mu_{11} = \mu_{12} = \mu_{13} = \mu_{14} = \mu_{15}) = (\mu_{21} = \mu_{22} = \mu_{23}) = (\mu_{31} = \mu_{32} = \mu_{23} = \mu_{34})$

$H_1 : (\mu_{11} = \mu_{12} = \mu_{13} = \mu_{14} = \mu_{15}) \neq (\mu_{21} = \mu_{22} = \mu_{23}) \neq (\mu_{31} = \ldots = \mu_{34})$

Level of significance = 1% (Given)

To test the above hypothesis, the procedure is as follows.

table

For y_1:-

① $SSE = \sum (y_{ij} - \bar{y}_i)^T (y_{ij} - \bar{y}_i) = \sum (y_{ij} - \bar{y}_i)^2$

$= (4-4.6)^2 + (3-4.6)^2 + (7-4.6)^2 + (4-4.6)^2 + (5-4.6)^2 +$
$(4-2.667)^2 + (2-2.667)^2 + (2-2.667)^2 + (3-3.5)^2 +$
$(6-3.5)^2 + (2-3.5)^2 + (3-3.5)^2 = 0.36 + 2.56 + 5.76 + 0.36 +$

$0.16 + 1.776 + 0.4356 + 0.4356 + 0.25 + 6.25 + 2.25 + 0.25 =$

$= 20.8472$ (20.87)

② $SST = \sum (y_{ij} - \bar{y})^T (y_{ij} - \bar{y}) = \sum (y_{ij} - \bar{y})^2$

$= (4-3.75)^2 + (3-3.75)^2 + (7-3.75)^2 + (4-3.75)^2 + (5-3.75)^2 +$

$(4-3.75)^2 + (2-3.75)^2 + (2-3.75)^2 + (3-3.5)^2 + (6-3.5)^2 +$

$(2-3.5)^2 + (3-3.5)^2 = 0.0625 + 0.5625 + 10.5625 + 0.0625 +$

$1.5625 + 0.0625 + 3.0625 + 3.0625 + 0.25 + 6.25 + 3.0625 +$

$0.25 = 28.8125$ (28.25)

③ $SSA = \sum n_i (\bar{y}_i - \bar{\bar{y}})^T (\bar{y}_i - \bar{\bar{y}}) = SST - SSE = 28.8125 - 20.8472$

$= 7.9653$ (7.38)

For y_2:-

① $SSE = \sum (y_{ij} - \bar{y}_i)^T (y_{ij} - \bar{y}_i) = \sum (y_{ij} - \bar{y}_i)^2$

$= (6-6.8)^2 + (7-6.8)^2 + (5-6.8)^2 + (8-6.8)^2 + (8-6.8)^2 + (5-5.667)^2 +$

$(7-5.667)^2 + (5-5.667)^2 + (6-3.25)^2 + (2-3.25)^2 + (1-3.25)^2 +$

$(4-3.25)^2 = 0.64 + 0.04 + 3.24 + 1.44 + 1.44 + 0.444 + 1.776 +$

$0.444 + 7.5625 + 1.5625 + 5.0625 + 0.5625 = 24.214$

② $SST = 52.66$

③ $SSA = 52.66 - 24.214 = 28.446.$

For both y_1 & y_2:-

① $SSE = (4 \times 6 - 4.6 \times 6.8) + (3 \times 7 - 4.6 \times 6.8) + (7 \times 5 - 4.6 \times 6.8)^2 +$

$\ldots\ldots + (3 \times 4 - 3.5 \times 3.25) = -7.32$

② $SST = (4 \times 6 - 3.75 \times 5.33) + (3 \times 7 - 3.75 \times 5.33) + \ldots$

$\ldots + (3 \times 4 - 3.75 \times 5.33) = 0.15$

③ $SSA = SST - SSE = 7.47$

MANOVA ONE WAY CLASSIFICATION TABLE:-

Source of variation	sum of squares	Degrees of freedom	Will's value	F-statistic
Regression	$\begin{bmatrix} 7.38 & 7.47 \\ 7.47 & 28.44 \end{bmatrix} = B$	$3-1=2$	$\lambda^* = \frac{\lvert W \rvert}{\lvert T \rvert}$	$F = \frac{12-3-1}{3-1} \cdot \frac{1-\sqrt{\lambda^*}}{\sqrt{\lambda^*}}$
error	$\begin{pmatrix} 20.87 & -7.32 \\ -7.32 & 24.22 \end{pmatrix} = W$	$12-3=9$	$= \frac{451.889}{1487.905}$	
Total	$\begin{pmatrix} 28.25 & 0.15 \\ 0.15 & 52.66 \end{pmatrix} = T$	$12-1=11$	$= 0.3037$	$= 3.26$

$\sim F_{2(3-1),2(12-3)}$

$\sim F_{4,6}$

Inference:-

The table value of F at 1% L.O.S for 4, 16 degrees of freedom is $F_{table} = 4.77$

$\therefore F_{cal} < F_{table}$, we accept H_0

Hence we conclude that there is homogenity in the groups and also among the groups.

i.e, the multivariate linear regression modal (MVLR) is applicable for future predictions.

Theory Questions:-

① Write the procedure to fit the Multivariate linear regression modal.

② Discuss the statistical Analysis of multivariate analysis of varience for ONE way classification (OR) What are the statistics commonly used in testing

the hypothesis in MANOVA.

③ what is the difference between ANOVA and MANOVA ?

Sol: ANOVA :-

→ ANOVA stands for analysis of variance.

→ In statistics, when two or more than two means are compared simultaneously, the statistical method used to make comparision is called ANOVA.

→ It is a method which gives values and the results which can be tested in order to determine if a relationship of any significance exists between different variables

→ the name ANOVA has been given to the comparision of means because in order to determine any relationship between the means, the variances are actually being compared under currently

procedure for ANOVA :-

$H_0 : \mu_1 = \mu_2 = \dots = \mu_K$

$H_1 : \mu_1 \neq \mu_2 \neq \dots \neq \mu_K$

Level of significance is α% (Given/choosen)

To test the above hypothesis, the ANOVA table is given by

Source of variation	sum of squares	degrees of freedom	mean sum of squares	variance ratio
Treatments	S^2_{tr}	$K-1$	$MS^2_{tr} = \frac{S^2_{tr}}{K-1}$	$F = \frac{MS^2_{tr}}{MS^2_e} \sim$
Error	S^2_e	$n-K$	$MS^2_e = \frac{S^2_e}{n-K}$	$F_{K-1, n-K}$
Total	S^2_T	$n-1$		

Inference:-

$F_{cal} \geq F_{\alpha, k-1, n-k}$, we reject H_0

Otherwise we accept H_0.

MANOVA :-

→ MANOVA stands for multivariate Analysis of Variance.

→ In statistics, the MANOVA methods contain multiple dependent variables. They help in determining the differences between either two or more dependent variables.

→ MANOVA assits in determining these differences simultaneously.

→ In MANOVA method determines the dependent variable gets significant effected by the change in independent variables, it also determines the interactions taking place amongst dependent variables.

Procedure:-

H_0: $(\mu_{11} = \mu_{12} \cdots) = (\mu_{21} = \mu_{22} \cdots) = \cdots = (\mu_{k1} = \mu_{k2} \cdots)$

H_1: $(\mu_{11} = \mu_{12} \cdots) \neq (\mu_{21} = \mu_{22} \cdots) \neq \cdots \neq (\mu_{k1} = \mu_{k2} \cdots)$

Level of significance α (Given/Choosen)

To test the above hypothesis the procedure is as follows:-

Source of variation	sum of squares	D.O.F	Will's value	F-statistic
Treatments	$\sum n_i(\bar{y}_i - \bar{\bar{y}})(\bar{y}_i - \bar{\bar{y}})^T = B$	K-1	$\Lambda^* = \frac{\lvert W \rvert}{\lvert T \rvert}$	$F = \frac{n-K-1}{n-K} * \frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}} \sim$
Error	$\sum\sum (y_{ij} - \bar{y}_i)^T(y_{ij} - \bar{y}_i) = W$	N-K		
Total	$\sum (y_{ij} - \bar{\bar{y}})^T(y_{ij} - \bar{\bar{y}}) = T$	N-1		$F_{p(K-1),\, p(n-K-1)}$

where

k indicates the number of treatments.

p indicates the number of values in each sub group.

n indicates total number of groups in the experiment.

Inference:-

1

Summary of MULR:-

① There will be more than one independent variable

② Simultaneous occurance of dependent variable

③ Independent variables are same

④ Estimation of the parameters is similar to multiple linear Regression model (MLR)

⑤ Model Adequacy Test is countless.

⑥ The Assumptions in MLR are equally important to MVLR.

⑦ Does not give much advantage over MLR but protects against the event handling.

⑧ The tests applied for Goodness of fit are

MLR - ANOVA, R^2.

MVLR - MANOVA.

Discriminant Analysis :- (Data classification)

Discriminant Ananlysis is a statistical technique used to classify observations into non overlapping groups based on one or more quantitative predictor variables.

There are several different ways to conduct the discriminant Analysis. One such approach is based on linear regression and the other is linear discriminant Analysis.

Discrimant Analysis Based on Linear Regression :-

When there are only 2 classification groups, the linear discrimant analysis based on linear regression is applied when

→ The dependent variable is a categorial variable (A categorical variable takes only 2 values like Yes/No, True/False, 0/1, Male/Female etc)

→ The dependent variable is expressed as a dummy variable (values having zeroes or ones)

→ Observations are assigned to groups based on wheather the predicted score is closure to 0 (or) 1.

→ The regression equation is called the discriminant function.

The biggest difference between discriminant analysis and standard multiple regression analysis is the use of categorical variable as a dependent variable.

Eg: The SAT is an apptitude test taken by high school juniors and seniors. The college administration use the SAT along with high school grade point avg. to predict the academic success in the college.

The following table shows the SAT and GPA scores for 10 students in a college and shows wheather each student ultimately graduated from the college.

SAT	GPA	Graduate
1300	2.7	Yes
1260	3.7	Yes
1220	2.9	Yes
1180	2.5	Yes
1060	3.9	Yes
1140	2.1	No
1100	3.5	No
1020	3.3	No
980	2.3	No
940	3.1	No

Define a discriminant function that classifies an incoming student with SAT 1000 and GPA 2.9 as graduate or non-graduate.

Sol:- The dependent variable (Graduate) is a categorial variable (having Yes/No).

To use that variable in regression Analysis, we have to convert it into a quantitative variable.

We can make a categorical variable into a quantitative variable through a dummy variable recoding, by replacing Yes with 1 and No with 0.

Now the data table looks like

x_1 SAT	x_2 GPA	y Graduate
1300	2.7	1
1260	3.7	1
1220	2.9	1
1180	2.5	1
1060	3.9	1
1140	2.1	0
1100	3.5	0
1020	3.3	0
980	2.3	0
940	3.1	0

$$y = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 1300 & 2.7 \\ 1 & 1260 & 3.7 \\ 1 & 1220 & 2.9 \\ 1 & 1180 & 2.5 \\ 1 & 1060 & 3.9 \\ 1 & 1140 & 2.1 \\ 1 & 1100 & 3.5 \\ 1 & 1020 & 3.3 \\ 1 & 980 & 2.3 \\ 1 & 940 & 3.1 \end{bmatrix}$$

let $\boxed{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2}$ where,

$$\hat{\beta} = (X^T X)^{-1}(X^T y) = \begin{bmatrix} -3.8392 \\ 0.0032 \\ 0.2395 \end{bmatrix}$$

$$\boxed{y = -3.8392 + 0.0032\,SAT + 0.2395\,GPA}$$

(SAT $\downarrow$ 1000, GPA $\downarrow$ 2.9)

Then $y = 0.05535$ and the value is near to zero

∴ The candidate will not be graduated.

Note:-
If the recoding of the categorical variable changes then the parametric values in the discriminant function changes. To have the commonality in the discriminant function, we apply the Fisher's Linear Discriminant Function.

17/08/2022

Eg:
and their qualities are measured in terms of curvature and diameter. Results of quality control by experts are given in the following table.

Curvature	Diameter	Quality control result
2.95	6.63	Passed
2.53	7.71	Passed
3.57	5.65	Passed
3.16	5.47	Passed
2.58	7.46	Not Passed
2.16	6.22	Not Passed
3.27	3.32	Not Passed

The new chip ring has curvature 2.81 and diameter 5.46. Can you solve this problem by employing Fisher's Linear Discriminant Analysis.

Sol:-

Step-1:-

$$X = \begin{bmatrix} 2.95 & 6.63 \\ 2.53 & 7.79 \\ 3.57 & 5.65 \\ 3.16 & 5.47 \\ 2.58 & 4.46 \\ 2.16 & 6.22 \\ 3.27 & 3.52 \end{bmatrix} \qquad y = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 2 \\ 2 \\ 2 \end{bmatrix}$$

$$u = [2.89 \quad 5.68]$$

$$x_1 = \begin{bmatrix} 2.95 & 6.63 \\ 2.53 & 7.79 \\ 3.57 & 5.65 \\ 3.16 & 5.47 \end{bmatrix}$$

$$u_1 = [3.0525 \quad 6.385]$$

$$x_2 = \begin{bmatrix} 2.58 & 4.46 \\ 2.16 & 6.22 \\ 3.27 & 3.52 \end{bmatrix}$$

$$u_2 = [2.67 \quad 4.733]$$

Step-2:- Polled co-variance Matrix.

$$C = \frac{(x-u)^T (x-u)}{n}$$

$$x - u = \begin{bmatrix} 0.06 & 0.95 \\ -0.36 & 2.11 \\ 0.68 & -0.03 \\ 0.27 & -0.21 \\ -0.31 & -1.22 \\ -0.73 & 0.54 \\ 0.38 & -2.16 \end{bmatrix}$$

$$C = \begin{bmatrix} 0.2059 & -0.2309 \\ -0.2309 & 1.6921 \end{bmatrix}$$

$$\bar{C}^{-1} = \begin{bmatrix} 5.7342 & 0.7824 \\ 0.7824 & 0.6977 \end{bmatrix}$$

- Fisher's Linear Discriminant Function is given by

$$f_i = u_i \bar{C}^{-1} x_k^T - \frac{1}{2} u_i \bar{C}^{-1} u_i^T + \ln(P_i)$$

$$f_1 = u_1 \bar{C}^{-1} x_k^T - \frac{1}{2} u_1 \bar{C}^{-1} u_1^T + \ln(P_1)$$

$$f_1 = [3.0525 \quad 6.385] \begin{bmatrix} 5.7342 & 0.7824 \\ 0.7824 & 0.6977 \end{bmatrix} \begin{bmatrix} 2.81 \\ 5.46 \end{bmatrix}$$

$$-\frac{1}{2}[3.0525 \quad 6.385]\begin{bmatrix}5.7342 & 0.7824\\ 0.7824 & 0.6977\end{bmatrix}\begin{bmatrix}3.0525\\ 6.385\end{bmatrix}+\ln\left(\frac{4}{7}\right)$$

$$f_1 = 100.5862 - 56.186 + (-0.2443) - 0.5596 = 43.8406$$

$$f_2 = \mu_2 \bar{c}^{-1} x_k^T - \frac{1}{2}\mu_2 \bar{c}^{-1}\mu_2^T + \ln\left(\frac{3}{7}\right)$$

$$f_2 = [2.67 \quad 4.733]\begin{bmatrix}5.7342 & 0.7824\\ 0.7824 & 0.6977\end{bmatrix}\begin{bmatrix}2.81\\ 5.46\end{bmatrix}$$

$$-\frac{1}{2}[2.67 \quad 4.733]\begin{bmatrix}5.7342 & 0.7824\\ 0.7824 & 0.6977\end{bmatrix}\begin{bmatrix}2.67\\ 4.733\end{bmatrix}+\ln\left(\frac{4}{7}\right)$$

$$f_2 = 43.86$$

$\therefore f_2 > f_1$, so we classify the new chip ring into the second group.

i.e, the new chip ring is not passed.

Eg: classify the following

samples of class $\omega_1 = X_1 = (x_1, x_2)$

$= \{(4,2)(2,4)(2,3)(3,6)(4,4)\}$

$\omega_2 = X_2 = (x_1, x_2)$

$= \{(9,10)(6,8)(9,5)(8,7)(10,8)\}$

$x = 4$

$$x = \begin{bmatrix} 4 & 2 \\ 2 & 4 \\ 2 & 3 \\ 3 & 6 \\ 4 & 4 \\ 9 & 10 \\ 6 & 8 \\ 9 & 5 \\ 8 & 7 \\ 10 & 8 \end{bmatrix} \quad x_1 = \begin{bmatrix} 4 & 2 \\ 2 & 4 \\ 2 & 3 \\ 3 & 6 \\ 4 & 4 \end{bmatrix} \quad x_2 = \begin{bmatrix} 9 & 10 \\ 6 & 8 \\ 9 & 5 \\ 8 & 7 \\ 10 & 8 \end{bmatrix}$$

$$u_1 = [3 \;\; 3.8] \qquad u_2 = [8.4 \;\; 7.6]$$

$$u = [5.7 \;\; 5.7]$$

The polled co-valience matrix,

$$\boxed{c = \frac{(x-u)^T (x-u)}{n}}$$

$$x - u = \begin{bmatrix} -1.7 & -3.7 \\ -3.7 & -1.7 \\ -3.7 & -2.7 \\ -2.7 & 0.3 \\ -1.7 & -1.7 \\ 3.3 & 4.3 \\ 0.3 & 2.3 \\ 3.3 & -0.7 \\ 2.3 & 1.3 \\ 4.3 & 2.3 \end{bmatrix}$$

$$c = \begin{bmatrix} 8.61 & 5.01 \\ 5.01 & 5.81 \end{bmatrix}$$

$$c^{-1} = \begin{bmatrix} 0.20331 & -0.2010 \\ -0.2010 & 0.3454 \end{bmatrix}$$

- Fisher's Linear Discriminant Function is given by

$$\boxed{f_i = u_i \bar{c}^{-1} x_k^T - \frac{1}{2} u_i \bar{c}^{-1} u_i^T + \ln(P_i)}$$

$$f_1 = u_1 \bar{c}^{-1} x_k^T - \frac{1}{2} u_1 \bar{c}^{-1} u_1^T + \ln(P_1)$$

$$f_1 = [3 \;\; 3.8] \begin{bmatrix} 0.2331 & -0.2010 \\ -0.2010 & 0.3454 \end{bmatrix} \begin{bmatrix} 5 \\ 6 \end{bmatrix} - \frac{1}{2} [3 \;\; 3.8]$$

$$\begin{bmatrix} 0.2331 & -0.2010 \\ -0.2010 & 0.3454 \end{bmatrix} \begin{bmatrix} 3 \\ 3.8 \end{bmatrix} + \ln\left(\frac{5}{10}\right)$$

$$= 3.9346 - 1.2513 + (-0.6931) = 1.9902$$

$$f_2 = \mu_2 C^{-1} x_k^T - \frac{1}{2} \mu_2 C^{-1} \mu_2^T + \ln\left(\frac{5}{10}\right)$$

$$= [8.4 \;\; 7.6] \begin{bmatrix} 0.2331 & -0.2010 \\ -0.2010 & 0.3454 \end{bmatrix} \begin{bmatrix} 5 \\ 6 \end{bmatrix}$$

$$- \frac{1}{2} [8.4 \;\; 7.6] \begin{bmatrix} 0.2331 & -0.2010 \\ -0.2010 & 0.3454 \end{bmatrix} \begin{bmatrix} 8.4 \\ 7.6 \end{bmatrix} + \ln\left(\frac{5}{10}\right)$$

$$= 7.7720 - 5.3670 - 0.6931 = 1.7119$$

$\therefore f_1 > f_2$, the observation (5,6) is classify into first group.

(3) compute the linear discriminant function and classify the observation 5.1 & 3.2.

Samples of class $\omega_1 = \{(1,2)\ (2,3)\ (3,3)\ (4,5)\ (5,5)\}$

$\omega_2 = \{(4,2)\ (5,0)\ (5,2)\ (3,2)\ (5,3)\ (6,3)\}$

$$x = \begin{bmatrix} 1 & 2 \\ 2 & 3 \\ 3 & 3 \\ 4 & 5 \\ 5 & 5 \\ 4 & 2 \\ 5 & 0 \\ 5 & 2 \\ 3 & 2 \\ 5 & 3 \\ 6 & 3 \end{bmatrix} \qquad \mu = [3.909 \;\; 2.7272]$$

$$x_1 = \begin{bmatrix} 1 & 2 \\ 2 & 3 \\ 3 & 3 \\ 4 & 5 \\ 5 & 5 \end{bmatrix} \qquad \mu_1 = [3 \;\; 3.6]$$

$$x_2 = \begin{bmatrix} 4 & 2 \\ 5 & 0 \\ 5 & 2 \\ 3 & 2 \\ 5 & 3 \\ 6 & 3 \end{bmatrix} \qquad \mu_2 = [4.666 \;\; 2]$$

the polled covariance matrix,

$$C = \frac{(x-u)^T (x-u)}{n}$$

$$x-u = \begin{bmatrix} -2.909 & -0.7272 \\ -1.909 & 0.2728 \\ -0.909 & 0.2728 \\ 0.1091 & 2.2728 \\ 1.1091 & 2.2728 \\ -0.091 & -0.7272 \\ 1.091 & -2.7272 \\ 1.091 & -0.7272 \\ -0.909 & -0.7272 \\ 1.091 & 0.2728 \\ 2.091 & 0.2728 \end{bmatrix}$$

$$C = \begin{bmatrix} 2.0826 & 0.1570 \\ 0.1570 & 1.8347 \end{bmatrix}$$

$$C^{-1} = \begin{bmatrix} 0.4832 & -0.0413 \\ -0.0413 & 0.5485 \end{bmatrix}$$

~~$u = [0.0744$~~

- Fishers linear Discriminant Function is given by

$$f_i = u_i \bar{C}^{-1} x_k^T - \frac{1}{2} u_i \bar{C}^{-1} u_i^T + \ln(P_i)$$

$$f_1 = u_1 \bar{C}^{-1} x_k^T - \frac{1}{2} u_1 \bar{C}^{-1} u_1^T + \ln(P_1)$$

$$f_1 = [3 \ 3.6] \begin{bmatrix} 0.4832 & -0.0413 \\ -0.0413 & 0.5485 \end{bmatrix} \begin{bmatrix} 5.1 \\ 3.2 \end{bmatrix}$$

$$- \frac{1}{2} [3 \ 3.6] \begin{bmatrix} 0.4832 & -0.0413 \\ -0.0413 & 0.5485 \end{bmatrix} \begin{bmatrix} 3 \\ 3.6 \end{bmatrix} + \ln\left(\frac{5}{11}\right)$$

$$= 6.486$$

$f_2 = 7.39$

$\therefore f_2 > f_1$

Therefore, the observation (5.1, 3.2) is classified into second group.

④ consider the 2 datasets

$x_1 = \begin{bmatrix} 3 & 7 \\ 2 & 4 \\ 4 & 7 \end{bmatrix}$ $x_2 = \begin{bmatrix} 6 & 9 \\ 5 & 7 \\ 4 & 8 \end{bmatrix}$ for which $\bar{x}_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}$ & $\bar{x}_2 = \begin{bmatrix} 5 \\ 8 \end{bmatrix}$

and spooled $= \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$. classify the observation

$x_0' = [2 \ 7]$ as a population π_1 (or) π_2.

Sol:- Given,

$x_1 = \begin{bmatrix} 3 & 7 \\ 2 & 4 \\ 4 & 7 \end{bmatrix}$ $x_2 = \begin{bmatrix} 6 & 9 \\ 5 & 7 \\ 4 & 8 \end{bmatrix}$ $C = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$

$u_1 = [3 \ 6]$ $u_2 = [5 \ 8]$ $C^{-1} = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}$

· Fisher's linear discriminant function:-

$$\boxed{f_i = u_i C^{-1} x_k^T - \frac{1}{2} u_i C^{-1} u_i^T - \ln(P_i)}$$

$$f_1 = u_1 C^{-1} x_k^T - \frac{1}{2} u_1 C^{-1} u_1^T - \ln(P_i)$$

$$= [3 \ 6] \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 7 \end{bmatrix} - \frac{1}{2} [3 \ 6] \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 6 \end{bmatrix} + \ln\left(\frac{3}{6}\right)$$

$$= 21 - \frac{1}{2}(18) - 0.693 = 11.30$$

$$f_2 = u_2 C^{-1} x_k^T - \frac{1}{2} u_2 C^{-1} u_2^T - \ln(P_2)$$

$$= [5 \ 8] \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 7 \end{bmatrix} - \frac{1}{2} [5 \ 8] \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 5 \\ 8 \end{bmatrix} - \ln\left(\frac{3}{6}\right)$$

$$= 7.30$$

$\therefore f_1 > f_2$

the observation (2,7) in to first group (or) population.

$\therefore$ It is into π_1 population.

(5) Consider the observations on $p=2$ variables from $g=3$ populations assuming that the population have a common covariance matrix (Σ), Given that $P_1 = P_2 = 0.25$ and $P_3 = 0.50$ · classify $[-2\ -1]$ using discriminant scores.

$$x_1 = \begin{bmatrix} -2 & 5 \\ 0 & 3 \\ -1 & 1 \end{bmatrix} \quad x_2 = \begin{bmatrix} 0 & 6 \\ 2 & 4 \\ 1 & 2 \end{bmatrix} \quad x_3 = \begin{bmatrix} 1 & -2 \\ 0 & 0 \\ -1 & -4 \end{bmatrix}$$

$$\mu_1 = [-1\ \ 3] \quad \mu_2 = [1\ \ 4] \quad \mu_3 = [0\ -2]$$

$$\boxed{c = \frac{(x-\mu)^T (x-\mu)}{n}}$$

$$x = \begin{bmatrix} -2 & 5 \\ 0 & 3 \\ 1 & 1 \\ 0 & 6 \\ 2 & 4 \\ 1 & 2 \\ 1 & -2 \\ 0 & 0 \\ -1 & -4 \end{bmatrix} \qquad x-\mu = \begin{bmatrix} -2 & 3.33 \\ 0 & 1.33 \\ 1 & -0.66 \\ 0 & 4.33 \\ 2 & 2.33 \\ 1 & 0.33 \\ 1 & -3.66 \\ 0 & -1.66 \\ -1 & -5.66 \end{bmatrix}$$

$$c = \begin{bmatrix} 1.33 & 0.11 \\ 0.11 & 9.56 \end{bmatrix}$$

$$c^{-1} = \begin{bmatrix} 0.75 & -0.008 \\ -0.008 & 0.10 \end{bmatrix}$$

$$\mu = [0\ \ 1.66]$$

· Fishers linear Discriminant Function is given by

$$\boxed{f_i = \mu_i c^{-1} x_k^T - \frac{1}{2} \mu_i c^{-1} \mu_i^T + \ln(P_i)}$$

$$f_1 = \mu_1 c^{-1} x_k^T - \frac{1}{2} \mu_1 c^{-1} \mu_1^T + \ln(P_i)$$

$$= [-1 \;\; 3]\begin{bmatrix} 0.75 & -0.008 \\ -0.008 & 0.10 \end{bmatrix}\begin{bmatrix} -2 \\ -1 \end{bmatrix} - \frac{1}{2}[-1 \;\; 3]\begin{bmatrix} 0.75 & -0.008 \\ -0.008 & 0.10 \end{bmatrix}\begin{bmatrix} -1 \\ 3 \end{bmatrix} + \ln(0.25)$$

$$= 1.24 - \frac{1}{2}(1.698) - 1.386$$

$$= -1.028$$

$$f_2 = \mu_2 \bar{c}^{-1} x_K^T - \frac{1}{2}\mu_2 \bar{c}^{-1} \mu_2^T + \ln(P_2)$$

$$= [1 \;\; 4]\begin{bmatrix} 0.75 & -0.008 \\ -0.008 & 0.10 \end{bmatrix}\begin{bmatrix} -2 \\ -1 \end{bmatrix} - \frac{1}{2}[1 \;\; 4]\begin{bmatrix} 0.75 & -0.008 \\ -0.008 & 0.10 \end{bmatrix}\begin{bmatrix} 1 \\ 4 \end{bmatrix} + \ln(0.25)$$

$$= -1.828 - \frac{1}{2}(2.286) - 1.386 = -4.357$$

$$f_3 = \mu_3 \bar{c}^{-1} x_K^T * - \frac{11}{2}\mu_3 \bar{c}^{-1} \mu_3^T + \ln(P_3)$$

$$= [0 \;\; -2]\begin{bmatrix} 0.75 & -0.008 \\ -0.008 & 0.10 \end{bmatrix}\begin{bmatrix} -2 \\ -1 \end{bmatrix} - \frac{1}{2}[0 \;\; -2]\begin{bmatrix} 0.75 & -0.008 \\ -0.008 & 0.10 \end{bmatrix}\begin{bmatrix} 0 \\ -2 \end{bmatrix} + \ln(05)$$

$$= -0.728$$

$\therefore f_3 > f_1 > f_2$,

The observation $(-2, -1)$ belongs to the 3rd group.

24/08/2022

⑥ Consider the observations on $P=2$ variables from $g=3$ populations. Assuming that the populations have common covariance matrix(Σ). Find the fishee's discriminants, the data path

$$x_1 = \begin{bmatrix} -2 & 5 \\ 0 & 3 \\ -1 & 1 \end{bmatrix} \qquad x_2 = \begin{bmatrix} 0 & 6 \\ 2 & 4 \\ 1 & 2 \end{bmatrix} \qquad x_3 = \begin{bmatrix} 1 & -2 \\ 0 & 0 \\ -1 & -4 \end{bmatrix}$$

$$\bar{x}_1 = \begin{bmatrix} -1 \\ 3 \end{bmatrix} \qquad \bar{x}_2 = \begin{bmatrix} 1 \\ 4 \end{bmatrix} \qquad \bar{x}_3 = \begin{bmatrix} 0 \\ -2 \end{bmatrix}$$

Find ii) solve $s \leq \min(g-1, p)$

i) $\bar{x}$, B and W

iii) and eigen values & fisher's discriminant functions

isd:- let $x = \begin{bmatrix} -2 & 5 \\ 0 & 3 \\ -1 & 1 \\ 0 & 6 \\ 2 & 4 \\ 1 & 2 \\ 1 & -2 \\ 0 & 0 \\ -1 & -4 \end{bmatrix}$

$$B = \sum (\bar{x}_i - \bar{x})^T (\bar{x}_i - \bar{x})$$

$$\bar{x}_i - \bar{x} = \begin{bmatrix} -1 & 1.33 \\ 1 & 2.33 \\ 0 & -3.66 \end{bmatrix}$$

$$B = \begin{bmatrix} 2 & 1 \\ 1 & 20.67 \end{bmatrix}$$

$$W = \sum (x_{ij} - \bar{x}_{ij})^T (x_{ij} - \bar{x}_{ij})$$

$$\bar{x} = \begin{bmatrix} 0 \\ 1.66 \end{bmatrix}$$

$$x_{ij} - \bar{x}_i = \begin{bmatrix} -2+1 & 5-3 \\ 0+1 & 3-3 \\ -1+1 & 1-3 \\ 0-1 & 6-4 \\ 2-1 & 4-4 \\ 1-1 & 2-4 \\ 1-0 & -2+2 \\ 0-0 & 0+2 \\ -1-0 & -4+2 \end{bmatrix} = \begin{bmatrix} -1 & 2 \\ 1 & 0 \\ 0 & -2 \\ -1 & 2 \\ 1 & 0 \\ 0 & -2 \\ 1 & 0 \\ 0 & 2 \\ -1 & -2 \end{bmatrix}$$

$$W = \begin{bmatrix} 6 & -2 \\ -2 & 24 \end{bmatrix}$$

ii →

Note:- To solve $s \leq \min(g-1, p)$

$$s \leq \min(3-1, 2) = \min(2, 2) = 2$$

we have to find two non-zero eigen values for $W^{-1}B$.

$$W^{-1}B = \frac{1}{140}\begin{bmatrix} 24 & 2 \\ 2 & 6 \end{bmatrix} * \begin{bmatrix} 2 & 1 \\ 1 & 20.67 \end{bmatrix}$$

$$W^{-1}B = \begin{bmatrix} 0.3571 & 0.4667 \\ 0.0714 & 0.8997 \end{bmatrix}$$

To Find the eigen values of $W^{-1}B$

$$|W^{-1}B - \lambda I| = 0$$

$$\left| \begin{bmatrix} 0.3571 & 0.4667 \\ 0.0714 & 0.8997 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right| = 0$$

$$\begin{vmatrix} 0.3571-\lambda & 0.4667 \\ 0.0714 & 0.8997-\lambda \end{vmatrix} = 0$$

$$(0.3571-\lambda)(0.8997-\lambda) - (0.0714)(0.4667) = 0.$$

$$\lambda^2 - 1.2571\lambda + 0.2881 = 0$$

$$\frac{-b \pm \sqrt{b^2-4ac}}{2a}$$

$$\boxed{\lambda_1 = 0.9556,\ \lambda_2 = 0.3015}$$

To Find the eigen vectors,

$$\begin{pmatrix} 0.3571-\lambda & 0.4667 \\ 0.0714 & 0.9-\lambda \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0$$

$$\begin{pmatrix} -0.5985 & 0.4667 \\ 0.0714 & -0.0555 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0$$

$$x_1(-0.5985) + (0.4667)x_2 = 0 \rightarrow (1)$$

$$0.0714x_1 - 0.0555x_2 = 0 \rightarrow (2)$$

$x_1 = 0, \; x_2 = 0$

$$\frac{|-0.5985|}{\sqrt{(-0.5985)^2 + (0.4667)^2}}, \quad \frac{|0.4667|}{\sqrt{(-0.5985)^2 + (0.4667)^2}}$$

$$(0.7886, 0.6149)$$

is a set of eigen vector for $\lambda_1 = 0.9556$.

$\lambda_2 = 0.3015$

$$\begin{pmatrix} 0.0556 & 0.4667 \\ 0.0714 & 0.5985 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0$$

$$0.0556 x_1 + 0.4667 x_2 = 0 \quad \checkmark$$

$$0.0714 x_1 + 0.5985 x_2 = 0$$

$$x_1 = 0, \; x_2 = 0$$

$$\frac{|0.0556|}{\sqrt{(0.0556)^2 + (0.4667)^2}}, \quad \frac{|0.4667|}{\sqrt{(0.0556)^2 + (0.4667)^2}}$$

$$(0.1184, 0.9930)$$

is a set of eigen vector for $\lambda_2 = 0.3015$.

Now, the fishers discriminant function

$$f_1 = 0.7886 x_1 + 0.6149 x_2$$

$$f_2 = 0.1184 x_1 + 0.9930 x_2$$

25/2
(11) Classify a new observation into one of the populations, let us assign an observation x_0 into one of $g=3$ populations π_1, π_2, π_3. Given the following hypothetical probabilities, misclassification class and density values.

		True Population		
		π_1	π_2	π_3
classify as:	π_1	$c(1/1)=0$	$c(1/2)=500$	$c(1/3)=100$
	π_2	$c(2/1)=10$	$c(2/2)=0$	$c(2/3)=50$
	π_3	$c(3/1)=50$	$c(3/2)=200$	$c(3/3)=0$
Pair Probab		$P_1=0.05$	$P_2=0.60$	$P_3=0.35$
Densities at x_0		$f_1(x_0)=0.01$	$f_2(x_0)=0.85$	$f_3(x_0)=2$

Use minimum expected cost of Misclassification (ECM) procedures to classify a new observation x_0

Procedure 1:-

ECM based on pair probabilities, densities and misclassification costs.

$$\sum_{\substack{i=1 \\ i \neq K}}^{n} P_i \cdot f_i(x_0) \cdot c(K|i)$$

Allot the new observation to the smallest value.

K=1, $P_2 \cdot f_2(x_0) \cdot c(1/2) + P_3 f_3(x_0) \cdot c(1/3)$

$0.60 * 0.85 * 500 + 0.35 * 2 * 100 = 325$

K=2, $P_1 f_1(x_0) \cdot c(2/1) + P_3 f_3(x_0)\, c(2/3)$

$0.05 * 0.01 * 10 + 0.35 * 2 * \frac{100}{50} = \cancel{70.005}\ 35.0005$

K=3, $P_1 f_1(x_0) \cdot c(3/1) + P_2 f_2(x_0)\, c(3/2)$

$0.05 * 0.01 * 50 + 0.60 * 0.85 * 200 = 102.025$

If class of misclassification are equal then the formula based on is given by

$P_1 f_1(x_0) = 0.0005$

$P_2 f_2(x_0) = 0.510$

$P_3 f_3(x_0) = 0.7$

Allot the observation to the biggest value, so allot the observation to the third population.

If all ~~classes~~ costs of misclassification are equal then we use baye's theorm.

$$P(\pi_1/x_0) = \frac{P_1 f_1(x_0)}{\Sigma P_i f_i(x_0)}$$

$$= \frac{0.05 * 0.01}{(0.05 * 0.01) + (0.60 * 0.85) + (0.35 * 2)} = 0.0004$$

$$P(\pi_2/x_0) = \frac{P_2 f_2(x_0)}{\sum_{i=1}^{3} P_i f_i(x_0)} = \frac{0.60 * 0.85}{\ldots} = 0.4197$$

$P(\pi_3/x_0) = 0.5761$

Allot the observation to the biggest value.
So allot to the third population.

Confusion Matrix :-

The apparent error rate in the confusion matrix is calculated by $\boxed{\frac{n_{1m} + n_{2m}}{n_1 + n_2}}$

Actual class	Predicted class π_1	Predicted class π_2	
π_1	n_{1c} → correct classif (Popul.)	n_{1m} → misclassif (Popul.)	n_1
π_2	n_{2m} → misclassification (Popu.)	n_{2c} → correct classific (Pop)	n_2

Eg:- Calculate apparent error rate.

Actual class	Predicted π_1	Predicted π_2	
π_1	10	2	12
π_2	2	10	12

$\therefore$ Apparent error rate $= \frac{2+2}{12+12} = 16.7\ \%$

Q.4 Principal Component Analysis:- (Data Reduct^n)

→ Regardless ones area of study, the first decision that the researcher bases concern to which variables are to be measured. Generally the researcher adopts a heat and miss strategy, collecting information on the tippically large number of variables that might be relavent

→ How many variables are to be measured, some pratical problems may arise for eg:- with as as 10 variables there are $^{10}C_2 = 45$ correlation coefficients, with 20 variables, there are $^{20}C_2 = 190$ correlation coefficients to be considered.

→ so, the number of correlation coefficients keeps on accelerating as the number of variables increases. Obviously, with large number of variables, the correlation coefficients are so large as to beyond the comprehension and some data reduction technique that can sysmatically summarize the large correlation matrices is needed which is possible with the principle component analysis.

Q:- Is there any technique which seeves the purpose of data reduction?

Sol:- The data reduction is possible by PCA (principal Component Analysis).

① The basic idea of principal component Analysis is to describe the variation of a set of multivariate data in terms of a set of unrelated variables each of which is a particular linear combination of a set of correlated variables.

② The new variables are derived in decreasing order of importance so that the first transformed variable accounts for as much as possible of the variation in the original data

③ The usual objective of this type of analysis is to see wheather the first few components accounts for most of the variation in the original data.

④ If they do so, then they can be used to summarize the data with little loss of information, thus providing a reduction in the dimensionality of the ~~data~~ Program which might be useful in simplifying the later analysis

Simple Algorithm for PCA:-

Step 1:- Standardize the data set

Step 2:- calculate the covariance matrix of the standardized dataset

Step 3:- Calculate the eigen values and eigen vectors. of the covariance matrix.

Step 4:- sort the eigen values and their corresponding eigen vectors.

Step 5:- pick up k eigen values to form a eigen matrix of eigen vectors

Step 6:- Transform the original matrix to find the new data set

Steps to PCA:-

Step 1:- Decide the variables to be studied

For eg:- Let $x_1, x_2, \ldots x_p$ be the p number of variables under step

Step 2:- collect the observations under each variable

$$\begin{matrix} x_{11} & x_{12} & \cdots & x_{p1} \\ x_{21} & x_{22} & \cdots & x_{p2} \\ \vdots & & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{pn} \end{matrix}$$

Step 3:- Calculate the variance covariance matrix which is given by

$$C = \frac{(x-\mu)^T(x-\mu)}{n}$$

Step 4:- Obtain the eigen values and eigen vectors of the variance covariance matrix

Let $\lambda_1, \lambda_2 \ldots \lambda_p$ be the eigen values.

Sort the eigen values in the descending order.

$$\lambda_{(1)} \geq \lambda_{(2)} \geq \ldots \lambda_{(p)}$$

Find the corresponding eigen vectors of the eigen values.

Step 5:- How many principal components to retain?

Principal component	Variance explained	
z_1	λ_1	$\frac{\lambda_1}{\sum \lambda_i} \times 100$
z_2	λ_2	$\frac{\lambda_1+\lambda_2}{\sum \lambda_i} \times 100$
⋮	⋮	⋮
z_p	λ_p	$\frac{\lambda_1+\lambda_2+\cdots+\lambda_p}{\sum \lambda_i} \times 100$

Note:-

Generally we take the threshholds as the 90%, 95% or 99% (depends upon the researcher or the statistician)

Step 6:- Transform the variables into new set of reduced variables say $z_1, z_2, \ldots z_q$ where $(q<p)$

Q:- Find the principal components of the following data.

x_1	x_2
2	4
1	3
0	1
-1	0.5

$$\text{let } X = \begin{bmatrix} 2 & 4 \\ 1 & 3 \\ 0 & 1 \\ -1 & 0.5 \end{bmatrix} \qquad x-\mu = \begin{bmatrix} 1.5 & 1.875 \\ 0.5 & 0.875 \\ -0.5 & -1.125 \\ -1.5 & -1.625 \end{bmatrix}$$

$$\mu = [0.5 \quad 2.125]$$

- Variance covariance matrix is given by

$$C = \frac{(x-\mu)^T(x-\mu)}{n}$$

$$C = \begin{bmatrix} 1.25 & 1.5625 \\ 1.5625 & 2.0468 \end{bmatrix}$$

To find the eigen values of varience covarience matrix C is

$$|C - \lambda I| = 0$$

$$\begin{vmatrix} 1.25-\lambda & 1.5625 \\ 1.5625 & 2.0468-\lambda \end{vmatrix} = 0$$

$$(1.25-\lambda)(2.0468-\lambda) - (1.5625)^2 = 0$$

$$2.5585 - 1.25\lambda - 2.0468\lambda + \lambda^2 - 2.4414 = 0$$

$$\lambda^2 - 3.2968\lambda + 0.1171 = 0$$

$$\boxed{\lambda_1 = 3.2608 \mid 0.0359 = \lambda_2}$$

~~To find the eig~~

Principal component Analysis Table:- (How many components are to be retained):

Principal component	varience explained	cumulative proportion of total varience
z_1	$\lambda_1 = 3.261$	$\frac{3.261}{3.2969} \times 100 = 98.91\%$
z_2	$\lambda_2 = 0.0359$	$\frac{3.2969}{3.2969} \times 100 = 100\%$
	$\Sigma\lambda_i = 3.2969$	

Note:- The threshold is fixed as 99% (depends upon the researcher)

Here we have to retain 2 principal components.

To find the eigen vectors for $\lambda_1 = 3.261$

$$\begin{pmatrix} 1.25 - 3.261 & 1.5625 \\ 1.5625 & 2.0468 - \lambda \\ & 3.261 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0$$

$$-2.011 x_1 + 1.5625 x_2 = 0 \rightarrow (1)$$

$$1.5625 x_1 - 1.2142 x_2 = 0 \rightarrow (2)$$

$$x_1 = 0, \quad x_2 = 0$$

$$+2.011 x_1 = +1.5625 x_2$$

$$\frac{x_1}{1.5625} = \frac{x_2}{2.011}$$

Normalization, $\frac{1.5625}{\sqrt{1.5625^2 + 2.011^2}}, \frac{2.011}{\sqrt{1.5625^2 + 2.011^2}}$

$$\boxed{0.6135, 0.7396}$$

is a set of eigen vector.

Find a eigen vector for $\lambda_2 = 0.0359$

$$\begin{pmatrix} 1.25 - 0.0359 & 1.5625 \\ 1.5625 & 2.0468 - 0.0359 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0$$

$$1.2141 x_1 + 1.5625 x_2 = 0 \rightarrow (1)$$

$$1.5625 x_1 + 2.0109 x_2 = 0 \rightarrow (2)$$

$$1.2141 x_1 = -1.5625 x_2$$

$$\frac{x_1}{-1.5625} = \frac{+x_2}{1.2141}$$

Normalization, $\frac{-1.5625}{\sqrt{1.5625^2 + 1.2141^2}}, \frac{1.2141}{\sqrt{1.5625^2 + 1.2141^2}}$

$[-0.7896, 0.6135]$

is a set of eigen vector when $\lambda_2 = 0.0359$

$\therefore z_1 = 0.6135 x_1 + 0.7896 x_2$

$z_2 = -0.7896 x_1 + 0.6135 x_2$

② Find the principal components of the following data.

x_1	x_2	x_3
90	60	90
90	90	30
60	60	60
60	60	90
30	30	30

sol:

$$x = \begin{bmatrix} 90 & 60 & 90 \\ 90 & 90 & 30 \\ 60 & 60 & 60 \\ 60 & 60 & 90 \\ 30 & 30 & 30 \end{bmatrix} \quad x - u = \begin{bmatrix} 24 & 0 & 30 \\ 24 & 30 & -30 \\ -6 & 0 & 0 \\ -6 & 0 & 30 \\ -36 & -30 & -30 \end{bmatrix}$$

$$u = [66 \quad 60 \quad 60]$$

- Variance covarience matrix is given by

$$c = \frac{(x-u)^T (x-u)}{n}$$

$$c = \begin{bmatrix} 504 & 360 & 180 \\ 360 & 3160 & 0 \\ 180 & 0 & 720 \end{bmatrix}$$

To find eigen values, the characteristic eqn $|A - \lambda I| = 0$

$$\Rightarrow \begin{vmatrix} 504-\lambda & 360 & 180 \\ 360 & 3160-\lambda & 0 \\ 180 & 0 & 720-\lambda \end{vmatrix} = 0$$

$$\Rightarrow (504-\lambda)((3160-\lambda)(720-\lambda)) - 360(360)(720-\lambda) \quad (\text{or})$$

$$+ 180(-(180)(3160-\lambda)) = 0$$

$$|A| = \begin{vmatrix} 504 & 360 & 180 \\ 360 & 360 & 0 \\ 180 & 0 & 720 \end{vmatrix}$$

$$\bar{A}^{-1} = \begin{bmatrix} 0.0101 & -0.0101 & 0.0025 \\ -0.0101 & 0.0128 & 0.0025 \\ -0.0025 & 0.0025 & 0.0202 \end{bmatrix}$$

$= 25660880$

$$\lambda^3 - \text{trace}(A)\cdot\lambda^2 + \text{trace}(\text{adj}(A))\cdot\lambda - \det(A) = 0$$

$$\therefore \text{adj}(A) = \bar{A}^{-1}|A|$$

$$\lambda^3 - 1584\lambda^2 + [20431][25660880]\cdot\lambda - 25660880 = 0$$

$$641263.39\lambda$$

$\lambda_1 = 911.027$, $\lambda_2 = 628.13$, $\lambda_3 = 44.842$ are the eigen values.

How many principal components to be Retained?

Principal Component	Variance explained	Cumulative proportion of total variance
z_1	$\lambda_1 = 911.027$	$\frac{911.027}{1584} \times 100 = 57.5\%$
z_2	$\lambda_2 = 628.13$	$\frac{911.027 + 628.13}{1584} \times 100 = 97.17\%$
z_3	$\lambda_3 = 44.84$	$\frac{911.07 + 628.13 + 44.842}{1584} \times 100 = 100\%$
	$\sum\lambda_1 = 1584$	

Note: If the threshold has fixed as 95%, we retain the first two principal components.

Cont. in lab cs observation books

Continuation:-

To find the eigen vector when $\lambda_1 = 911.02$

$$\begin{pmatrix} 504-911.02 & 360 & 180 \\ 360 & 360-911.02 & 0 \\ 180 & 0 & 720-911.02 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = 0$$

$$-407.02x_1 + 360x_2 + 180x_3 = 0 \rightarrow (1)$$

$$360x_1 - 551.02x_2 = 0 \rightarrow (2)$$

$$180x_1 - 191.02x_3 = 0 \rightarrow (3)$$

$$\begin{array}{cccc} x_1 & x_2 & x_3 & x_1 \\ -407.02 & 360 & 180 & -407.02 \\ 360 & -551.02 & 0 & 360 \end{array}$$

$$\frac{x_1}{0+(180)(551.02)} = \frac{x_2}{-(180)(360)} = \frac{x_3}{(360)(360)-(551.02)(407.02)}$$

$$\frac{x_1}{99183.6} = \frac{+x_2}{+64800} = \frac{x_3}{-94676.16}$$

Normalizing the value

$$= \frac{99183.6}{\sqrt{9837386509 + 4199040000 + 8963575272}}$$

$$= \frac{99183.6}{151657.51} = 0.6539$$

$\boxed{0.6539, 0.4292, 0.6210}$ is the eigen vector when $\lambda_1 = 911.02$

To find the eigen vector for $\lambda_2 = 628.13$

$$\begin{pmatrix} 504-628.13 & 360 & 180 \\ 360 & 360-628.13 & 0 \\ 180 & 0 & 720-628.13 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = 0$$

$$-121.13 + 360x_2 + 180x_3 = 0 \rightarrow (1)$$

$$360x_1 * -268.13x_2 = 0 \rightarrow (2)$$

$$180x_1 - 91.87x_3 = 0 \rightarrow (3)$$

$$\begin{array}{cccc} x_1 & x_2 & x_3 & x_1 \\ -121.13 & 360 & 180 & -121.13 \\ 360 & -268.73 & 0 & 360 \end{array}$$

$$\frac{x_1}{32478.58 - 129600} = \frac{+x_2}{+64800} = \frac{x_3}{-97121.42}$$

$$48263.4$$

Normalizing the value

$$= \frac{48263.4}{\sqrt{2329355780 + 4199040000 + 9432570223}}$$

$$= \frac{48263.4}{126336.71} = 0.3820$$

0.3820, 0.5129, -0.7687 is the eigen vector when $\lambda_2 = 628.13$

- Now, the principal components equations are

$$z_1 = 0.6539x_1 + 0.4278x_2 + 0.6243x_3$$

$$z_2 = 0.3820x_1 + 0.5164x_2 - 0.7644x_3$$

03/09/22

Q3. Find the principal components as follows.

x_1	x_2	x_3
7	4	3
4	1	8
6	3	5
8	6	1
8	5	7
7	2	9
5	3	3
9	5	8
7	4	5
8	2	2

$$x = \begin{bmatrix} 7 & 4 & 3 \\ 4 & 1 & 8 \\ 6 & 3 & 5 \\ 8 & 6 & 1 \\ 8 & 5 & 7 \\ 7 & 2 & 9 \\ 5 & 3 & 3 \\ 9 & 5 & 8 \\ 7 & 4 & 5 \\ 8 & 2 & 2 \end{bmatrix}$$

$$\mu = [6.9 \quad 3.5 \quad 5.1]$$

$$x - \mu = \begin{bmatrix} 0.1 & 0.5 & -2.1 \\ -2.9 & -2.5 & 2.9 \\ -0.9 & -0.5 & 1.9 \\ 1.1 & 2.5 & -4.1 \\ 1.1 & 1.5 & 1.9 \\ 0.1 & -1.5 & 3.9 \\ -1.9 & -0.5 & -2.1 \\ 2.1 & 1.5 & 2.9 \\ 0.1 & 0.5 & -0.1 \\ 1.1 & -1.5 & -3.1 \end{bmatrix}$$

$$C = \frac{(x-\mu)^T (x-\mu)}{n}$$

$$C = \begin{bmatrix} 2.09 & 1.45 & -0.39 \\ 1.45 & 2.25 & -1.15 \\ -0.39 & -1.15 & 7.09 \end{bmatrix}$$

To find the eigen values of varience covarience matrix;

$|C - \lambda I| = 0$ is the characteristic equation

$$\begin{vmatrix} 2.09-\lambda & 1.45 & -0.39 \\ 1.45 & 2.25-\lambda & -1.15 \\ -0.39 & -1.15 & 7.09-\lambda \end{vmatrix} = 0$$

$\lambda^3 - \text{trace}(A).\lambda^2 + \text{trace}(|A|A^{-1})\lambda - |A| = 0$

↳ characteristic equation for 3x3 matrix.

$\lambda^3 - 11.43\lambda^2 + 31.886\lambda - 16.6284 = 0$

$\lambda_1 = 7.4465 \quad \lambda_2 = 3.3085 \quad \lambda_3 = +0.6749$

To find the eigen vector,

$$\begin{pmatrix} 2.09-7.4465 & 1.45 & -0.39 \\ 1.45 & 2.25-7.4465 & -1.15 \\ -0.39 & -1.15 & 7.09-\lambda \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = 0$$

$-5.3565x_1 + 1.45x_2 - 0.39x_3 = 0 \rightarrow (1)$

$1.45x_1 - 5.1965x_2 - 1.15x_3 = 0 \rightarrow (2)$

$-0.39x_1 - 1.15x_2 - 0.3572x_3 = 0 \rightarrow (3)$

$\frac{x_1}{}$ $\quad$ $\frac{x_2}{}$ $\quad$ $\frac{x_3}{}$ $\quad$ $\frac{x_1}{}$

$-5.3565 \quad 1.45 \quad -0.39 \quad -5.3565$

$1.45 \quad -5.1965 \quad -1.15 \quad 1.45$

$-0.39 \quad -1.15 \quad -0.3572 \quad -0.39$

$$\frac{x_1}{27.8335 - 2.1025} = \frac{x_2}{6.1599 + 0.5655} = \frac{x_3}{2.1025 - 27.8350}$$

Normalizing the value

$$= \frac{25.731}{\sqrt{662.0843 + 45.2310 + 662.161}}$$

$$= \frac{25.731}{} = -0.1375$$

$-0.1375, -0.2504, 0.9583$ is the eigen vector when $\lambda_1 = 7.4465$.

To find the eigen vector for $\lambda_2 = 3.3085$.

$$\begin{pmatrix} 2.09 - 3.3085 & 1.45 & -0.39 \\ 1.45 & 2.25 - 3.3085 & -1.15 \\ -0.39 & -1.15 & 7.09 - 3.3085 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = 0$$

$$-1.2185x_1 + 1.45x_2 - 0.39x_3 = 0 \rightarrow (1)$$
$$1.45x_1 - 1.0585x_2 - 1.15x_3 = 0 \rightarrow (2)$$
$$-0.39x_1 - 1.15x_2 + 3.7815x_3 = 0 \rightarrow (3)$$

x_1	x_2	x_3	x_1
-1.2185	1.45	-0.39	-1.2185
1.45	-1.0585	-1.15	1.45

$\Rightarrow \mp 0.6990 \quad \mp 0.6608 \quad \mp 0.2730$ are the eigen vector when $\lambda_2 = 3.3085$.

Q: Find the principal components of the following.

Screen Plot / Elbow Plot :-

There is always a question of how many components to retain. There is no definite answer to this question, values to consider include the amount of total sample variance explain, the relative sizes of the eigen values. A component with less eigen value deemed unimportant and may indicate an

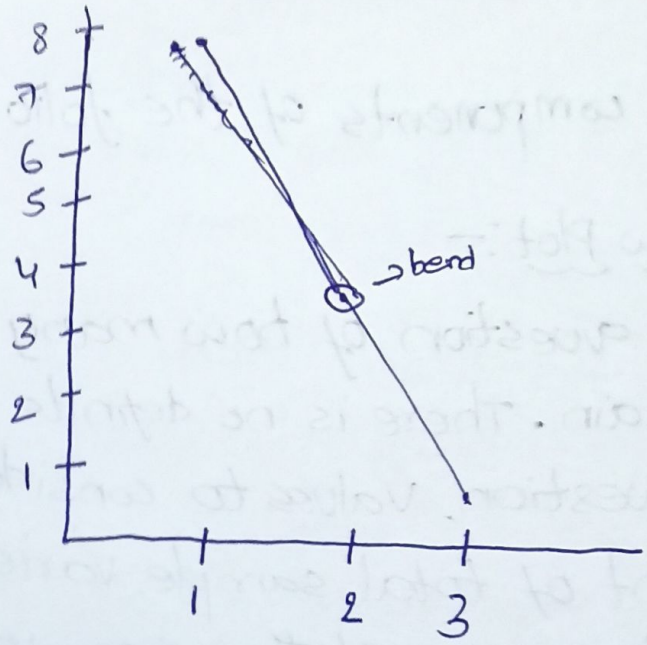
unsuspected linear dependency in the data.

An usual aid to determine an appropriate number of principal components is a scree plot/elbow plot with the eigen values ordered from largest to smallest, a scree plot is a plot of λ_i vs i^{th} magnitude of eigen value and its number. To determine the appropriate number of components we look for an elbow(bend) in the scree block. Plot.

Scree plot for Q3:-

The 3 eigen values are

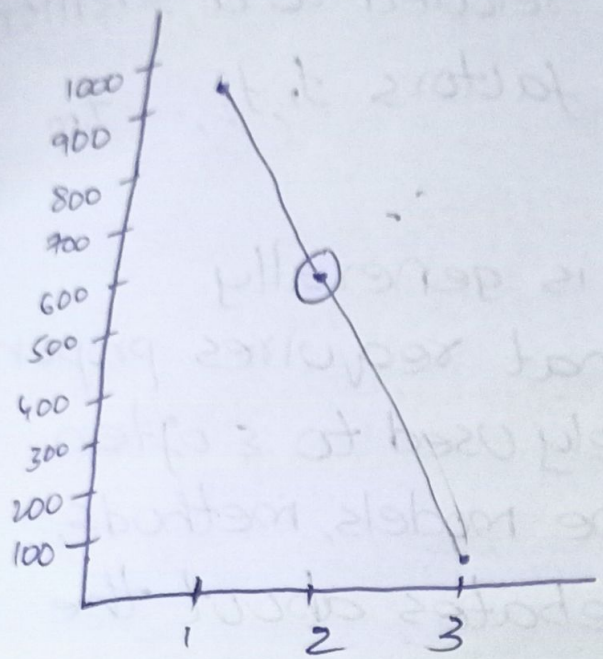
$\lambda_1 = 7.74465$, $\lambda_2 = 3.3085$, $\lambda_3 = 0.6749$



Scree plot for Q2:-

The 3 eigen values are

$\lambda_1 = 911.02$ $\lambda_2 = 688.13$ $\lambda_3 = 44.84$



Q4:- Find the principal components for the following data.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
16	16	13	18	16	15	14	16	16
18	19	15	16	18	18	18	17	19
17	17	14	14	17	17	20	14	15
17	17	17	16	18	18	16	20	14
16	15	17	17	18	18	19	16	19
15	17	16	17	18	18	15	19	16
17	16	16	18	18	18	17	15	18
20	18	16	20	15	15	19	14	17
14	16	18	17	19	19	18	17	18
16	16	15	19	18	18	18	15	14
18	19	16	14	14	14	17	16	13
19	15	15	18	16	16	18	19	17

Factor Analysis:- (Classification of Groups)

Factor analysis is a method for investing wheather a number of variables of interest $x_1, x_2, \ldots x_p$ are linearly related to a smaller number of unobservable factors $f_1, f_2, \ldots f_m$ $(m<p)$.

Factor Analysis is generally exploratory/ disuriptive that reqvires proper judgements. It is widely used to & often contraversial because the models, methods, are so flexible that debates about the interpretations that occur.

Notations for Factor Analysis:-

Let all the independent variables be denoted by

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} \qquad u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_p \end{pmatrix}$$

let the population mean vector is denoted by Here $E(x_i) = u_i$.

Consider m unobseevable common factors $f_1, f_2, \ldots f_m$ denoted by $f = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{pmatrix}$

Factor Analysis Modal:-

The factor modal can be taught of as a series of multiple regressions predicting each of the observable variables x_i from the values of unobseevable common factors f_i;

$$x_1 = \mu_1 + l_{11}f_1 + l_{12}f_2 + \cdots + l_{1m}f_m + \epsilon_1$$
$$x_2 = \mu_2 + l_{21}f_1 + l_{22}f_2 + \cdots + l_{2m}f_m + \epsilon_2$$
$$\vdots$$
$$x_p = \mu_p + l_{p1}f_p + l_{p2}f_2 + \cdots + l_{pm}f_m + \epsilon_p$$

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} + \begin{pmatrix} l_{11} & l_{12} & \cdots & l_{1m} \\ l_{21} & l_{22} & \cdots & l_{2m} \\ \vdots & \vdots & & \vdots \\ l_{p1} & l_{p2} & \cdots & l_{pm} \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_p \end{pmatrix}$$

$$X_{px1} = \mu_{px1} + L_{pxm} \cdot f_{mx1} + \epsilon_{px1}$$

$$\boxed{X - \mu = Lf - \epsilon} \rightarrow (1)$$

is known as factor modal.

- If the factors in the F matrix are dependent then it is known as oblique factor analysis.
- If the factors in the F matrix are independent then it is known as orthogonal factor analysis