

Multivariate analysis consists of a collection of methods that can be used when several measurements are made on each individual or object in one or more samples. Each sample is called an observation or unit.

Units                          Variables  
1. students                  several exam scores in a single course.

### THE VARIATE:-

The variate is a linear combination of variables with empirically determined weights. The building block of multivariate analysis is the variate. A variate of  $n$  weighted variables ( $x_1, x_2, \dots, x_n$ ) can be stated mathematically as

$$\text{Variate value} = w_1 x_1 + w_2 x_2 + \dots + w_n x_n$$

where  $x_n$  is the observed variable and  $w_n$  is the weight determined by the multivariate technique.

### Mean Vector:-

Mean vector consists of the mean of each variable

Let  $\bar{Y}$  represent a vector of  $p$  variables measured in a set of  $n$  sample units denoted by  $y_1, y_2, \dots, y_n$  where each  $y$  is a  $p$  valued vector. The mean vector is a vector of all the averages.

$$\bar{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} & \dots & y_{1j} & \dots & y_{1p} \\ y_{21} & y_{22} & \dots & y_{2j} & \dots & y_{2p} \\ \vdots & & & & & \\ y_{n1} & y_{n2} & \dots & y_{nj} & \dots & y_{np} \end{bmatrix}$$

(Variables)

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_p \end{bmatrix}$$

$$E(y) = E \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix} = \begin{bmatrix} E(y_1) \\ E(y_2) \\ \vdots \\ E(y_p) \end{bmatrix} = \begin{bmatrix} H_1 \\ H_2 \\ \vdots \\ H_p \end{bmatrix} = H$$

### Covariance Matrix:-

A matrix of sample covariances and variances.

$$S = (S_{jk}) = \begin{bmatrix} S_{11} & S_{12} & \dots & S_{1p} \\ S_{21} & S_{22} & \dots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \dots & S_{pp} \end{bmatrix}$$

covariance is a statistical tool that is used to determine the relationship between the movement of two asset pieces.

$$S_{jj} = S_{j^2} = \frac{1}{n-1} \sum_{i=1}^n (y_{ij} - \bar{y}_j)^2$$

$$= \frac{1}{n-1} \left( \sum_i y_{ij}^2 - n \bar{y}_j^2 \right)$$

$$S_{jk} = \frac{1}{n-1} \sum_{i=1}^n (y_{ij} - \bar{y}_j)(y_{ik} - \bar{y}_k)$$

$$= \frac{1}{n-1} \left( \sum_i y_{ij} y_{ik} - n \bar{y}_j \bar{y}_k \right)$$

### Problem:-

Three variables were measured (in milliequivalents per 100g) at 10 different locations in the south. The variables are  $y_1$  = available soil calcium,  $y_2$  = exchangeable soil calcium,  $y_3$  = leaching calcium

$$\bar{y} = (28.1, 7.18, 3.089)$$

Location Number	$y_1$	$y_2$	$y_3$
1	35	3.5	2.80
2	35	4.9	2.70
3	40	30.0	4.38
4	10	2.8	3.21
5	6	2.7	2.73
6	20	2.8	2.81
7	35	4.6	2.88
8	35	10.9	2.90
9	35	8.0	3.28
10	30	1.6	3.20

Solution:-

$$\text{Given } \bar{y}_1 = 28.1, \bar{y}_2 = 7.18, \bar{y}_3 = 3.089$$

$$S_{11} = \frac{1}{10-1} \left( \sum_{i=1}^{10} y_{i1}^2 - 10 \bar{y}_1^2 \right)$$

$$= \frac{1}{9} \left[ \sum_{i=1}^{10} y_{i1}^2 - 10 \bar{y}_1^2 \right]$$

$$= \cancel{\frac{1}{9}} \left[ (35)^2 - 10(28.1)^2 \right]$$

$$= \frac{1}{9} \left[ 1225 - 10(789.61) \right]$$

$$= \frac{1}{9} \left[ (35)^2 + (35)^2 + (40)^2 + \dots + (30)^2 - 10(789.61) \right]$$

$$= \frac{1}{9} [ 9161 - 7896.1 ] = 140.54$$

$$S_{12} = \frac{1}{10-9} \left[ \sum_{i=1}^{10} \cancel{y_{i1}} y_{i1} y_{i2} - 10 \bar{y}_1 \bar{y}_2 \right]$$

$$= \frac{1}{9} \left[ (35) \cancel{(3.5)} + (35) \cancel{(4.9)} + \dots + (30) \cancel{(3.2)} \right. \\ \left. - 10(28.1)(7.18) \right]$$

$$= \frac{1}{9} \left[ 2464.7 - 2017.58 \right] = 49.68$$

Continuing in this fashion, we obtain

$$S = \begin{bmatrix} 140.54 & 49.68 & 1.94 \\ 49.68 & 72.25 & 3.68 \\ 1.94 & 3.68 & 0.25 \end{bmatrix}$$

Correlation matrices :-

1 matrix of correlation between the variables

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj} s_{kk}}} = \frac{s_{jk}}{s_j s_k}$$

$$R = (r_{jk}) = \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \dots & 1 \end{bmatrix}$$

The diagonal elements of R are  $r_{jj} = 1$  and  $-1 \leq r_{jk} \leq 1$  for each  $j, k$ .

The total variation in a multivariate scatter is

Total variation =  $\text{tr}(S) = \text{trace}(S) = \text{sum of eigenvalues of } S$

= sum of diagonal elements of  $S$

The generalized variation =  $|S| = \text{product of eigenvalues of } S$

Problem:-

1) Two measurements  $x_1, x_2$  made at the same position on each of 3 cans of food, resulted in the following

$X$ -matrix  $X = \begin{bmatrix} 4 & 1 \\ -1 & 3 \\ 3 & 5 \end{bmatrix}$ , find the mean vector and

the covariance matrix  $S$ .

solution :-

$$X = \begin{bmatrix} 4 & 1 \\ -1 & 3 \\ 3 & 5 \end{bmatrix}$$

$$= [x_1, x_2, x_3]^T$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^3 x_i = \frac{1}{3} \left[ (4) + (-1) + (3) \right] = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

$$x' = X - \bar{x}$$

$$= \begin{bmatrix} 4 & 1 \\ -1 & 3 \\ 3 & 5 \end{bmatrix} - \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

$$= \begin{bmatrix} 2 & -2 \\ -3 & 0 \\ 1 & 2 \end{bmatrix}$$

$$S = \frac{1}{3} x'^T x'$$

$$= \begin{bmatrix} 14/3 & -2/3 \\ -2/3 & 8/3 \end{bmatrix} = \begin{bmatrix} 4.67 & -0.67 \\ -0.67 & 2.67 \end{bmatrix}$$

### Multiple Linear Regression:-

In most research problems where regression analysis is applied, more than one independent variable is needed in the regression model. The complexity of most scientific mechanisms is such that in order to be able to predict an important response, a multiple regression model is needed.

When this model is linear in the coefficients, it is called a multiple linear regression model.

$$\mu_{Y|x_1, x_2, \dots, x_k} = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

where  $x_1, x_2, \dots, x_k$  are  $k$  independent variables and the estimated response is obtained from the sample regression equation

$$\hat{y} = b_0 + b_1 x_1 + \dots + b_k x_k$$

where each regression coefficient  $\beta_i$  is estimated by  $b_i$  from the sample data using the method of least squares.

### Estimating the coefficients:-

In this section, we obtain the least squares estimates of the parameters  $\beta_0, \beta_1, \dots, \beta_k$  by fitting the multiple linear regression model.

$$\mu_{Y|x_1, x_2, \dots, x_k} = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

to the data points

$$\{(x_{1i}, x_{2i}, \dots, x_{ki}, y_i); i=1, 2, \dots, n\}$$

where  $y_i$  is the observed response to the values  $x_{1i}, x_{2i}, \dots, x_{ki}$  of the  $k$  independent variables  $x_1, x_2, \dots, x_k$ . Each observation  $(x_{1i}, x_{2i}, \dots, x_{ki}, y_i)$  is assumed to satisfy the following equation

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \epsilon_i$$

$$(or) \quad y_i = \hat{y}_i + \epsilon_i \quad \hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \dots + b_k x_{ki} + \epsilon_i$$

where  $\epsilon_i$  and  $e_i$  are the random error and residual respectively associated with the response  $y_i$  and fitted value  $\hat{y}_i$ .

In using the concept of least squares to arrive at estimates  $b_0, b_1, \dots, b_K$ , we minimize the equation

$$\begin{aligned} SSE &= \sum_{i=1}^n e_i^2 \\ &= \sum_{i=1}^n (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - \dots - b_K x_{Ki})^2 \end{aligned}$$

Differentiating SSE in terms with respect to  $b_0, b_1, \dots, b_K$  and equating to zero, we generate the set of  $K+1$  normal equations for multiple linear regression

$$nb_0 + b_1 \sum_{i=1}^n x_{1i} + b_2 \sum_{i=1}^n x_{2i} + \dots + b_K \sum_{i=1}^n x_{Ki} = \sum_{i=1}^n y_i$$

$$b_0 \sum_{i=1}^n x_{1i} + b_1 \sum_{i=1}^n x_{1i}^2 + b_2 \sum_{i=1}^n x_{1i} x_{2i} + \dots + b_K \sum_{i=1}^n x_{1i} x_{Ki} = \sum_{i=1}^n x_{1i} y_i$$

$$b_0 \sum_{i=1}^n x_{Ki} + b_1 \sum_{i=1}^n x_{Ki} x_{1i} + b_2 \sum_{i=1}^n x_{Ki} x_{2i} + \dots + b_K \sum_{i=1}^n x_{Ki}^2 = \sum_{i=1}^n x_{Ki} y_i$$

These equations can be solved for  $b_0, b_1, \dots, b_K$  by any appropriate method for solving system of linear equations.

Problem:-

1) A study was done on a diesel-powered light-duty pickup truck to see if humidity, air temperature and barometric pressure influence emission of nitrous oxide (in ppm). Emission measurements were taken at different times, with varying experimental conditions. The data are given in the following table. The model is

$$Y_i/x_1, x_2, x_3 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

(or) equivalently,

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i \quad i=1, 2, \dots, 20$$

Fit this multiple linear regression model to the given data and then estimate the amount of nitrous oxide emitted for the conditions where humidity is 50%, temperature is 76°F and barometric pressure is 29.30.

Nitrous Oxide (y)	Humidity (x <sub>1</sub> )	Temperature (x <sub>2</sub> )	Pressure (x <sub>3</sub> )
0.90	72.4	76.3	29.18
0.91	41.6	70.3	29.35
0.96	34.3	77.1	29.24
0.89	35.1	68.0	29.27
1.00	10.7	79.0	29.78
1.10	12.9	67.4	29.39
1.15	8.3	66.8	29.69
1.03	20.1	76.9	29.48
0.77	72.2	77.7	29.09
1.07	24.0	67.7	29.60
1.07	23.2	76.8	29.38
0.94	47.4	86.6	29.35
1.10	31.5	76.9	29.63
1.10	10.6	86.3	29.56
1.10	11.2	86.0	29.48
0.91	73.3	76.3	29.40

0.87	75.4	77.9	29.28
0.78	96.6	78.7	29.29
0.82	107.4	86.8	29.03
0.95	54.9	70.9	29.87

Solution

$$nb_0 + b_1 \sum_{i=1}^n x_{1i} + b_2 \sum_{i=1}^n x_{2i} + b_3 \sum_{i=1}^n x_{3i} = \sum_{i=1}^n y_i$$

$$b_0 \sum_{i=1}^n x_{1i} + b_1 \sum_{i=1}^n x_{1i}^2 + b_2 \sum_{i=1}^n x_{1i} x_{2i} + b_3 \sum_{i=1}^n x_{1i} x_{3i} = \sum_{i=1}^n x_{1i} y_i$$

$$b_0 \sum_{i=1}^n x_{2i} + b_1 \sum_{i=1}^n x_{2i} x_{1i} + b_2 \sum_{i=1}^n x_{2i}^2 + b_3 \sum_{i=1}^n x_{2i} x_{3i} = \sum_{i=1}^n x_{2i} y_i$$

$$b_0 \sum_{i=1}^n x_{3i} + b_1 \sum_{i=1}^n x_{3i} x_{1i} + b_2 \sum_{i=1}^n x_{3i} x_{2i} + b_3 \sum_{i=1}^n x_{3i}^2 = \sum_{i=1}^n x_{3i} y_i$$

$$\Rightarrow \cancel{20b_0} + \sum_{i=1}^{20} x_{1i} = 72.4 + 41.6 + \dots + 54.9 = 863.100$$

$$\sum_{i=1}^{20} x_{2i} = 76.3 + 70.3 + \dots + 70.9 = 1630.4$$

$$\sum_{i=1}^{20} x_{3i} = 29.18 + 29.35 + \dots + 29.37 = 587.840$$

$$\sum_{i=1}^{20} x_{1i}^2 = (72.4)^2 + (41.6)^2 + (34.3)^2 + \dots + (54.9)^2 = 54876.89$$

$$\sum_{i=1}^{20} x_{1i} x_{2i} = (72.4)(76.3) + (41.6)(70.3) + \dots + (54.9)(70.9) \\ = 67000.09$$

$$\sum_{i=1}^{20} x_{1i} x_{3i} = (72.4)(29.18) + (41.6)(29.35) + \dots + (54.9)(29.37) \\ = 25283.395$$

$$\sum_{i=1}^{20} x_{2i}^2 = (76.3)^2 + (70.3)^2 + \dots + (70.9)^2 = 117912.32$$

$$\sum_{i=1}^{20} x_{2i} x_{3i} = (76.3)(29.18) + (70.3)(29.35) + \dots + (70.9)(29.37) \\ = 44976.867$$

$$\sum_{i=1}^{20} x_{3i}^2 = 29.18^2 + 29.35^2 + 29.24^2 + \dots + 29.87^2 = 17278.509$$

$$\sum_{i=1}^{20} y_i = 0.90 + 0.91 + 0.96 + \dots + 0.95 = 19.42$$

$$\sum_{i=1}^{20} x_{ii} y_i = (12.4)(0.9) + (41.6)(0.91) + \dots + (54.9)(0.95) \\ = 779.477$$

$$\sum_{i=1}^n x_i y_i = (76.3)(0.9) + (70.3)(0.91) + \dots + (70.9)(0.95) \\ = 1483.437$$

$$\sum_{i=1}^6 x_i y_i = (29 \cdot 18)(0.9) + (29 \cdot 35)(0.91) + \dots + (29 \cdot 37)(0.95) \\ = 571.122$$

$$\Rightarrow 20b_0 + 863.1b_1 + 1530.4b_2 + \cancel{558.55}b_3 = 19.42$$

$$863.1 b_0 + 54876.89 b_1 + 67000.09 b_2 + 25283.895 b_3 = 779.477$$

$$1530.4b_0 + 67000.09b_1 + 117912.32b_2 + 44976.867b_3 = 1483.437$$

$$587.84 - 558.56 b_0 + 25283.395 b_1 + 44976.867 b_2 + 14278.509 b_3$$

By solving these equations, we get

$$b_0 = -3.5115$$

$$b_1 = -0.0026244$$

$$b_2 = 0.00079895$$

$$b_3 = 0.15428$$

Therefore, the egression equation is

$$\hat{y} = -3.5115 - 0.0026244 x_1 + 0.00079895 x_2 + 0.15428 x_3$$

For 50% humidity, a temperature of  $76^{\circ}\text{F}$ , and a barometric pressure of 29.3, the estimated amount of nitrous oxide emitted is

$$\begin{aligned} \hat{y} &= -3.5115 - 0.0026244x_1 + 0.00049895x_2 + 0.15428x_3 \\ &= -3.5115 - 0.0026244(50) + 0.00049895(76) + 0.15428(29.3) \\ &\approx 0.938 \text{ ppm.} \end{aligned}$$

### Polynomial Regression:-

Polynomial regression is a special case of linear regression where we fit a polynomial equation on the data with a curvilinear relationship between the target variable and the independent variables.

Now suppose that we wish to fit the polynomial equation

$$y_{\text{fit}} = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_r x^r$$

to the  $n$  pairs of observations  $\{(x_i, y_i); i = 1, 2, \dots, n\}$ . Each observation,  $y_i$ , satisfies the equation

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \dots + \beta_r x_i^r + \epsilon_i$$

$$(1) \quad y_i = \hat{y}_i + \epsilon_i = b_0 + b_1 x_i + b_2 x_i^2 + \dots + b_r x_i^r + \epsilon_i$$

where  $r$  is the degree of the polynomial and  $\epsilon_i$  and  $\hat{y}_i$  are again the random error and residual associated with the response  $y_i$  and fitted value  $\hat{y}_i$  respectively.

### Problem:-

- 1) Given the data

x	0	1	2	3	4	5	6	7	8	9
y	9.1	7.3	3.2	4.6	4.8	2.9	5.7	7.1	8.8	10.2

fit a regression curve of the form  $Uy/x = \beta_0 + \beta_1 x + \beta_2 x^2$   
and then estimate  $Uy_{12}$ .

Solution:

$$x \cdot y$$

Estimating the regression coefficients is based on the solution of the following system

$$\begin{bmatrix} \sum_{i=1}^N x_i & \sum_{i=1}^N x_i^2 & \sum_{i=1}^N x_i^3 \\ \sum_{i=1}^N x_i^2 & \sum_{i=1}^N x_i^4 & \sum_{i=1}^N x_i^5 \\ \sum_{i=1}^N x_i^3 & \sum_{i=1}^N x_i^5 & \sum_{i=1}^N x_i^6 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^N y_i \\ \sum_{i=1}^N xy_i \\ \sum_{i=1}^N x^2 y_i \end{bmatrix}$$

Problems:-

1) Given the data

x	0	1	2	3	4	5	6	7	8	9
y	9.1	7.3	3.2	4.6	4.8	2.9	5.7	7.1	8.8	10.2

fit a regression curve of the form  $Uy/x = \beta_0 + \beta_1 x + \beta_2 x^2$   
and then estimate  $Uy_{12}$ .

Solution:

$$\begin{bmatrix} \sum_{i=1}^N x_i & \sum_{i=1}^N x_i^2 & \sum_{i=1}^N x_i^3 \\ \sum_{i=1}^N x_i^2 & \sum_{i=1}^N x_i^4 & \sum_{i=1}^N x_i^5 \\ \sum_{i=1}^N x_i^3 & \sum_{i=1}^N x_i^5 & \sum_{i=1}^N x_i^6 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^N y_i \\ \sum_{i=1}^N xy_i \\ \sum_{i=1}^N x^2 y_i \end{bmatrix}$$

$x$	$y$	$x^2$	$x^3$	$x^4$	$xy$	$x^2y$
0	9.1	0	0	0	0	0
1	-1.3	1	1	1	-1.3	-1.3
2	3.2	4	8	16	6.4	12.8
3	4.6	9	27	81	13.8	41.4
4	4.8	16	64	256	19.2	76.8
5	2.9	25	125	625	14.5	72.5
6	5.7	36	216	1296	34.2	205.2
7	7.1	49	343	2401	49.7	347.9
8	8.8	64	512	4096	70.4	563.2
9	10.2	81	729	6561	91.8	826.2
<b>45</b>	<b>63.7</b>	<b>285</b>	<b>2025</b>	<b>15333</b>	<b>307.3</b>	<b>2153.3</b>

$$\begin{bmatrix} 10 & 45 & 285 \\ 45 & 285 & 2025 \\ 285 & 2025 & 15333 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 63.7 \\ 307.3 \\ 2153.3 \end{bmatrix}$$

$$10\beta_0 + 45\beta_1 + 285\beta_2 = 63.7$$

$$45\beta_0 + 285\beta_1 + 2025\beta_2 = 307.3$$

$$285\beta_0 + 2025\beta_1 + 15333\beta_2 = 2153.3$$

By solving the above system of equations.

$$\beta_0 = 8.698, \beta_1 = -2.341, \beta_2 = 0.288$$

∴ The regression equation is

$$\hat{y} = 8.698 - 2.341x + 0.288x^2$$

when  $x=2$ , our estimate of  $y_{12}$  is

$$\hat{y} = 8.698 - 2.341(2) + 0.288(4) = 5.168$$

2)  $X: 1 \ 3 \ 4 \ 7 \ 9$

$\times Y: 1 \ 6 \ 1 \ 8 \ 20$

Find the regression equation for the following data

Let us consider the equation  $y = b_0 + b_1 x$ .

Solution:

$x$	$y$	$x^2$	$x^2 y$	$xy$
1	1	1	1	1
3	6	9	54	18
4	1	16	16	4
7	8	49	392	56
9	20	81	1620	180
24	36	156	2083	259

$$\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

$$\begin{bmatrix} 5 & 24 \\ 24 & 156 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} 36 \\ 259 \end{bmatrix}$$

$$5b_0 + 24b_1 = 36$$

$$24b_0 + 156b_1 = 259$$

we get the values  $b_0 = -2.92$ ,  $b_1 = 2.11$  by solving the equations.

$\therefore \hat{y} = -2.92 + 2.11x$  is a linear square fit

3) Develop a 2nd order polynomial curve fit for the following data set:

$$\begin{matrix} x & -3 & -2 & -1 & -0.2 & 1 & 3 \\ y & 0.9 & 0.8 & 0.4 & 0.2 & 0.1 & 0 \end{matrix}$$

solution :-

The second order polynomial curve equation is

$$y = a_0 + a_1 x + a_2 x^2$$

$$\begin{bmatrix} 6 & -2.2 & 24.08 \\ -2.2 & 24.04 & -8.008 \\ 24.04 & -8.008 & 180.0016 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 2.4 \\ -4.64 \\ 11.808 \end{bmatrix}$$

Using Cramer's rule,

$$M_0 = \begin{bmatrix} 2.4 & -2.2 & 24.08 \\ -4.64 & 24.04 & -8.008 \\ 11.808 & -8.008 & 180.0016 \end{bmatrix}$$

$$M_1 = \begin{bmatrix} 6 & 2.4 & 24.08 \\ -2.2 & -4.64 & -8.008 \\ 24.04 & 11.808 & 180.0016 \end{bmatrix}$$

$$M_2 = \begin{bmatrix} 6 & -2.2 & 2.4 \\ -2.2 & 24.04 & -4.64 \\ 24.04 & -8.008 & 11.808 \end{bmatrix}$$

$$a_0 = \frac{\det M_0}{\det M} = \frac{2671.20}{11661.27} = 0.2291$$

$$a_1 = \frac{\det M_1}{\det M} = \frac{-1898.46}{11661.27} = -0.1628$$

$$a_2 = \frac{\det M_2}{\det M} = \frac{323.76}{11661.27} = 0.0278$$

Hence the regression equation is

$$Y = 0.2291 - 0.1628 X + 0.0278 X^2$$

### LINEAR REGRESSION MODEL USING MATRICES:-

In fitting a multiple linear regression model, particularly when the number of variables exceed two, a knowledge of matrix theory can facilitate the mathematical manipulations considerably. Suppose that the experimenter has  $k$  independent variables  $x_1, x_2, \dots, x_k$  and  $n$  observations  $y_1, y_2, \dots, y_n$  each of which can be expressed by the equation.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \epsilon_i$$

This model essentially represents  $n$  equations describing how the response values are generated in the scientific process. Using matrix notation, we can write the following equation.

$$y = X\beta + \epsilon,$$

where,

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

$$\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Then the least squares method of estimation of  $\beta$ , involves finding  $b$  for which

$$SSE = (y - xb)'(y - xb)$$

is minimized. This minimization process involves solving for  $b$  in the equation

$$\frac{\partial}{\partial b} (SSE) = 0.$$

The result reduces to the solution of  $b$  in

$$(x'x)b = x'y$$

Notice the nature of the  $x$  matrix. Apart from the initial element, the  $i^{\text{th}}$  row represents the  $x$  values that give rise to the response  $y_i$ . Writing

$$A = x'x = \begin{bmatrix} n & \sum_{i=1}^n x_{1i} & \sum_{i=1}^n x_{2i} & \cdots & \sum_{i=1}^n x_{ki} \\ \sum_{i=1}^n x_{1i} & \sum_{i=1}^n x_{1i}^2 & \sum_{i=1}^n x_{1i}x_{2i} & \cdots & \sum_{i=1}^n x_{1i}x_{ki} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{ki} & \sum_{i=1}^n x_{ki}x_{1i} & \sum_{i=1}^n x_{ki}x_{2i} & \cdots & \sum_{i=1}^n x_{ki}^2 \end{bmatrix}$$

and

$$g = x'y = \begin{bmatrix} g_0 = \sum_{i=1}^n y_i \\ g_1 = \sum_{i=1}^n x_{1i}y_i \\ \vdots \\ g_k = \sum_{i=1}^n x_{ki}y_i \end{bmatrix}$$

allows the normal equations to be put in the matrix form

$$Ab = g$$

If the matrix A is nonsingular, we can write the solution for the regression coefficient as

$$\hat{b} = A^{-1}g = (x'x)^{-1}x'y.$$

Problems:-

- 1) The percent survival rate of sperm in a certain type of animal semen, after storage, was measured at various combinations of concentrations of three materials used to increase chance of survival. Estimate the multiple linear regression model to the given data.

$y$ (% survival)	$x_1$ (weight %)	$x_2$ (weight %)	$x_3$ (weight %)
25.5	1.74	5.3	10.80
31.2	6.32	5.42	9.40
25.9	6.22	8.41	7.2
38.4	10.52	4.63	8.50
18.4	1.19	11.60	9.40
26.7	1.22	5.85	9.90
26.4	4.10	6.62	8
25.9	6.32	8.72	9.1
32	4.08	4.42	8.7
25.2	4.15	7.60	9.20
39.7	10.15	4.83	9.40
35.7	1.72	3.12	7.60
26.5	1.70	5.30	8.20

Solution:-

The least squares estimating equations,  $n=13$

$$(X'X)b = X'y$$

$$\sum_{i=1}^{13} x_{1i} = 59.43 \quad \sum_{i=1}^{13} x_{2i} = 81.82 \quad \sum_{i=1}^{13} x_{3i} = 115.40$$

$$\sum_{i=1}^{13} x_{1i}^2 = 394.7255 \quad \sum_{i=1}^{13} x_{2i}^2 = 576.7264 \quad \sum_{i=1}^{13} x_{3i}^2 = 1035.9600$$

$$\sum_{i=1}^{13} x_{1i}x_{2i} = 360.6621 \quad \sum_{i=1}^{13} x_{2i}x_{3i} = 728.3100 \quad \sum_{i=1}^{13} x_{3i}x_{1i} = 522.0780$$

$$\sum_{i=1}^{13} y_i = 377.5 \quad \sum_{i=1}^{13} x_{1i}y_i = 1877.567 \quad \sum_{i=1}^{13} x_{2i}y_i = 2246.661$$

$$\sum_{i=1}^{13} x_{3i}y_i = 3337.780$$

$$\begin{bmatrix} n \\ \sum_{i=1}^n x_{1i} \\ \sum_{i=1}^n x_{2i} \\ \sum_{i=1}^n x_{3i} \\ \vdots \\ \sum_{i=1}^n x_{4i} \end{bmatrix} \begin{bmatrix} \sum_{i=1}^n x_{ii} \\ \sum_{i=1}^n x_{1i}^2 \\ \sum_{i=1}^n x_{2i}x_{1i} \\ \sum_{i=1}^n x_{3i}x_{1i} \\ \vdots \\ \sum_{i=1}^n x_{3i}x_{2i} \end{bmatrix} \begin{bmatrix} \sum_{i=1}^n x_{3i} \\ \sum_{i=1}^n x_{1i}x_{3i} \\ \sum_{i=1}^n x_{2i}x_{3i} \\ \sum_{i=1}^n x_{3i}x_{2i} \\ \vdots \\ \sum_{i=1}^n x_{3i}^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{1i}y_i \\ \sum_{i=1}^n x_{2i}y_i \\ \sum_{i=1}^n x_{3i}y_i \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} 13 & 59.43 & 81.82 & 115.4 \\ 59.43 & 394.7255 & 360.6621 & 522.078 \\ 81.82 & 360.6621 & 576.7264 & 728.31 \\ 115.4 & 522.078 & 728.31 & 1035.96 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 377.5 \\ 1877.567 \\ 2246.661 \\ 3337.780 \end{bmatrix}$$

$$(x'x)^{-1} = \begin{bmatrix} 8.0648 & -0.0826 & -0.0942 & -0.7905 \\ -0.0826 & 0.0085 & 0.0017 & 0.0035 \\ -0.0942 & 0.0017 & 0.0166 & -0.0021 \\ -0.7905 & 0.0035 & -0.0021 & 0.0886 \end{bmatrix}$$

and then, using the relation  $b = (x'x)^{-1} x'y$ ,

$$\begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 8.0648 & -0.0826 & -0.0942 & -0.7905 \\ -0.0826 & 0.0085 & 0.0017 & 0.0035 \\ -0.0942 & 0.0017 & 0.0166 & -0.0021 \\ -0.7905 & 0.0035 & -0.0021 & 0.0886 \end{bmatrix}$$

$$\begin{bmatrix} 371.5 \\ 1877.567 \\ 2246.66 \\ 3337.780 \end{bmatrix}$$

$$= \begin{bmatrix} 39.1574 \\ 1.0161 \\ -1.8616 \\ -0.8433 \end{bmatrix} \begin{bmatrix} 39.224 \\ 0.944 \\ -2.083 \\ -0.851 \end{bmatrix}$$

Hence, our estimated regression equation is.

$$\hat{y} = 39.1574 + 1.0161 x_1 - 1.8616 x_2 - 0.8433 x_3.$$

- 8) Determine the least squares regression line using a matrix. The price is  $x$  and  $y$  is the monthly sales.

Price (x)	Demand (y)
49	124
69	95
89	71
99	45
109	18

solution:-

The matrix equation is  $f(x) = Ax + E$

$$\text{where } x = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

From the given data

$$x = \begin{bmatrix} 1 & 49 \\ 1 & 69 \\ 1 & 89 \\ 1 & 99 \\ 1 & 109 \end{bmatrix} \quad y = \begin{bmatrix} 124 \\ 95 \\ 71 \\ 45 \\ 18 \end{bmatrix}$$

$$\theta = (x^T x)^{-1} x^T y$$

$$x^T = \begin{bmatrix} 1 & 1 & 1 & -1 & 1 \\ 49 & 69 & 89 & 99 & 109 \end{bmatrix}$$

$$x^T x = \begin{bmatrix} 1 & 1 & 1 & 1 & -1 \\ 49 & 69 & 89 & 99 & 109 \end{bmatrix} \begin{bmatrix} 1 & 49 \\ 1 & 69 \\ 1 & 89 \\ 1 & 99 \\ 1 & 109 \end{bmatrix}$$

$$x^T x = \begin{bmatrix} 5 & 415 \\ 415 & 36765 \end{bmatrix}$$

$$(x^T x)^{-1} = \frac{1}{11600} \begin{bmatrix} 36765 & -415 \\ -415 & 5 \end{bmatrix}$$

$$x^T y = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 49 & 69 & 89 & 99 & 100 \end{bmatrix} \begin{bmatrix} 124 \\ 95 \\ 71 \\ 45 \\ 18 \end{bmatrix}$$

$$= \begin{bmatrix} 353 \\ 25365 \end{bmatrix}$$

$$A = (x^T x)^{-1} x^T y$$

$$= \frac{1}{11600} \begin{bmatrix} 36765 & -415 \\ -415 & 5 \end{bmatrix} \begin{bmatrix} 353 \\ 25365 \end{bmatrix}$$

$$= \begin{bmatrix} 211 \\ -1.7 \end{bmatrix}$$

$$\therefore f(x) = 211 - 1.7x$$

Now let us find the sum of squares of the error.

x	y	$f(x_i)$	$e_i = y_i - f(x_i)$
49	124	127.7	-3.7
69	95	93.7	1.3
89	71	59.7	11.3
99	45	42.7	2.3
100	18	25.7	-7.7

$$E = \begin{bmatrix} -3.7 \\ 1.3 \\ 11.3 \\ 2.3 \\ -7.7 \end{bmatrix}$$

$$SSE = E^T E = \begin{bmatrix} -3.7 & 1.3 & 11.3 & 2.3 & -7.7 \end{bmatrix} \begin{bmatrix} -3.7 \\ 1.3 \\ 11.3 \\ 2.3 \\ -7.7 \end{bmatrix} = 207.66$$

### Properties of the Least Squares Estimates:

The means and variances of the estimates  $b_0, b_1, \dots, b_k$  are obtained under certain assumptions on the random errors  $e_1, e_2, \dots, e_K$  that are identical to those made in the case of simple linear regression.

When we assume these errors to be independent, each with mean 0 and variance  $\sigma^2$ , it can be shown that  $b_0, b_1, \dots, b_k$  are respectively, unbiased estimators of the regression coefficients  $\beta_0, \beta_1, \dots, \beta_k$ .

In addition, the variances of the  $b$ 's are obtained through the elements of the inverse of the  $A$  matrix. Note that the off-diagonal elements of  $A = x'x$  represent sum of products of elements in the columns of  $x$ , while the diagonal elements of  $A$  represent sum of squares of elements in the columns of  $x$ .

The inverse matrix,  $A^{-1}$ , apart from the multiples of, represents the variance-covariance matrix of the estimated regression coefficients. That is, the elements of the matrix  $A^{-1}\sigma^2$  display the variances of  $b_0, b_1, \dots, b_k$  on the main diagonal and covariances on the off-diagonal.

For example, in a  $K=2$  multiple linear regression problem,

$$(x'x)^{-1} = \begin{bmatrix} c_{00} & c_{01} & c_{02} \\ c_{10} & c_{11} & c_{12} \\ c_{20} & c_{21} & c_{22} \end{bmatrix}$$

with the elements below the main diagonal determined through the symmetry of the matrix. Then we can write

$$\sigma_{bi}^2 = c_{ii} \sigma^2, \quad i=0, 1, 2.$$

$$\begin{aligned}\sigma_{bibj} &= \text{cov}(b_i, b_j) \\ &= c_{ij} \sigma^2, \quad i \neq j\end{aligned}$$

The estimates of the variances and hence the standard errors of these estimates are obtained by replacing  $\sigma^2$  with the appropriate estimate obtained through experimental data.

Theorem:

For the linear regression equation  $y = X\beta + \epsilon$ , an unbiased estimate of  $\sigma^2$  is given by the error or residual mean square.

$$s^2 = \frac{\text{SSE}}{n-k-1}, \text{ where SSE} = \sum_{i=1}^n e_i^2$$

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

The error and regression sum of squares take on the same form and play the same role. ~~as in the~~

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

continues to hold, and we retain our previous notation, namely

$$\text{SST} = \text{SSR} + \text{SSE}$$

with  $SST = \sum_{i=1}^n (y_i - \bar{y})^2$  = total sum of squares

and

$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$  = regression sum of squares.

There are  $k$  degrees of freedom associated with  $SSR$ , and as always,  $SST$  has  $n-1$  degrees of freedom. Therefore, after subtraction,  $SSE$  has  $n-k-1$  degrees of freedom.

Analysis of Variance in Multiple Regression:-

The partition of the total sum of squares into its components, the regression and error sums of squares, plays an important role. An analysis of variance can be conducted to shed light on the quality of the regression equation. A useful hypothesis that determines if a significant amount of variation is explained by the model is

$$H_0: \beta_1 = \beta_2 = \beta_3 = \dots = \beta_k = 0.$$

The analysis of variance involves an F-test via a table given as below follows:

Source	Sum of Squares	Degrees of Freedom	Mean Square	F
Regression	SSR	$K$	$MSR = \frac{SSR}{K}$	$f = \frac{MSR}{MSE}$
Error	SSE	$n-K-1$	$MSE = \frac{SSE}{n-K-1}$	
Total	SST	$n-1$		

a conclusion or opinion that is formed because of known  
Inferences in Multiple Linear Regression:- facts

A knowledge of the distributions of the individual coefficient estimators enables the experimenter to construct confidence intervals for the coefficients and to test hypotheses about them.

$$t = \frac{b_j - \beta_{j0}}{s\sqrt{c_{jj}}}$$

The  $b_j$  ( $j=0, 1, 2, \dots, k$ ) are normally distributed with mean  $\beta_j$  and variance  $c_{jj}\sigma^2$  with  $n-k-1$  degrees of freedom to test hypotheses and construct confidence intervals on  $\beta_j$ .

For example, if we wish to test

$$H_0 : \beta_j = \beta_{j0}$$

$$H_1 : \beta_j \neq \beta_{j0},$$

we compute the above t-statistic and do not reject  $H_0$  if  $-t_{\alpha/2} < t < t_{\alpha/2}$ , where  $t_{\alpha/2}$  has  $n-k-1$  degrees of freedom.

Problem:-

i) for the model of problem 1. (survival rate of sperm), test the hypothesis that  $\beta_2 = -2.5$  at the 0.05 level of significance against the alternative that  $\beta_2 > -2.5$ .

Solution:-

First we have to find the multiple linear regression equation.

$$\hat{y} = 39.1574 + 1.0161 x_1 - 1.8616 x_2 - 0.3433 x_3$$

$y$	$x_1$	$x_2$	$x_3$	$\hat{y}_i$	$e_i = y_i - \hat{y}_i$
25.5	1.74	5.30	10.80	27.3513	-1.8513
31.2	6.32	5.42	9.4	32.2623	-1.0623
25.9	6.22	8.41	7.2	27.3497	-1.4497
38.4	10.52	4.63	8.5	38.3095	0.0905
18.4	1.19	11.60	9.4	15.5450	2.855
26.7	1.22	5.85	9.9	26.1080	0.592
26.4	4.1	6.62	8	28.2532	-1.8532
25.9	6.32	8.72	9.1	26.2220	-0.322
32	4.08	4.42	8.7	32.0881	-0.881
25.2	4.15	7.6	9.2	26.0677	2.4477
39.7	10.15	4.83	9.4	37.2523	3.2122
35.7	1.72	3.12	7.6	32.4878	-1.7032
26.5	1.7	5.3	8.2	28.2032	

$$E = \begin{bmatrix} -1.8513 \\ -1.0623 \\ \vdots \\ -1.7032 \end{bmatrix}$$

$$E^T E = \begin{bmatrix} -1.8513 & -1.0623 & -1.4497 & \dots & -1.7032 \end{bmatrix} \begin{bmatrix} -1.8513 \\ -1.0623 \\ \vdots \\ -1.7032 \end{bmatrix}$$

$$= 38.6761$$

$$SSE = 38.6761$$

$$s^2 = \frac{SSE}{n-k-1} = \frac{38.6761}{13-3-1} = 4.2973$$

$$s = 2.073$$

$$H_0 : \beta_2 = -2.5$$

$$H_1 : \beta_2 > -2.5$$

$$t = \frac{b_2 - \beta_{20}}{s \sqrt{c_{22}}} = \frac{-1.8616 + 2.5}{2.073 \sqrt{c_{22}}}$$

To find  $c_{22}$ :

$$A = (x^T x)^{-1} x^T y$$

we have to find  $(x^T x)^{-1}$

$$x^T = \begin{bmatrix} 1 & 1.74 & 5.30 & 10.8 \\ 1 & 6.32 & 5.42 & 9.4 \\ 1 & 6.22 & 8.21 & 7.2 \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1.70 & 5.30 & 8.2 \end{bmatrix} \quad y = \begin{bmatrix} 25.5 \\ 31.2 \\ 25.9 \\ \vdots \\ \vdots \\ 26.5 \end{bmatrix}$$

$$x^T x = \begin{bmatrix} 13 & 59.43 & 81.82 & 115.40 \\ 59.43 & 394.7255 & 360.6621 & 522.0780 \\ 81.82 & 360.6621 & 576.7264 & 728.3100 \\ 115.40 & 522.0780 & 728.81 & 1085.9600 \end{bmatrix}$$

$$(x^T x)^{-1} = \begin{bmatrix} c_{00} & c_{01} & c_{02} & c_{03} \\ c_{01} & -0.0826 & -0.0942 & -0.7905 \\ c_{02} & -0.0942 & 0.0017 & 0.0037 \\ c_{03} & -0.7905 & 0.0087 & 0.0886 \end{bmatrix}$$

$$t = -\frac{1.8616 + 2.5}{2.073 \sqrt{0.0166}} = 2.390$$

degrees of freedom =  $n-1 = 13-1=12$ ,

$$t_{\alpha} = 0.04$$

$$\therefore t > t_{\alpha}$$

$\therefore H_0$  is rejected. Hence  $\beta_2 > -2.5$ .

### Inferences on Mean Response and Prediction:-

One of the most useful inferences that can be made regarding the quality of the predicted response  $y_0$  corresponding to the values  $x_{10}, x_{20}, \dots, x_{k0}$  is the confidence interval on the mean response

confidence interval on the mean response for the set of conditions  $H_4 | x_{10}, x_{20}, \dots, x_{k0}$ . By constructing a confidence interval on the mean response for the set of conditions given by

$$x'_0 = [1, x_{10}, x_{20}, \dots, x_{k0}]$$

We augment the conditions on the  $x$ 's by the number 1 in order to facilitate the matrix notation. Normality in the  $\epsilon_i$  produces normality in the  $b_i$  and the mean and variance are still the same. The covariance between  $b_i$  and  $b_j$  for  $i \neq j$ . Hence.

$$\hat{y} = b_0 + \sum_{j=1}^k b_j x_{j0}$$

is likewise normally distributed and is, in fact, an unbiased estimator for the mean response on which we are attempting to attach a confidence interval. The

variance of  $\hat{y}_0$  written in matrix notation simply as a function of  $\sigma^2$ ,  $(x'x)^{-1}$ , and the condition vector  $x_0'$  is

$$\sigma_{\hat{y}_0}^2 = \sigma^2 x_0' (x'x)^{-1} x_0.$$

After  $\sigma^2$  is replaced by  $s^2$  as given by theorem, the  $100(1-\alpha)\%$  confidence interval on  $\mu_{y|x_0, x_2, \dots, x_k}$  can be constructed from the statistic

$$T = \frac{\hat{y}_0 - \mu_{y|x_0, x_2, \dots, x_k}}{s \sqrt{x_0' (x'x)^{-1} x_0}}$$

which has a t-distribution with  $n-k-1$  degrees of freedom.

confidence Interval for  $\mu_{y|x_0, x_2, \dots, x_k}$  :-

a  $100(1-\alpha)\%$  confidence interval for the mean response  $\mu_{y|x_0, x_2, \dots, x_k}$  is

$$\hat{y}_0 - t_{\alpha/2} s \sqrt{x_0' (x'x)^{-1} x_0} < \mu_{y|x_0, x_2, \dots, x_k}$$

$$< \hat{y}_0 + t_{\alpha/2} s \sqrt{x_0' (x'x)^{-1} x_0}.$$

where  $t_{\alpha/2}$  is a value of the t-distribution with  $n-k-1$  degrees of freedom.

The quantity  $s \sqrt{x_0' (x'x)^{-1} x_0}$  is often called the standard error of prediction. ~~and appears on the~~

Problem:-

- 2) Using the data of previous problem, construct a 95% confidence interval for the mean response when  $x_1 = 3$ ,  $x_2 = 8$ , and  $x_3 = 9$ .

Solution:-

From the regression equation

$$\hat{y} = 39.1574 + 1.0161 x_1 - 1.8616 x_2 - 0.3433 x_3$$

the estimated percent survival when  $x_1 = 3$ ,  $x_2 = 8$  and  $x_3 = 9$  is

$$\begin{aligned}\hat{y} &= 39.1574 + 1.0161(3) - 1.8616(8) - 0.3433(9) \\ &= 24.2282\end{aligned}$$

$$x_0^T (x^T x)^{-1} x_0 = [1 \ 3 \ 8 \ 9] \begin{bmatrix} 8.0648 & -0.0826 & -0.0942 \\ -0.0826 & 0.0085 & 0.0017 \\ -0.0942 & 0.0017 & 0.0166 \\ -0.7905 & 0.0037 & -0.0021 \end{bmatrix}$$

$$\begin{bmatrix} -0.7905 \\ 0.0037 \\ -0.0021 \\ 0.0886 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ 8 \\ 9 \end{bmatrix}$$

$$= 0.1267$$

Using the mean square error,  $s^2 = 4.298$ ,  $s = 2.073$

$t_{0.025} = 2.262$  for 9 degrees of freedom.

$\therefore$  A 95% confidence interval for the mean percent survival for  $x_1 = 3$ ,  $x_2 = 8$  and  $x_3 = 9$  is given by:

$$24.2232 - (2.262)(2.073)\sqrt{0.1267} < \mu_{Y|1,3,8,9}$$

$$< 24.2232 + (2.262)(2.073)\sqrt{0.1267}$$

$$\Rightarrow 22.5541 < \mu_{Y|1,3,8,9} < 25.8923.$$

A prediction interval for a single predicted response  $y_0$  is once again established by considering the difference  $\hat{y}_0 - y_0$ . The sampling distribution can be shown to be normal with mean:

$$\mu_{\hat{y}_0 - y_0} = 0$$

and variance

$$\sigma^2_{\hat{y}_0 - y_0} = \sigma^2 [1 + x_0^T (x^T x)^{-1} x_0].$$

Thus, a  $100(1-\alpha)\%$  prediction interval for a single prediction value  $y_0$  can be constructed from the statistic

$$T = \frac{\hat{y}_0 - y_0}{s \sqrt{1 + x_0^T (x^T x)^{-1} x_0}}$$

which has a t-distribution with  $n-k-1$  degrees of freedom.

Prediction interval is less certain than CI

P.I predicts an individual number whereas a CI

predicts the mean value

P.I focuses on future events

(I)

past or current event

### Prediction interval for $y_0$ :

$\pm 100(1-\alpha)\%$ . prediction for a single response  $y_0$  is given by

$$\hat{y}_0 - t_{\alpha/2} s \sqrt{1 + x_0^T (x^T x)^{-1} x_0} < y_0 < \hat{y}_0 + t_{\alpha/2} s \sqrt{1 + x_0^T (x^T x)^{-1} x_0}$$

where  $t_{\alpha/2}$  is a value of the t-distribution with  $n-k-1$  degrees of freedom.

### Problem:-

- 3) Using the previous problem data, construct a 95% prediction interval for an individual percent survival response when  $x_1 = 3$ ,  $x_2 = 8$  and  $x_3 = 9$ .

### Solution:-

The 95% prediction interval for the response  $y_0$ , when  $x_1 = 3\%$ ,  $x_2 = 8\%$ , and  $x_3 = 9\%$  is

$$24.2232 - (2.262)(2.073) \sqrt{1.267} < y_0 <$$

$$24.2232 + (2.262)(2.073) \sqrt{1.1267}$$

$$(i.e.) 19.2459 < y_0 < 29.2005.$$

The prediction interval is wider than the confidence interval for mean percent survival. This test is used to determine the influence that independent variables have on the dependent variable in a regression analysis of variance in multiple regression:- study

The partition of the total sum of squares

$$\sum_{i=1}^n (y_i - \bar{y})^2 \text{ into its two components, the regression}$$

model and error sum of squares. The analysis of variance leads to a test of

$$H_0: \beta_1 = \beta_2 = \beta_3 = \dots = \beta_k = 0.$$

Rejection of the null hypothesis has an important interpretation for the scientist or engineer.

The vector of least squares estimator is given by

$$\hat{b} = (x'x)^{-1} x'y$$

\* partition of the uncorrected sum of squares

$$y'y = \sum_{i=1}^n y_i^2$$

into two components is given by

$$y'y = \hat{b}'x'y + (y'y - \hat{b}'x'y)$$

$$= y'x(x'x)^{-1}x'y + [y'y - y'x(x'x)^{-1}x'y]$$

The second term on the right-hand side is simply the error sum of squares  $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ . The alternative expression for the error sum of squares is

$$SSE = y' [I_n - x(x'x)^{-1}x']y.$$

The term  $y'x(x'x)^{-1}x'y$  is called the regression sum of squares.

To test if the regression equation differs significantly from zero, that is

$$H_0: \beta_0 = \beta_1 = \beta_2 = \dots = \beta_K = 0.$$

it states that the mean response is a constant, not necessarily zero.

### Degrees of Freedom:-

The partition of sum of squares and degrees of freedom reduces to

source	sum of squares	d.f
Regression	$\sum_{i=1}^n \hat{y}_i^2 = y^T x (x^T x)^{-1} x^T y$	$K+1$
Error	$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = y^T [I_n - x(x^T x)^{-1} x^T] y$	$n - K - 1$
Total	$\sum_{i=1}^n y_i^2 = y^T y$	$n$

### Coefficient of determination ( $R^2$ ):-

The coefficient of determination is used to illustrate the sufficiency adequacy of a fitted regression model.

$$\begin{aligned} R^2 &= \frac{SSR}{SST} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= 1 - \frac{SSE}{SST} \end{aligned}$$

$R^2$  is the square of the measure of association which indicates the percent of overlap between the

$R^2$  is the percentage of the response variable variation that is explained by a linear model.

It is always between 0 and 100%.

The adjusted coefficient of determination ( $R^2_{adj}$ ):

### Problems:-

1) Compute and interpret the coefficient of determination, for the variables.

$y$	$x_1$	$x_2$
6.40	1.32	1.15
15.05	2.69	3.40
18.75	3.56	4.10
30.25	4.41	8.75
44.85	5.36	14.82
48.94	6.20	15.15
51.55	7.12	15.32
61.50	8.87	18.18
100.44	9.80	35.19
111.42	10.65	40.40

Solution:-

$$\bar{y} = \frac{6 \cdot 40 + 15 \cdot 05 + \dots + 111 \cdot 42}{10} = \frac{489 \cdot 15}{10} = 48.915$$

$$n b_0 + b_1 \sum_{i=1}^n x_{1i} + b_2 \sum_{i=1}^n x_{2i} = \sum_{i=1}^n y_i$$

$$b_0 \sum_{i=1}^n x_{1i} + b_1 \sum_{i=1}^n x_{1i}^2 + b_2 \sum_{i=1}^n x_{1i}x_{2i} = \sum_{i=1}^n x_{1i}y_i$$

$$b_0 \sum_{i=1}^n x_{2i} + b_1 \sum_{i=1}^n x_{2i} x_{3i} + b_2 \sum_{i=1}^n x_{2i}^2 = \sum_{i=1}^n x_{2i} y_i$$

$y_i$	$x_{1i}$	$x_{2i}$	$x_{1i}^2$	$x_{2i}^2$	$x_{1i}x_{2i}$	$x_{1i}y_i$	$x_{2i}y_i$
6.40	1.32	1.15	1.7424				
15.05	2.69	3.40					
18.75	3.56	4.10					
30.25	4.41	8.75					
44.85	5.85	14.82					
48.94	6.20	15.15					
51.55	7.12	15.32					
61.50	8.87	18.18					
100.44	9.8	35.19					
111.42	10.65	40.40					
489.15	59.97	156.46	446.9965	3991.1208	1282.5215	3875.93	11749.8781

$$10b_0 + 59.97b_1 + 156.46b_2 = 489.15$$

$$59.97b_0 + 446.9965b_1 + 1282.5215b_2 = 3875.93$$

$$156.46b_0 + 1282.5215b_1 + 3991.1208b_2 = 11749.8781$$

$$b_0 = 0.51998, \quad b_1 = 2.71223, \quad b_2 = 2.049707$$

$$\hat{y} = 0.51998 + 2.71223x_1 + 2.049707x_2$$

$y_i$	$\hat{y}_i$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$\hat{y}_i - \bar{y}$	$(\hat{y}_i - \bar{y})^2$
6.40	6.517	-42.515	1801.525	-42.898	1791.590
15.05	14.845	-34.07	1160.765		
18.75	18.639	27.2	446.838	-34.07	1160.765
30.25	30.476	-30.105	916.636		
44.85	45.467	44.2	909.924	-30.276	916.636
48.94	48.449	-18.655	340.107		
51.55	51.292	-3.448	848.882	-18.439	339.997
61.50	61.901	18.666	11.916		
100.44	99.289	5.374	16.524	-3.448	11.889
111.42	112.213	62.358	3906.845	0.217	

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

$$= 1797.690 + 1160.765 + \dots + 4014.236.$$

$$= 10953.156$$

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

$$= 1807.525 + 1146.838 + \dots + 3906.875.$$

$$= 10956.223$$

$$R^2 = \frac{SSR}{SST} = \frac{10953.156}{10956.223} = 0.99$$

The adjusted coefficient of determination :- ( $R^2_{adj}$ )

Adjusted  $R^2$  is a variation on  $R^2$  that provides coefficient of determination. Adjusted  $R^2$  is

$R^2$  does not decrease as the error degrees of freedom  $n-k-1$  are reduced.

$$R^2_{adj} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$

Test on an Individual coefficient:-

The addition of any single variable to a regression system will increase the regression sum of squares and thus reduce the error sum of squares.

Problem:- (Testing ↑) (survival rate of species)

solution:-

$$\hat{y} = 39.1574 + 1.0161x_1 - 1.8616x_2 - 0.3433x_3$$

Let us test the hypothesis. By assuming

$$H_0 : \beta_3 = 0$$

$$H_1 : \beta_3 \neq 0.$$

$$\begin{aligned} t &= \frac{b_3 - \beta_{30}}{s\sqrt{c_{33}}} = \frac{b_3 - 0}{s\sqrt{c_{33}}} \\ &= \frac{-0.3433}{(2.073)\sqrt{0.0886}} = -0.556. \end{aligned}$$

which indicates that  $\beta_3$  does not differ significantly from zero, and hence we may very well justified in removing  $x_3$  from the model.

Let us consider the regression of  $y$  on the set  $(x_1, x_2)$ , the least squares normal equations now

reducing to

$$\begin{bmatrix} 13.0 & 59.43 & 81.82 \\ 59.43 & 394.7255 & 360.6621 \\ 81.82 & 360.6621 & 576.7264 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 877.50 \\ 1877.5670 \\ 2246.6610 \end{bmatrix}$$

The estimated regression coefficients for this reduced model are

$$b_0 = 36.094, b_1 = 1.031, b_2 = -1.810.$$

$$\hat{y} = 36.094 + 1.031x_1 - 1.810x_2$$

$R(\beta_1, \beta_2)$  - This can be calculated regression sum of squares.

$$R(\beta_1, \beta_2) = 398.12$$

Here we use the notation  $R(\beta_1, \beta_2)$  to indicate the regression sum of squares of the restricted model; it should not be confused with SSR, the regression sum of squares of the original model with 3 degrees of freedom.

The new error sum of squares is then

$$SST - R(\beta_1, \beta_2) = 438.13 - 398.12 = 40.01$$

and the resulting mean square error with 10 degrees of freedom becomes

$$s^2 = \frac{40.01}{10} = 4.001$$

### Special case of Orthogonality:-

Consider the  $x$  matrix as given below.

$$x = [1, x_1, x_2, \dots, x_k]$$

where 1 represents a column of ones and  $x_j$  is a column vector representing the levels of  $x_j$ . If

$$x_p' x_q = 0 \text{ for } p \neq q$$

the variables  $x_p$  and  $x_q$  are said to be orthogonal to each other. There are certain obvious advantages to having a completely orthogonal situation where  $x_p' x_q = 0$ .

for all possible  $p$  and  $q$ ,  $p+q$  and, in addition

$$\sum_{i=1}^n x_{ji} = 0, \quad j=1, 2, \dots, k$$

The resulting  $\mathbf{x}'\mathbf{x}$  is a diagonal matrix, and the normal equations reduce to

$$nb_0 = \sum_{i=1}^n y_i, \quad b_1 \sum_{i=1}^n x_{1i}^2 = \sum_{i=1}^n x_{1i} y_i, \dots$$

$$b_k \sum_{i=1}^n x_{ki}^2 = \sum_{i=1}^n x_{ki} y_i$$

An important advantage is that one is easily able to partition SSR into single-degree-of-freedom components, each of which corresponds to the amount of variation in  $y$  accounted for by a given controlled variable.

In the orthogonal situation, we can write

$$\begin{aligned} SSR &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^n (b_0 + b_1 x_{1i} + \dots + b_k x_{ki} - b_0)^2 \\ &= b_1^2 \sum_{i=1}^n x_{1i}^2 + b_2^2 \sum_{i=1}^n x_{2i}^2 + \dots + b_k^2 \sum_{i=1}^n x_{ki}^2 \end{aligned}$$

$$\text{#} = R(\beta_1) + R(\beta_2) + \dots + R(\beta_k)$$

The quantity  $R(\beta_i)$  is the amount of the regression sum of squares associated with a model involving a single independent variable  $x_i$ .

To test simultaneously for the significance of a set of  $m$  variables in an orthogonal situation, the regression sum of squares becomes

$$R(\beta_1, \beta_2, \dots, \beta_m) | \beta_{m+1}, \beta_{m+2}, \dots, \beta_K)$$

$$= R(\beta_1) + R(\beta_2) + \dots + R(\beta_m)$$

and thus we have the further simplification

$$R(\beta_1 | \beta_2, \beta_3, \dots, \beta_K) = R(\beta_1)$$

when evaluating a single independent variable.

### Analysis of variance for Orthogonal Variables

Source of Variation	Sum of squares	Degrees of freedom	Mean square	computed f
$\beta_1$	$R(\beta_1) = b_1^2 \sum_{i=1}^n x_{1i}^2$	1	$R(\beta_1)$	$\frac{R(\beta_1)}{s^2}$
$\beta_2$	$R(\beta_2) = b_2^2 \sum_{i=1}^n x_{2i}^2$	1	$R(\beta_2)$	$\frac{R(\beta_2)}{s^2}$
$\vdots$				
$\beta_K$	$R(\beta_K) = b_K^2 \sum_{i=1}^n x_{Ki}^2$	1	$R(\beta_K)$	$\frac{R(\beta_K)}{s^2}$
Error	SSE	$n-K-1$	$s^2 = \frac{SSE}{n-K-1}$	9.22
Total	$SST = Syy$	$n-1$		

### Problem:-

- 1) suppose that a scientist takes experimental data on the radius of a propellant grain  $y$  as a function of powder temperature  $x_1$ , extension rate  $x_2$  and die temperature  $x_3$ . Fit a linear regression model for predicting grain radius and determine the effectiveness of each variable in the model.

Braun radius	Powder temperature	extension rate	Die temperature
82	150 (-1)	12 (-1)	220 (-1)
93	190 (+1)	12 (-1)	220 (-1)
114	150 (-1)	24 (+1)	220 (-1)
124	150 (-1)	12 (-1)	250 (+1)
111	190 (+1)	24 (+1)	250 (+1)
129	150 (-1)	12 (-1)	250 (+1)
157	190 (+1)	24 (+1)	250 (+1)
164			

solution:-

Note that each variable is controlled at two levels, and the experiment is composed of the eight possible combinations.

The data on the independent variables are coded for convenience by means of the following formulae.

$$x_1 = \frac{\text{Powder Temperature} - 170}{20}$$

$$x_2 = \frac{\text{extension rate} - 18}{6}$$

$$x_3 = \frac{\text{die temperature} - 235}{15}$$

The resulting levels of  $x_1$ ,  $x_2$  and  $x_3$  take on the values -1 and +1.

$$X = \begin{bmatrix} 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & +1 & -1 \\ 1 & +1 & -1 & 1 \\ 1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & +1 & 1 & 1 \\ 1 & -1 & 1 & 1 \end{bmatrix}$$

We can now compute coefficients:

$$b_0 = \frac{1}{8} \sum_{i=1}^8 y_i = 121.75$$

$$b_1 = \frac{1}{8} \sum_{i=1}^8 x_{1i} y_i = \frac{20}{8} = 2.5$$

$$b_2 = \frac{1}{8} \sum_{i=1}^8 x_{2i} y_i = \frac{118}{8} = 14.75$$

$$b_3 = \frac{1}{8} \sum_{i=1}^8 x_{3i} y_i = \frac{144}{8} = 21.75$$

so in terms of coded variables, the prediction eqn is

$$\hat{y} = 121.75 + 2.5x_1 + 14.75x_2 + 21.75x_3$$

Source of Variation	Sums of squares	Degrees of freedom	Mean square	Computed f	P-Value
$\beta_1$	$(2.5)^2(8) = 50$	1	50	2.16	0.2156
$\beta_2$	$(14.75)^2(8) = 1740.5$	1	1740.5	15.26	0.0016
$\beta_3$	$(21.75)^2(8) = 3784.5$	1	3784.5	168.25	0.0002
Error	<u>92.50</u>	4	23.13		
	<u>5667.5</u>	7			

The variable  $x_2$  and  $x_8$  both impact the grain yield in a positive fashion, with  $x_2$  being the more important factor based on the smallness of its P-value.

### CATEGORICAL OR INDICATOR VARIABLES:-

An extremely important special case application of multiple linear regression occurs when one or more of the regressor variables are categorical, indicator or dummy variables.

### MLR with categorical variable:-

Consider the data

i	Gender	Meet pay
Bob	M	9.6
Paul	M	8.3
Maeve	F	4.2
John	M	8.8
Nancy	F	2.1
Tim	M	6.0
George	M	1.1
Alan	M	9.2
Idea	F	3.3
Anne	F	2.7

Here the dependent variable  $y$ , is meet pay increase measured in percent and the independent variable is gender which is quite obviously a nominal or categorical variable. We need to convert the categorical variable gender into a form that "makes sense" to regression

analysis. One way to represent a categorical variable is to code the categories 0 and 1 as follows:

Let  $x = 1$  if gender is "male"  
0 otherwise

for Eg:-

Bob is scored '1' because he is male; Mary is such  $x$  are called dummy variables. A dummy variable, in other words, is a numerical representation of the categories of a nominal or ordinal variable. By coding  $x$  with scores of 1 and 0 we can transform the above table into a set of data that can be analyzed with regular regression

i	Gender	y	i	Gender	y
		Meet pay			Meet pay
Bob	1	9.6	Tim	1	6.0
Paul	1	8.3	George	1	1.1
Mary	0	4.2	Alan	1	9.2
John	1	8.8	Lisa	0	3.3
Nancy	0	2.1	Anne	0	2.7

The reason for using 1 and 0 and not any other values is that it is easier for usage and interpretation as  $x=1$  gives interpretation for males and 0 for females.

The regression equation for the above data is  $y = 3.08 + 4.09x$  with  $R^2 = 0.428$

The ~~and~~ estimate of the average merit pay increase for women in the population is 3.08 percent.

Men, on average, get 4.09 percent more than women. Hence their average increase is

$$\begin{aligned}\hat{\gamma}_{men} &= 3.08 + 4.09(1) \\ &= 7.17\end{aligned}$$

The effect of being male is 4.09 percent greater merit pay than what women get.

### Multicollinearity:-

In regression, "multicollinearity" refers to predictors that are correlated with each other predictors.

Multicollinearity occurs when your model includes multiple factors that are correlated not just to your response variable, but also to each other.

### Types of multicollinearity:-

\* Structural multicollinearity is a mathematical artifact caused by creating new predictors from other predictors such as, creating the predictor  $x^2$  from the predict  $x$ .

\* Data-based multicollinearity is a result of a poorly designed experiment, reliance on purely observational data, or inability to manipulate the system on which the data are collected.

## Variance Inflation Factor :-

The variance inflation factor is another way to express exactly the same information found in the coefficient of multiple correlation. A variance inflation factor is computed for each independent variable.

$$VIF_k = \frac{1}{1 - R_k^2}$$

where  $VIF_k$  is the variance inflation factor for variable  $k$ , and  $R_k^2$  is the coefficient of multiple determination for variable  $k$ .

VIF	state of Predictor
$VIF = 1$	Not correlated
$1 < VIF < 5$	Moderately correlated
$VIF > 5$ to $10$	Highly correlated

## STEPWISE REGRESSION:-

One standard procedure for searching for the "optimum subset" of variables is the absence of orthogonality is a technique called stepwise regression. It is based on the procedure of sequentially introducing the variables into the model one at a time. Given a predetermined  $\alpha$ , the description of the stepwise routine will be better understood if the methods of forward selection and backward elimination are described first.

### Forward selection:-

Forward selection procedure begins with no explanatory variable in the model and sequentially adds a variable according to the criterion of partial F-statistic. At each step, a variable is added; whose partial F-statistic yields the smallest p-value. Variables are entered as long as the partial F-statistic p-value remains below a specific maximum value,  $(PIN)$ . The procedure stops when the addition of any of the remaining variables yields a partial p-value  $> PIN$ . This procedure has two limitations. Some of the variables never get into the model and hence their importance is never determined. Another limitation is that a variable once included in the model remains there throughout the process, even if it loses its stated significance, after the inclusion of the other variables.

### Backward selection:-

The backward elimination procedure begins with all the variables in the model and proceeds by eliminating the least useful variable at a time. A variable, whose partial F p-value is greater than a perceived value,  $(POUT)$ , is the least useful variable and is therefore removed from the regression model. The process continues, until no variable can be removed according to the elimination criterion.

### Stepwise regression:

The stepwise procedure is a modified forward selection method which later in the process permits the elimination of variables that become statistically non-significant. At each step of the process, the p-values are computed for all variables in the model. If the largest of these p-values  $> P_{OUT}$ , then that variable is eliminated. After the included variables have been examined for exclusion, the excluded variables are re-examined for inclusion. At each step of the process, there can be at most one exclusion, followed by one inclusion. It is necessary that  $P_{IN} < P_{OUT}$  to avoid infinite cycling of the process.

### Problem:-

- Using the techniques of stepwise regression, find an appropriate linear regression model for predicting the length of infants for the data.

Infant length (y)	age ( $x_1$ )	Length at Birth ( $x_2$ )	Weight at Birth, ( $x_3$ )	Chest size at Birth, ( $x_4$ )
51.5	78	48.2	2.75	29.5
52.8	69	45.5	2.15	26.3
61.3	77	46.3	4.41	32.2
67.0	78	49.0	5.52	36.5
53.5	67	43.0	3.21	27.2
62.7	80	48.0	4.32	27.7
56.2	74	48.0	2.31	28.3
68.5	94	53.0	4.30	30.3
69.2	102	58.0	3.71	28.7

Solution:-

Step 1:-

considering each variable separately, four individual simple linear regression equations are fitted. The following pertinent regression sums of squares are computed.

$$R(\beta_1) = 288.1468, \quad R(\beta_2) = 215.3013$$

$$R(\beta_3) = 186.1065, \quad R(\beta_4) = 100.8594$$

Variable  $x_1$  clearly gives the largest regression sum of squares. The mean square error for the equation involving only  $x_1$  is  $s^2 = 4.7276$ , and since

$$f = \frac{R(\beta_1)}{s^2} = \frac{288.1468}{4.7276} = 60.9600,$$

which exceeds  $t_{0.05}(1,7) = 5.59$ , the variable  $x_1$  is

significant and is entered into the model.

Step 2:-

Three regression equations are fitted at this stage, all containing  $x_1$ . The important results for the combinations  $(x_1, x_2)$ ,  $(x_1, x_3)$  and  $(x_1, x_4)$  are

$$R(\beta_2 | \beta_1) = 23.8703, \quad R(\beta_3 | \beta_1) = 29.3086,$$

$$R(\beta_4 | \beta_1) = 13.8178.$$

Variable  $x_3$  displays the largest regression sum of squares in the presence of  $x_1$ . The regression involving  $x_1$  and  $x_3$  gives a new value of  $s^2 = 0.6307$ , and

since

$$f = \frac{R(\beta_3 | \beta_1)}{s^2} = \frac{29.3086}{0.6307} = 46.47$$

which exceeds  $f_{0.05}(1, 6) = 5.99$ , the variable  $x_3$  is significant and is included along with  $x_1$  in the model. Now we must subject  $x_1$  in the presence of  $x_3$  to a significance test. We find that  $R(\beta_1 | \beta_3) = 131.349$  and hence.

$$f = \frac{R(\beta_1 | \beta_3)}{s^2} = \frac{131.349}{0.6307} = 208.26$$

which is highly significant. Therefore,  $x_1$  is retained along with  $x_3$ .

Step 3 :-

With  $x_1$  and  $x_3$  already in the model, we now require  $R(\beta_2 | \beta_1, \beta_3)$  and  $R(\beta_4 | \beta_1, \beta_3)$  in order to determine which, if any, of the remaining two variable is entered at this stage. From the regression analysis using  $x_2$  along with  $x_1$  and  $x_3$ , we find  $R(\beta_2 | \beta_1, \beta_3) = 0.7948$  and when  $x_4$  is used along with  $x_1$  and  $x_3$ , we obtain

$$R(\beta_4 | \beta_1, \beta_3) = 0.1855$$

The value of  $s^2$  is 0.5979 for the  $(x_1, x_2, x_3)$  combination and 0.7198 for the  $(x_1, x_3, x_4)$  combination.

since neither f-value is significant at the  $\alpha = 0.05$  level, the final regression model includes only the variables  $x_1$  and  $x_3$ .  
The estimating equation is found to be

$$\hat{y} = 20.1084 + 0.4136 x_1 + 2.0253 x_3$$

and the coefficient of determination for this model is  $R^2 = 0.9882$ .

Although  $(x_1, x_3)$  is the combination chosen by the stepwise regression.

## UNIT-II

### TIME SERIES FORECASTING

#### Regression Model for forecasting:-

A forecast is merely a prediction about values of data. However, most extrapolative models assume that the past is a proxy for the future. There are many traditional models for forecasting, such as smoothing, regression, time series, and econometric models. Forecasting often involves expert judgement or model forecasts, often involving expert judgement.

Regression analysis is a statistical method used to analyze quantitative data to estimate parameters and make forecasts. Regression can be expanded to include more than one variable. Regressions involving more than one variable are called multiple regressions.

variable are referred to as multiple regressions.

The case of simple, linear, least square regression may be written in the form

$$Y = \alpha + \beta X + \epsilon,$$

where  $Y$ , the dependent variable, is a linear function  $X$ , the independent variable. The parameters  $\alpha$  and  $\beta$  characterize the population regression line and  $\epsilon$  is the randomly distributed error term. The regression estimates of  $\alpha$  and  $\beta$  will be derived from the principle of least squares.

#### Method of Least Square:-

The trend project method fits a trend line to a series of historical data points and then projects the line into the future for medium-to-long range forecasts. Several mathematical trend equations can be developed, depending upon movement of time-series data.

#### Linear Trend Model:

For linear equations, it is found by the simultaneous solution for  $a$  and  $b$  of the two normal equations:

$$\sum y = na + b \sum x$$

$$\sum xy = a \sum x + b \sum x^2$$

Problem:-

1) Below are given the figures of production of a sugar factory:

Year :-	1992	1993	1994	1995	1996	1997	1998
Production :-	80	90	92	83	94	99	92

a) Fit a straight line trend to these figures.

b) Plot these figures on a graph and show the trend line.

c) Estimate the production in 2001.

Solution:-

Year	Production (y)	Time period (x)	$x^2$	$xy$
1992	80	1	1	80
1993	90	2	4	180
1994	92	3	9	276
1995	83	4	16	332
1996	94	5	25	470
1997	99	6	36	594
1998	92	7	49	644
	630	28	140	2576

$$1a + 28b = 630$$

$$28a + 140b = 2576$$

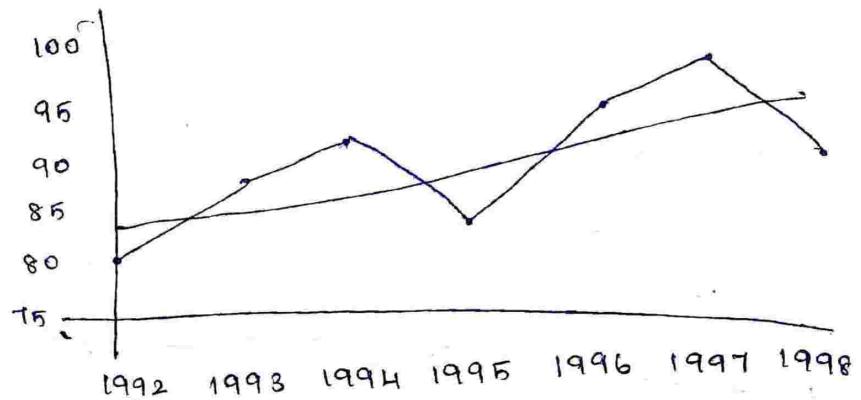
From solving these equations,  $a = 82$ ,  $b = 2$

Therefore, linear trend component for the production

of sugar is,

$$y = a + bx = 82 + 2x$$

b)



c) The production of sugar for year 2001 will be  $\hat{y} = 82 + 2(10) = 102$  thousand quintals.

d) calculate the linear trend values from the following data using the method of least squares.

Year :-	1984	1985	1986	1987	1988	1989
Production :-	7	9	12	15	18	23

Parabolic trend model:-

The curvilinear relationship for estimating the value of a dependent variable  $y$  from an independent variable  $x$  might take the form

$$\hat{y} = a + bx + cx^2$$

This trend line is called the parabola.

For a non-linear equation  $\hat{y} = a + bx + cx^2$ , the values of constants  $a, b$  and  $c$  can be determined by solving three normal equations.

$$\Sigma y = na + b \Sigma x + c \Sigma x^2$$

$$\Sigma xy = a \Sigma x + b \Sigma x^2 + c \Sigma x^3$$

$$\Sigma x^2 y = a \Sigma x^2 + b \Sigma x^3 + c \Sigma x^4$$

Problems:-

- 1) The prices of a commodity during 1999 - 2004 are given below. Fit a parabola to these data. Estimate the price of the commodity for the year 2005.

Year	Price
1999	100
2000	107
2001	128
2002	140
2003	181
2004	192

Also plot the actual and trend values on graph.

Year	(x)	y	$x^2$	$x^3$	$x^4$	$\Sigma y$	$\Sigma x^2 y$
1999	1	100	1	1	1	100	100
2000	2	107	4	8	16	214	428
2001	3	128	9	27	81	384	1152
2002	4	140	16	64	256	560	2240
2003	5	181	25	125	625	905	4525
2004	6	192	36	216	1296	1152	6912
	21	848	91	441	2275	3315	15357

$$848 = ba + 21b + 91c = 848$$

$$21a + 91b + 441c = 3315$$

$$91a + 441b + 2275c = 15357$$

$$a = 88.6, b = 7.33, c = 1.785$$

$$\hat{Y} = 88.6 + 7.33x + 1.785x^2$$

Forecasting time series data with seasonal variation:-

Method of averages:-

Moving averages are very widely used in business and economics. They are used to smoother the data.

Problems:-

- 1) Using three-yearly moving averages, determine the trend and short-term errors:

Year	Production	Year	Production
1987	21	1992	22
1988	22	1993	25
1989	23	1994	26
1990	25	1995	27
1991	24	1996	26

Solution:-

The moving average calculation for the first 3 years is:

Year	Production (Y)	3-Year moving average	Forecast error ( $y - \hat{y}$ )
1987	21	-	-
1988	22	22	0
1989	23	23.33	-0.33
1990	25	24	1
1991	24	23.67	0.33
1992	22	23.67	-1.67
1993	25	24.33	0.67
1994	26	25	0.67
1995	27	26.33	-
1996	26	-	-

2) Assume a four-yearly cycle and calculate the trend by the method of moving average from the following data relating to the production of tea in India.

Year	1987	1988	1989	1990	1991	1992
Production :-	464	515	518	467	502	540
1993	557	571	586	612		
1994						

Year	Production	4-Yearly moving average	4-Yearly moving average centered
1987	464	-	-
1988	515	491	496.75
1989	518	500.5	503.62
1990	467	506.75	511.62
1991	502	516.5	529.5
1992	540	542.5	553
1993	557	563.5	572.5
1994	571	581.5	-
1995	586	-	-
1996	612	-	-

## Weighted moving averages:-

In moving averages, each observation has equal importance (weight). A weighted average may be expressed mathematically as

$$\text{Weighted moving average} = \frac{\sum \text{weight for per value}}{\sum \text{weights}}$$

Problems:-

data.

### Methods of Measuring Seasonal Variations:-

The measurement of seasonal variation isolating them from other components of a time series is commonly used for the study of seasonal variation. There are three methods commonly used for the study of seasonal variation. These methods are:

- 1) Method of simple averages
- 2) Ratio to trend method
- 3) Ratio to moving average method.

#### Method of Simple averages:-

This method is used when the time series consists of only the seasonal and random component. The effect of taking average of data corresponding to same period is to eliminate the effect of random component and thus, the resulting average will contain only seasonal component. These averages are called seasonal averages.

into seasonal indices.

Problems:-

- 1) Assuming that trend and cyclical variations are absent  
compute the seasonal index for each month of the  
following data of sales of a company.

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1987	H6	H5	H4	H6	H5	H7	H6	H3	H0	H0	H1	H5
1988	H5	H4	H3	H6	H6	H5	H7	H2	H3	H2	H3	H4
1989	H2	H1	H0	H4	H5	H5	H6	H3	H1	H0	H2	H5

Solution:-

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1987	H6	H5	H4	H6	H5	H7	H6	H3	H0	H0	H1	H5
1988	H5	H4	H3	H6	H6	H5	H7	H2	H3	H2	H3	H4
1989	H2	H1	H0	H4	H5	H5	H6	H3	H1	H0	H2	H5
A <sub>t</sub>	44.33	43.33	42.33	45.33	45.33	45.67	46.33	42.67	41.33	40.67	H2	H4.67
S.I	101.5	99.2	96.9	103.8	103.8	104.6	106.1	97.7	94.7	93.1	96.2	102.3

$$\text{Seasonal indices} = \frac{A_t}{G_1} \times 100 \quad \text{where } G_1 = \frac{\sum A_t}{12}$$

$$G_1 = \frac{\sum A_t}{12} = \frac{523.96}{12} = 43.66$$

$$\text{Further } \sum S.I = 1199.9 \neq 1200$$

Thus, we have to adjust these values such that their total is 1200.

$$\text{Each figure} \times \frac{1200}{1199.9}$$

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
S.I	101.5	99.2	96.9	103.8	103.8	104.6	106.1	97.7	94.7	93.1	96.2	102.3

∴ No adjustment is needed.

2) Compute the seasonal index from the following data by the method of simple averages.

Year	Quarter	y	Year	Quarter	y	Year	Quarter	y
1980	I	106	1982	I	90	1984	I	80
	II	124		II	112		II	104
	III	104		III	101		III	95
	IV	90		IV	85		IV	83
1981	I	84	1983	I	76	1985	I	104
	II	114		II	94		II	112
	III	107		III	91		III	102
	IV	88		IV	76		IV	84

Solution:-

Years	1 <sup>st</sup> Qtr	2 <sup>nd</sup> Qtr	3 <sup>rd</sup> Qtr	4 <sup>th</sup> Qtr
1980	106	124	104	90
1981	84	114	107	88
1982	90	112	101	85
1983	76	94	91	76
1984	80	104	95	83
1985	104	112	102	84
A <sub>t</sub>	90	110	100	84.33
S.I	93.67	114.49	104.08	87.77

$$G_1 = \frac{A_t}{4} = 96.08$$

$$S.I = \frac{A_t \times 100}{G_1}$$

Further, since the sum of terms in the last row of the table is 400, no adjustment is needed.

max 15

## Auto regressive Models:-

In a multiple regression model the variable of interest using a set of predictors. In an auto regression m

forecast the variable of interest using a linear combination of past values of the variable. The term auto regression indicates that it is a regression of the variable against itself measured at different time points.

Hence, an autoregressive model of order  $P$  can

be written as

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \dots + \beta_P Y_{t-P} + \epsilon_t$$

where  $\epsilon_t$  is white noise. This is like a multiple regression but with lagged values of  $Y_t$  as predictors. This is an AR( $K$ ) model, an autoregressive model of order  $K$ .

First-order autoregressive model:

$$\text{AR}(1): Y_t = \beta_0 + \beta_1 Y_{t-1} + \epsilon_t$$

Second order autoregressive model:

$$\text{AR}(2): Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \epsilon_t$$

Problems:-

- Given an AR(1) model where  $\beta_0 = 2$  and  $\beta_1 = 1.8$ , the one-step-ahead forecast of  $Y_1$  when  $Y_0 = 2$

Solution:-

$$Y_t = \beta_0 + \beta_1 Y_{t-1} \quad (1)$$

$$Y_1 = \beta_0 + \beta_1 Y_0$$

$$= 2 + 1.8(2) = 5.6$$

Two-step

Let us calculate the two-step ahead forecast  
of  $y_2$

Replace  $t$  by  $t+1$  in (1)

$$y_{t+1} = \beta_0 + \beta_1 y_t$$

put  $t=1$ ,

$$\begin{aligned} y_2 &= \beta_0 + \beta_1 y_1 \\ &= 2 + (1.8)(5.6) = 12.08. \end{aligned}$$

Q) Consider an AR(1) model with the following prediction equation:-

$$x_t = 0.8 + 0.5 x_{t-1}$$

If the current value of  $x$  is 4.0, the two-step-ahead forecast is closest to:

Solution:-

One-step ahead forecast.

$$\text{If } x_t = 4; \text{ then } x_{t+1} = 0.8 + 0.5(4) = 1.6.$$

Two-step ahead forecast.

$$\begin{aligned} \text{If } x_{t+1} = 1.6, \text{ then } x_{t+2} &= 0.8 + 0.5(1.6) \\ &= 0.64. \end{aligned}$$

Auto correlation and Partial Auto correlation

function:-

The coefficient of correlation between two values in a time series is called the auto-correlation function (ACF).

For example, the ACF for a time series  $y_t$  is given by

$$s_k = \text{cov}(y_t, y_{t-p}) = \frac{\text{cov}(y_t, y_{t-p})}{\sqrt{\text{var} y_t} \sqrt{\text{var} y_{t-p}}} = \frac{r_p}{r_0}$$

This value of  $k$  is the time gap being considered and is called the lag.

At lag 1, auto correlation (i.e.  $k=1$ ) is the correlation between values that are one time period apart.

More generally, a lag  $k$  autocorrelation is the correlation between values that are  $k$  time periods apart.

The ACF is a way to measure the linear relationship between an observation at time  $t$  and the observations

at previous times. If we assume an AR( $k$ ) model, then we may wish to only measure the association between  $y_t$  and  $y_{t-k}$  and filter out the linear influence of

the random variables that lie in between [i.e.]  $y_{t-1}, y_{t-2}, \dots, y_{t-(k-1)}$ , which requires a transformation

on the time series. Then by calculating the correlation of the transformed time series, we obtain the partial autocorrelation function (PACF).

The PACF is most useful for identifying the order of an autoregressive model.

A plot of partial autocorrelation for different values of  $k$  is called PACF.

### Moving average model :-

A moving average model is one where  $y_t$  depends on the random error terms which follow a white noise process. (i.e.)

A common representation of moving average model where it depends on  $q$  of its past values is called MA( $q$ ) model and is represented as follows

$$y_t = \cancel{\phi} + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \dots + \phi_q \varepsilon_{t-q}$$

The error terms  $\varepsilon_t$  are assumed to be white noise processes with mean zero and ~~variate~~ variance  $\sigma^2$ .

### Auto regressive moving average model :-

The notation ARMA( $p, q$ ) refers to the model with  $p$  autoregressive terms and  $q$  moving average terms. This model contains the AR( $p$ ) and MA( $q$ ) models.

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \dots + \beta_p y_{t-p} + \varepsilon_t$$

$$+ \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \phi_3 \varepsilon_{t-3} + \dots + \phi_q \varepsilon_{t-q}$$

### Auto regressive Integrated moving average model :-

An autoregressive integrated moving average model or ARIMA is a statistical analysis model that uses time series data to either better understand the data set or to predict future trends.

An autoregressive integrated moving average model is a form of regression analysis that gauges the strength of one dependent variable relative to other changing variables. The model's goal is to predict future securities or financial market moves by examining the differences between values in the series instead of through actual values.

An ARIMA model can be understood by

outlining each of its components as follows:

\* Autoregression (AR) refers to a model that shows a changing variable that regresses on its own lagged, or prior values.

\* Integrated (I) represents the differencing of raw observations to allow for the time series to become stationary (ie) data values are replaced by the difference between the data values and the previous values.

\* Moving Average (MA) incorporates the dependency between an observation and a residual error from a moving average model applied to lagged observations.

Each component functions as a parameter with a

standard notation. For ARIMA models, a standard notation would be ARIMA with p, d and q where p and q are integer values substitute for the parameters to indicate the type of ARIMA model used. The parameters can be defined as:

- p : the no. of lag observations in the model; also known as the lag order.
- d : the no. of times that the raw observations are differenced; also known as the degree of differencing.
- q : the size of the moving average window; also known as the order of the moving average.

### Stationary:

A stationary time series is one whose properties do not depend on the time at which the series is observed. Thus, time series with trends or with seasonality are not stationary - the trend and seasonality will affect the value of the time series at different times.

### Differencing:

Differencing is a popular and widely used data transform for making time series data stationary. It helps to remove trend and seasonal patterns. The first difference of a time series is the series of changes from one period to the next. If  $y_t$  denotes the value of the time series  $y$  at period  $t$ , then the first difference of  $y$  at period  $t$  is equal to

$$y_t - y_{t-1}$$

to get rid of it  
is helpful at no value.

### UNIT ROOT TEST:-

A unit root test tests whether a time series is not stationary and consists of a unit root in time series analysis. The presence of a unit root in time series defines the null hypothesis and the alternative hypothesis defines time series as stationary.

Mathematically the unit root test can be represented as

$$y_t = D_t + z_t + \varepsilon_t$$

where,  $D_t$  is deterministic component

$z_t$  is the stochastic component

$\varepsilon_t$  is the stationary error process.

The unit root test's basic concept is to determine whether the  $z_t$  consists of a unit root or not.

### DICKEY FULLER TEST:-

The simplest approach to test for a unit root begins with AR(1) model

$$y_t = \beta_0 + \beta y_{t-1} + \varepsilon_t, \text{ where}$$

$y_t$  is variable of interest at the time  $t$

$\beta$  is a coefficient that defines the unit root.

$\varepsilon_t$  is noise or can be considered as an error term.

Test

$$Y_t - Y_{t-1} = \beta_0 + \beta Y_{t-1} - Y_{t-1} + \varepsilon_t$$

$$\Delta Y_t = \beta_0 + (\beta - 1) Y_{t-1} + \varepsilon_t$$

$$\text{Hence } \Delta Y_t = \beta_0 + \delta Y_{t-1} + \varepsilon_t$$

where  $\delta$  is a difference operator,

$$\delta = \beta - 1$$

If  $\beta = 1$ , we will get the differencing as the error term and if the coefficient has some values smaller than 1 or bigger than one, we will see the changes according to the past observation.

There can be three versions of the test

$$\bullet \Delta Y_t = \delta Y_{t-1} + \varepsilon_t \text{ test for a unit root}$$

$$\bullet \Delta Y_t = \beta_0 + \delta Y_{t-1} + \varepsilon_t \text{ test for a unit root with constant}$$

$$\bullet \Delta Y_t = \beta_0 + \beta_1 t + \delta Y_{t-1} + \varepsilon_t \text{ test for a unit root with the constant and deterministic terms with time}$$

### AUGMENTED DICKEY-FULLER TEST:-

The augmented dickey-fuller test is an extension of the dickey-fuller test, which is used to remove autocorrelation from the series and then tests similar to the procedure of the dickey-fuller test.

The augmented dickey fuller test works on the statistic, which gives a negative number and rejection of the hypothesis depends on that negative number; the more negative magnitude of the number represents the confidence of presence of unit root at some level in the time series.

We apply ADF on a model, and it can be represented mathematically as

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \dots + \delta_{p-1} \Delta y_{t-p+1} + \epsilon_t$$

where,

$\alpha$  is a constant

$\beta$  is the coefficient at time

$p$  is the lag order of the autoregressive process.

Here in the mathematical representation of ADF, we have added the differencing terms that make changes between ADF and the Dickey-Fuller test.

ARIMA (p, d, q) model building :-

Given a time series  $x_t$ ,  $t=1, 2, \dots, n$ , the objective is to select the best fitted ARIMA (p, d, q) model.

The methodology of building ARIMA models according to Box and Jenkins consists of three stages

Stage 1:- Identification Stage or Model Selection

Autocorrelation and Partial Autocorrelation

### Stage 2: Estimation Stage or Parameter Estimation

Estimate the unknown autoregressive and moving average parameters.

### Stage 3: Checking Stage or Model checking

check the estimated results.

### Stage 1: Identification:-

- \* The two most useful tools in model identification are
  - the sample autocorrelation function and
  - the sample partial autocorrelation function
- \* These two functions will provide valuable information with regard to stationarity and to the true generating process.

### Stage II: Parameter Estimation

- \* Assuming that a particular ARIMA  $(p, d, q)$  model is selected the next step is to estimate the unknown parameters  $p, d$  and  $q$ .
- \* If  $d \neq 0$ , it is necessary to difference the series  $d$  times so that to get stationarity.
- \* The new series  $w_t = (1 - B)^d x_t$  will then be used to obtain estimates for the autoregressive, moving average and the mean of the series parameters.

- \* Frequently, after differencing, it is reasonable to assume that the new series has mean zero, in which case the parameter  $\mu$  is dropped.
- \* The estimation procedure for time series is not unique.

### Stage III:- Model checking

- \* fit a model that contains additional parameters.
- \* If an ARMA (p,q) model is fitted, then fit an alternative model with one or two autoregressive parameters and then with one or two moving average parameters.
- \* The resulting parameter estimators of the augmented model will have very large variances.

### Ljung - Box test:-

Ljung - Box is used to test for serial correlation in a time series. It is widely used in econometrics to determine whether there is structure in a time series to make it worth modelling. It is also applied to residuals after a forecast model has been fit the data.

$$\text{Ljung - Box statistic : } Q = n(n+2) \sum_{i=1}^K \frac{r_i^2}{n-i}$$

where  $r_i$  be the estimated autocorrelation between observations separated by  $i$  time periods.

### Theil's coefficient:

Theil's coefficient is an accuracy that emphasizes the importance of large errors as well as providing a relative basis for comparison with naive forecasting ~~method~~ methods.

$$U = \sqrt{\frac{\sum_{t=1}^{n-1} \left[ \frac{F_{t+1} - Y_{t+1}}{Y_t} \right]^2}{\sum_{t=1}^{n-1} \left[ \frac{Y_{t+1} - Y_t}{Y_t} \right]^2}}$$

where  $U$  = Theil's  $U$ -statistic

$F$  = forecast

$Y$  = observation.

Theil's  $U$  statistic can be interpreted as dividing the RMSE (Root mean square error or square root of the MSE) of the proposed forecasting method by the RMSE of a no-change model.

If  $U=1$ , it means that the proposed model is as good as the naive model. If  $U>1$ , there is no point in using the proposed forecasting model. ~~since a~~ ~~method~~ It is worthwhile to consider using the proposed model only when  $U<1$  (the smaller the better).

## UNIT - II

Discriminant analysis is defined as an appropriate technique when the dependent variable is categorical (nominal or nonmetric) and the independent variables are metric. The single dependent variable can have two, three or more categories.

Discrimination and classification are multivariate techniques concerned with separating distinct sets of objects and with allocating new objects to previously defined groups.

Discriminant analysis is rather exploratory in nature. Classification procedures are less exploratory in the sense that they lead to well-defined rules.

Classification ordinarily requires more problem structure than discrimination does. The immediate goals of discrimination and classification are as follows:

- 1) To describe, either graphically or algebraically, the differential features of objects from several known collections.
- 2) To sort objects into two or more labeled classes. The emphasis is on deriving a rule that can be used to optimally assign new objects to the labeled classes.

### The Discriminant function for two groups:-

We assume that the two populations to be compared have the same covariance matrix  $\Sigma$  but distinct mean vectors  $\mu_1$  and  $\mu_2$ . We work with samples  $y_{11}, y_{12}, \dots, y_{1n_1}$  and  $y_{21}, y_{22}, \dots, y_{2n_2}$  from the two population. Each vector  $y_{ij}$  consists of measurements on  $p$  variables. The discriminant function is the linear combination of these  $p$  variables that maximizes the distance between the two group mean vectors. A linear combination  $z = a'y$  transforms each observation vector to a scalar.

$$z_{1i} = a'y_{1i} = a_1 y_{1i1} + a_2 y_{1i2} + \dots + a_p y_{1ip}$$

$$z_{2i} = a'y_{2i} = a_1 y_{2i1} + a_2 y_{2i2} + \dots + a_p y_{2ip}$$

$$i = 1, 2, \dots, n_1, \quad i = 1, 2, \dots, n_2$$

Hence  $n_1 + n_2$  observation vectors in the two samples,

$$\begin{matrix} y_{11} & y_{21} \\ y_{12} & y_{22} \\ \vdots & \vdots \\ y_{1n_1} & y_{2n_2} \end{matrix}$$

are transformed to scalars.

$$\begin{matrix} z_{11} & z_{21} \\ z_{12} & z_{22} \\ \vdots & \vdots \\ z_{1n_1} & z_{2n_2} \end{matrix}$$

we find the means  $\bar{z}_1 = \sum_{i=1}^n \frac{z_{1i}}{n_1} = a' \bar{y}_1$

and  $\bar{z}_2 = a' \bar{y}_2$

where  $\bar{y}_1 = \sum_{i=1}^{n_1} \frac{y_{1i}}{n_1}$  and  $\bar{y}_2 = \sum_{i=1}^{n_2} \frac{y_{2i}}{n_2}$

The vector  $a$  that maximizes the standardized difference  $\frac{(\bar{z}_1 - \bar{z}_2)}{s_z}$ . Since  $\frac{(\bar{z}_1 - \bar{z}_2)}{s_z}$  can be negative,

we use squared distance  $\frac{(\bar{z}_1 - \bar{z}_2)^2}{s_z^2}$

$$\frac{(\bar{z}_1 - \bar{z}_2)^2}{s_z^2} = \frac{[a' (\bar{y}_1 - \bar{y}_2)]^2}{a' s_p a} \quad (1)$$

The maximum of (1) occurs when

$$a = s_p^{-1} (\bar{y}_1 - \bar{y}_2)$$

### Fisher's Discriminant Function:

Linear discriminant analysis is a generalized form of Fisher's linear discriminant. The basic idea of Fisher's linear discriminant is to project data points onto a line to maximize the between-class scatter and minimize the within-class scatter.

### LDA: Two classes

1) Compute the Linear discriminant projection for the following two dimensional dataset.

- Samples for class  $w_1$ :

$$x_1 = (x_1, x_2) = \{(4, 2), (2, 4), (2, 3), (3, 6), (4, 4)\}$$

- Samples for class  $w_2$ :

$$x_2 = (x_1, x_2) = \{(9, 10), (6, 8), (9, 5), (8, 7), (10, 8)\}$$

Solution:

The classes mean are:

$$\mu_1 = \sum_{i=1}^{n_1} \frac{x_{1i}}{n_1}$$

$$= \frac{1}{5} \left[ \begin{pmatrix} 4 \\ 2 \end{pmatrix} + \begin{pmatrix} 2 \\ 4 \end{pmatrix} + \begin{pmatrix} 2 \\ 3 \end{pmatrix} + \begin{pmatrix} 3 \\ 6 \end{pmatrix} + \begin{pmatrix} 4 \\ 4 \end{pmatrix} \right] = \begin{bmatrix} 3 \\ 3.8 \end{bmatrix}$$

$$\mu_2 = \sum_{i=1}^{n_2} \frac{x_{2i}}{n_2}$$

$$= \frac{1}{5} \left[ \begin{pmatrix} 9 \\ 10 \end{pmatrix} + \begin{pmatrix} 6 \\ 8 \end{pmatrix} + \begin{pmatrix} 9 \\ 5 \end{pmatrix} + \begin{pmatrix} 8 \\ 7 \end{pmatrix} + \begin{pmatrix} 10 \\ 8 \end{pmatrix} \right] = \begin{bmatrix} 8.4 \\ 7.6 \end{bmatrix}$$

Covariance matrix of the first class

$$S_1 = \sum_{x \in w_1} (x - \mu_1)(x - \mu_1)^T$$

$$= \left[ \begin{pmatrix} 4 \\ 2 \end{pmatrix} - \begin{pmatrix} 3 \\ 3.8 \end{pmatrix} + \begin{pmatrix} 2 \\ 4 \end{pmatrix} - \begin{pmatrix} 3 \\ 3.8 \end{pmatrix} + \begin{pmatrix} 2 \\ 3 \end{pmatrix} - \begin{pmatrix} 3 \\ 3.8 \end{pmatrix} + \begin{pmatrix} 3 \\ 6 \end{pmatrix} - \begin{pmatrix} 3 \\ 3.8 \end{pmatrix} \right. \\ \left. + \begin{pmatrix} 4 \\ 4 \end{pmatrix} - \begin{pmatrix} 3 \\ 3.8 \end{pmatrix} \right]$$

$$= \left[ \begin{pmatrix} 1 \\ -1.8 \end{pmatrix} (1 - 1.8) + \begin{pmatrix} -1 \\ 0.2 \end{pmatrix} (-1 - 0.2) + \begin{pmatrix} -1 \\ -0.8 \end{pmatrix} (-1 - 0.8) \right. \\ \left. \begin{pmatrix} 0 \\ 2.2 \end{pmatrix} (0 - 2.2) + \begin{pmatrix} 1 \\ 0.2 \end{pmatrix} (1 - 0.2) \right]$$

$$= \begin{bmatrix} 1 & -1.8 \\ -1.8 & 3.24 \end{bmatrix} + \begin{bmatrix} 1 & -0.2 \\ -0.2 & 0.04 \end{bmatrix} + \begin{bmatrix} 1 & 0.8 \\ 0.8 & 0.64 \end{bmatrix}$$

$$+ \begin{bmatrix} 0 & 0 \\ 0 & 4.84 \end{bmatrix} + \begin{bmatrix} 1 & 0.2 \\ 0.2 & 0.04 \end{bmatrix}$$

$$= \begin{bmatrix} 4 & -1 \\ -1 & 8.8 \end{bmatrix}$$

covariance matrix of the second class

$$S_2 = \sum_{x_i \in \omega_2} (x_2 - \mu_2)(x_2 - \mu_2)^T$$

$$= \left[ \begin{pmatrix} 9 \\ 10 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix} + \begin{pmatrix} 6 \\ 8 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix} + \begin{pmatrix} 9 \\ 5 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix} + \begin{pmatrix} 8 \\ 7 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix} \right]$$

$$= \left[ \begin{pmatrix} 0.6 \\ 2.4 \end{pmatrix} (0.6 \ 2.4) + \begin{pmatrix} -2.4 \\ 0.4 \end{pmatrix} (-2.4 \ 0.4) + \begin{pmatrix} 0.6 \\ -2.6 \end{pmatrix} (0.6 \ -2.6) \right. \\ \left. - 0.4 (-0.4 \ -0.6) + \begin{pmatrix} 1.6 \\ 0.4 \end{pmatrix} (1.6 \ 0.4) \right]$$

$$= \begin{bmatrix} 0.36 & 1.44 \\ 1.44 & 5.76 \end{bmatrix} + \begin{bmatrix} 5.76 & -0.96 \\ -0.96 & 0.16 \end{bmatrix} + \begin{bmatrix} 0.36 & -1.56 \\ -1.56 & 6.76 \end{bmatrix}$$

$$+ \begin{bmatrix} 0.16 & 0.24 \\ 0.24 & 0.36 \end{bmatrix} + \begin{bmatrix} 2.56 & 0.64 \\ 0.64 & 0.16 \end{bmatrix}$$

$$= \begin{bmatrix} 9.2 & -0.2 \\ -0.2 & 13.2 \end{bmatrix}$$

Within-class scatter matrix

$$S_W = S_1 + S_2 = \begin{bmatrix} 4 & -1 \\ -1 & 8.8 \end{bmatrix} + \begin{bmatrix} 9.2 & -0.2 \\ -0.2 & 13.2 \end{bmatrix}$$

$$= \begin{bmatrix} 13.2 & -1.2 \\ -1.2 & 22 \end{bmatrix}$$

Between-class scatter matrix :-

$$S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$$

$$= \left[ \begin{pmatrix} 5 \\ 3.8 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix} \right] \left[ \begin{pmatrix} 5 \\ 3.8 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix} \right]^T$$

$$= \begin{pmatrix} -5.4 \\ -3.8 \end{pmatrix} \begin{pmatrix} -5.4 & -3.8 \end{pmatrix}$$

$$= \begin{bmatrix} 29.16 & 20.52 \\ 20.52 & 14.44 \end{bmatrix}$$

The LDA projection is then obtained as the solution of the generalized eigen value problem.

$$S_w^{-1} S_B w = \lambda w$$

$$\Rightarrow |S_w^{-1} S_B - \lambda I| = 0$$

$$\Rightarrow \begin{vmatrix} 13.2 & -1.2 \\ -1.2 & 2.2 \end{vmatrix}^{-1} \begin{bmatrix} 29.16 & 20.52 \\ 20.52 & 14.44 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = 0$$

$$\Rightarrow \begin{bmatrix} 0.08 & 0.004 \\ 0.004 & 0.04 \end{bmatrix} \begin{bmatrix} 29.16 & 20.52 \\ 20.52 & 14.44 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = 0$$

$$\begin{bmatrix} 3.154 & 2.219 \\ 11.016 & 7.752 \end{bmatrix} + \begin{bmatrix} -\lambda & 0 \\ 0 & -\lambda \end{bmatrix} = 0, \begin{bmatrix} 3.154 & 1.699 \\ 1.143 & 0.804 \end{bmatrix}$$

$$\begin{bmatrix} 3.154 - \lambda & 2.219 \\ 11.016 & 7.752 - \lambda \end{bmatrix} = 0, \begin{bmatrix} 3.154 - \lambda & 1.699 \\ 1.143 & 0.804 - \lambda \end{bmatrix}$$

$$\begin{aligned}
 & (3.154 - \lambda)(7.752 - \lambda) - 24 \cdot 4.45 = 0 \\
 \Rightarrow & (3.154 - \lambda)(0.804 - \lambda) - 1.942 = 0 \\
 \Rightarrow & 24 \cdot 4.45 - 3.154\lambda - 7.752\lambda + \lambda^2 - 24 \cdot 4.45 = 0 \\
 \Rightarrow & 2.536 - 3.154\lambda - 0.804\lambda + \lambda^2 - 1.942 = 0 \\
 \Rightarrow & \lambda^2 - 10.906\lambda + 0.594 = 0 \\
 \Rightarrow & \lambda_1 = 0, \lambda_2 = 10.906, \lambda_1 = 2.3053, \lambda_2 = 0.7448
 \end{aligned}$$

$$w^* = S_w^{-1} (\mu_1 - \mu_2)$$

$$\begin{aligned}
 & = \begin{bmatrix} 13.2 & -1.2 \\ -1.2 & 2.2 \end{bmatrix}^{-1} \left[ \begin{pmatrix} 9 \\ 3.8 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix} \right] \\
 & = \begin{bmatrix} 0.08 & 0.04 \\ 0.04 & 0.48 \end{bmatrix} \begin{bmatrix} -5.4 \\ -3.8 \end{bmatrix} \quad \frac{1}{288.96} \begin{bmatrix} 22.2 & 1.2 \\ 1.2 & 13.2 \end{bmatrix} \\
 & = \begin{bmatrix} -0.584 \\ -2.04 \end{bmatrix} \quad \begin{bmatrix} 0.08 & 0.04 \\ 0.04 & 0 \end{bmatrix} \\
 & \quad \begin{bmatrix} 0.08 & 0.004 \\ 0.004 & 0.05 \end{bmatrix}
 \end{aligned}$$

Logistic Regression:

The Logistic model:

Let us assume that the outcome has two classes.  
 Logistic regression starts with different model setup than linear regression: instead of modeling  $y$  as a function of  $x$  directly, we model the probability that  $y$  is equal to class 1, given  $x$

The logistic model is

$$\begin{bmatrix} 22.86 & 1.70 \\ 1.14 & 0.804 \end{bmatrix}$$

$$\pi(x) = \frac{\exp(\beta^T x)}{1 + \exp(\beta^T x)} \quad (1)$$

$$= \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

$$= \frac{1}{1 + e^{-\beta_0 - \beta_1 x}}$$

The function on the RHS above is called the sigmoid of  $\beta^T x$ .

By rearranging, the eqn (1) becomes,

$$\log \left( \frac{p(x)}{1-p(x)} \right) = \beta^T x$$

the LHS above is called the log odds (or) logit of  $p(x)$ , and is written as logit  $p(x)$ .

$$\text{In general, logit } (a) = \log \left( \frac{a}{1-a} \right)$$

Interpreting coefficients:

To interpret the role of the coefficients

$$\beta = (\beta_1, \dots, \beta_p) \in \mathbb{R}^p \text{ in (2)}$$

$$\frac{p(x)}{1-p(x)} = e^{\beta^T x} = e^{\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p}$$

$$\log \left( \frac{p(x)}{1-p(x)} \right) = \beta^T x = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

A transformation of  $\pi(x)$  that is central to our study of logistic regression is the logit transformation. This transformation is defined, in terms of  $\pi(x)$ , as:

$$g(x) = \ln \left[ \frac{\pi(x)}{1-\pi(x)} \right] = \beta_0 + \beta_1 x$$

The importance of this transformation is that  $g(x)$  has many of the desirable properties of a linear regression model. The logit,  $g(x)$ , is linear in its parameters, may be continuous, and may range from

$\rightarrow \infty$  to  $+\infty$ , depending on the range of  $x$ .

### Fitting the Logistic Regression model:

Suppose we have a sample of  $n$  independent observations of the pair  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$  where  $y_i$  denotes the value of a dichotomous outcome variable and  $x_i$  is the value of the independent variable for the  $i^{\text{th}}$  subject.

Furthermore, assume that the outcome variable has been coded as 0 or 1, representing the absence or the presence of the characteristic, respectively.

To fit the logistic regression model in (1)

to a set of data requires that we estimate the values of  $\beta_0$  and  $\beta_1$ , the unknown parameters.

In linear regression, the method used most often for estimating unknown parameters is least squares. The general method of estimation that leads to the least squares function under the linear regression model is called maximum likelihood.

This method will provide the foundation for our approach to estimation with the logistic regression model. In a very general sense the method of maximum likelihood yields values for the unknown parameters which maximize the probability of obtaining the observed set of data.

In order to apply this method we must first construct a function, called the likelihood function.

This function expresses the probability of the observed data as a function of the unknown parameters. The maximum likelihood estimators of these parameters are chosen to be those values that maximize this function. Thus, the resulting estimators are those which agree most closely with the observed data.

If  $y$  is coded as 0 or 1 then the expression for  $\pi(x)$  given in (1) provides (for an arbitrary value of  $\beta = (\beta_0, \beta_1)$ , the vector of parameters) the conditional probability that  $y$  is equal to 1 given  $x$ . This will be denoted as  $P(y=1|x)$ . It follows that the quantity  $1 - \pi(x)$  gives the conditional probability that  $y$  is equal to zero given  $x$ ,  $P(y=0|x)$ .

Thus, for those pairs  $(x_i, y_i)$ , where  $y_i=1$ , the contribution to the likelihood function is  $\pi(x_i)$ , and for those pairs where  $y_i=0$ , the contribution to the likelihood function is  $1 - \pi(x_i)$ , where the quantity  $\pi(x_i)$  denotes the value of  $\pi(x)$  computed at  $x_i$ . A convenient way to express the contribution to the likelihood function for the pair  $(x_i, y_i)$  is through the expression

$$\pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i} \quad (2)$$

Since the observations are assumed to be independent, the likelihood function is obtained as the product of the terms given in expression (2) as follows:

$$l(\beta) = \prod_{i=1}^n \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i} \quad (3)$$

The principle of maximum likelihood states that we use as our estimate of  $\beta$  the value which maximizes the expression in (3). However, it is easier mathematically to work with the log of equation (3). This expression, the log likelihood is defined as

$$L(\beta) = \ln [l(\beta)] = \sum_{i=1}^n \{ y_i \ln [\pi(x_i)] + (1-y_i) \ln [1 - \pi(x_i)] \} \quad (4)$$

To find the value of  $\beta$  that maximizes  $L(\beta)$  we differentiate  $L(\beta)$  with respect to  $\beta_0$  and  $\beta_1$  and set the resulting expressions equal to zero. These equations, known as the likelihood equations, are:

$$\sum [y_i - \pi(x_i)] = 0 \quad (5)$$

$$\sum x_i [y_i - \pi(x_i)] = 0 \quad (6)$$

For logistic regression the expressions in (5) + (6) are nonlinear in  $\beta_0$  and  $\beta_1$ , and thus require special methods for their solution. In particular, they show that the solution to (5) + (6) may be obtained using an iterative weighted least squares procedure.

The value of  $\beta$  given by the solution to (5) and (6) is called the maximum likelihood estimate and will be denoted as  $\hat{\beta}$ . In general,

The use of the symbol ' $\hat{\cdot}$ ' denotes the maximum likelihood estimate of the respective quantity.

### The multiple Logistic Regression model:-

Consider a collection of  $p$  independent variables denoted by the vector  $x = (x_1, x_2, \dots, x_p)$ . For the moment we will assume that each of these variables is at least interval scale. Let the conditional probability that the outcome is present be denoted by  $P(Y=1 | x) = \pi(x)$ . The logit of the multiple logistic regression model is given by the equation

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p \quad (1)$$

in which case the logistic regression model is

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}}$$

### Problems :-

- For the customer service data, the proportion of customers who would recommend the service in the sample of customers is  $\hat{P} = 0.84$ , so the proportion of customers who would not recommend the service department.

### Solution :-

#### Note :-

Logistic regression work with odds rather than proportions. The odds are simply the ratio of

the proportions for the two possible outcomes. If  $\hat{p}$  is the proportion for one outcome, then  $1 - \hat{p}$  is the proportion for the second outcome:

$$\text{odds} = \frac{\hat{p}}{1 - \hat{p}}$$

The odds of recommending the service department

are  $\text{odds} = \frac{\hat{p}}{1 - \hat{p}}$

$$= \frac{0.84}{1 - 0.84} = 5.25$$

Q) The sample proportion of women who are Instagram users is given as 61.08%, and the sample proportion for men is 43.98%.

Solution:-

Another way to analyze these data is to use logistic regression. The explanatory variable is gender, a categorical variable.

$$x = \begin{cases} 1 & \text{if the person is a woman} \\ 0 & \text{if the person is a man} \end{cases}$$

The response variable is the proportion of Instagram users.

For women,

$$\text{odds} = \frac{\hat{p}}{1 - \hat{p}}$$

$$= \frac{0.6108}{1 - 0.6108} = 1.5694$$

for men,

$$\text{odds} = \frac{\hat{P}}{1-\hat{P}}$$
$$= \frac{0.4398}{1-0.4398} = 0.7851$$

The logistic regression solution is to transform the odds  $\frac{P}{1-P}$  using the natural logarithm.

We use the term log odds or logit for this transformation.

for women,

$$y = \log(\text{odds}) = \log(1.5694) = 0.4507$$

For men,

$$y = \log(\text{odds}) = \log(0.7851) = -0.2419$$

For our Instagram example, there are  $n=1069$  young persons in the sample. The explanatory variable is gender, which we have coded using an indicator variable with values  $x=1$  for women and  $x=0$  for men.

$$\log \left( \frac{P_{\text{women}}}{1-P_{\text{women}}} \right) = \beta_0 + \beta_1 \quad (\text{for women})$$

$$\log \left( \frac{P_{\text{men}}}{1-P_{\text{men}}} \right) = \beta_0 \quad (\text{for men})$$

Note, that there is a  $\beta_1$  term in the equation for women because  $x=1$ , but it is missing in the eqn for men because  $x=0$ ,

*Fitting and interpreting the logistic regression model.*

$$\log \left( \frac{\hat{P}_{\text{women}}}{1 - \hat{P}_{\text{women}}} \right) = 0.4507 \quad (\text{for women})$$

$$\log \left( \frac{\hat{P}_{\text{men}}}{1 - \hat{P}_{\text{men}}} \right) = -0.2419 \quad (\text{for men})$$

To find estimates  $b_0$  and  $b_1$ ,

$$b_0 = -0.2419 \quad (1)$$

$$b_0 + b_1 = 0.4507$$

$$b_1 = 0.4507 + 0.2419$$

$$= 0.6926$$

$\therefore$  The fitted logistic regression model is

$$\log(\text{odds}) = -0.2419 + 0.6926x.$$

The slope in this logistic regression model is the difference between the log odds for men and the log odds for women.

$$\frac{\text{odds}_{\text{women}}}{\text{odds}_{\text{men}}} = e^{0.6926} \\ = 1.999.$$

## PRINCIPAL COMPONENT ANALYSIS

Principal component analysis, or PCA, is a dimensionality reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

Reducing the number of variables of a data set naturally comes at the expense of accuracy, but the trick in dimensionality reduction is to trade a little accuracy for simplicity. Because smaller data sets are easier to explore and visualize and make analyzing data much easier and faster for machine learning algorithms without extraneous variable to process.

The idea of principal component analysis is to reduce the number of variables of a data set, while preserving as much information as possible.

### Geometric base of principal components :-

Principal component analysis deals with a single sample of  $n$  observation vectors  $y_1, y_2, \dots, y_n$  that form a swarm of points in a  $p$ -dimensional space. Principal component analysis can be applied to any distribution of  $y$ , but it will be easier to visualize geometrically if the swarm of points is ellipsoidal.

If the variables  $y_1, y_2, \dots, y_p$  in  $y$  are correlated

Step 1: Standardization:-

The aim of this step is to standardize the range of the continuous initial variables so that each one of them contributes equally to the analysis.

Mathematically, this can be done by subtracting the mean and dividing by the standard deviation for each value of each variable.

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$$

Once the standardization is done, all the variables will be transformed to the same scale.

Step 2: Covariance matrix computation:-

The aim of this step is to understand how the variables of the input data set are varying from the mean with respect to each other. Because, variables are highly correlated in such a way that they contain redundant information.

The covariance matrix  $P \times P$  symmetric matrix (where  $P$  is the number of dimensions) that has as entries the covariances associated with all possible pairs of the initial variables. For example, for a 3-dimensional data set with 3 variables  $x, y$  and  $z$ , the covariance matrix is a  $3 \times 3$  matrix of the form:

$$\begin{bmatrix} \text{cov}(x,x) & \text{cov}(x,y) & \text{cov}(x,z) \\ \text{cov}(y,x) & \text{cov}(y,y) & \text{cov}(y,z) \\ \text{cov}(z,x) & \text{cov}(z,y) & \text{cov}(z,z) \end{bmatrix}$$

Since the covariance of a variable with itself is its ( $\text{cov}(a,a) = \text{Var}(a)$ ) in the main diagonal.

Step 3:-

Compute the eigenvectors and eigenvalues of the covariance matrix to identify the principal components.

Eigenvectors and eigenvalues are the linear algebra concepts that we need to compute from the covariance matrix in order to determine the principal components of the data.

#### POPULATION PRINCIPAL COMPONENTS:-

Algebraically, principal components are particular linear combinations of the  $p$  random variables  $x_1, x_2, \dots, x_p$ . Geometrically, these linear combinations represent the selection of a new coordinate system obtained by rotating the original system with  $x_1, x_2, \dots, x_p$  as the coordinate axes.

Let the random vector  $x^t = \{x_1, x_2, \dots, x_p\}$  have the covariance matrix  $S$  with eigenvalues  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ .

Consider the linear combinations

$$y_1 = a_1' x = a_{11} x_1 + a_{12} x_2 + \dots + a_{1p} x_p$$

$$y_2 = a_2' x = a_{21} x_1 + a_{22} x_2 + \dots + a_{2p} x_p \quad (1)$$

$$\vdots$$

$$y_p = a_p' x = a_{p1} x_1 + a_{p2} x_2 + \dots + a_{pp} x_p$$

we obtain,

$$\text{Var}(y_i) = a_i' \Sigma a_i \quad i=1, 2, \dots, p \quad (2)$$

$$\text{cov}(y_i, y_k) = a_i' \Sigma a_k \quad i, k = 1, 2, \dots, p \quad (3)$$

The principal components are those uncorrelated linear combinations  $y_1, y_2, \dots, y_p$  whose variances in (2) are as large as possible.

The first principal component is the linear combination with maximum variance. That is, it maximizes  $\text{Var}(y_1) = a_1' \Sigma a_1$ . It is clear that  $\text{Var}(y_1) = a_1' \Sigma a_1$  can be increased by multiplying any  $a_1$  by some constant. To eliminate this indeterminacy, it is convenient to restrict attention to coefficient vectors of unit length. We therefore define

First principal component = linear combination  $a_1' x$   
that maximizes  $\text{Var}(a_1' x)$  subject to  $a_1' a_1 = 1$

Second principal component = linear combination  $a_2' x$   
that maximizes  $\text{Var}(a_2' x)$  subject to  $a_2' a_2 = 1$   
and  $\text{cov}(a_1' x, a_2' x) = 0$

At the  $i^{\text{th}}$  component,

$i^{\text{th}}$  principal component = linear combination  $a_i'x$  that maximizes  $\text{Var}(a_i'x)$  subject to  $a_i'a_i = 1$  and  $\text{Cov}(a_i'x, a_k'x) = 0$  for  $k < i$ .

Problem:-

calculating the population principal components suppose the random variables  $x_1, x_2$  and  $x_3$  have the covariance matrix

$$\Sigma = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix} \quad \begin{bmatrix} \text{Var}(x_1) & \text{Cov}(x_1, x_2) & \text{Cov}(x_1, x_3) \\ \text{Cov}(x_1, x_2) & \text{Var}(x_2) & \text{Cov}(x_2, x_3) \\ \text{Cov}(x_1, x_3) & \text{Cov}(x_2, x_3) & \text{Var}(x_3) \end{bmatrix}$$

Solution:-

It may be verified that the eigenvalue-eigenvector pairs are

$$\lambda_1 = 5.83 \quad e_1' = [0.383 \ -0.924 \ 0]$$

$$\lambda_2 = 2 \quad e_2' = [0 \ 0 \ 1]$$

$$\lambda_3 = 0.17 \quad e_3' = [0.924 \ 0.383 \ 0]$$

Therefore, the principal components become

$$Y_1 = e_1' x = 0.383 x_1 - 0.924 x_2$$

$$Y_2 = e_2' x = x_3$$

$$Y_3 = e_3' x = 0.924 x_1 + 0.383 x_2$$

The variable  $x_3$  is one of the principal components, because it is uncorrelated with the other two variables.

$$\text{Var}(Y_i) = e_i^T \Sigma e_i = \lambda_i$$

$$\text{Cov}(Y_i, Y_k) = e_i^T \Sigma e_k = 0 \quad i \neq k$$

$$\begin{aligned}\text{Var}(Y_1) &= \text{Var}(0.383X_1 - 0.924X_2) \\ &= (0.383)^2 \text{Var}X_1 + (-0.924)^2 \text{Var}X_2 \\ &\quad + 2(0.383)(-0.924) \text{Cov}(X_1, X_2)\end{aligned}$$

$$\text{Var } X_1 = 1, \text{Var } X_2 = 5$$

$$\begin{aligned}\text{Cov}(X_1, X_2) &= -2 \\ &= (0.383)^2(1) + (-0.924)^2(5) + 2(0.383)(-0.924)(-2) \\ &= 5.83 = \lambda_1\end{aligned}$$

$$\begin{aligned}\text{Cov}(Y_1, Y_2) &= \text{Cov}(0.383X_1 - 0.924X_2, X_3) \\ &= 0.383 \text{Cov}(X_1, X_3) - 0.924 \text{Cov}(X_2, X_3) \\ &= 0.383(0) - 0.924(0) = 0.\end{aligned}$$

$$\text{Total population variance} = \sigma_{11} + \sigma_{22} + \dots + \sigma_{pp}$$

$$= \lambda_1 + \lambda_2 + \dots + p\lambda_p$$

$$\sigma_{11} + \sigma_{22} + \sigma_{33}$$

$$\Rightarrow 1 + 5 + 2 = 8$$

$$\Rightarrow \lambda_1 + \lambda_2 + \lambda_3 = 5.83 + 2 + 0.17 = 8$$

$$\begin{aligned}\text{Proportion of total population variance due to } k^{\text{th}} \text{ principal component} &= \frac{\lambda_k}{\lambda_1 + \lambda_2 + \dots + \lambda_p} = k = 1, 2, \dots, p\end{aligned}$$

$$\text{First principal component} = \frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3}$$

$$= \frac{5.83}{8} = 0.73$$

$$\text{Proportion of first two components} = \frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \lambda_3}$$

$$= \frac{5.83 + 2}{8}$$

$$= 0.98$$

Correlation coefficients  $\rho_{Y_i, X_k}$

$$\rho_{Y_i, X_k} = \frac{e_{ik} \sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \quad i, k = 1, 2, \dots, p$$

$$\rho_{Y_1, X_1} = \frac{e_{11} \sqrt{\lambda_1}}{\sqrt{\sigma_{11}}} = \frac{0.883 \sqrt{5.83}}{\sqrt{5}} = 0.925$$

$$\rho_{Y_1, X_2} = \frac{e_{12} \sqrt{\lambda_1}}{\sqrt{\sigma_{22}}} = \frac{-0.924 \sqrt{5.83}}{\sqrt{5}} = -0.998.$$

Notice here that the variable  $X_2$ , with coefficient  $-0.924$ , receives the greatest weight in the component  $Y_1$ .

Since, both coefficients are reasonably large and they have opposite signs. Finally,

$$\rho_{Y_2, X_1} = \rho_{Y_2, X_2} = 0, \quad \rho_{Y_2, X_3} = \frac{\sqrt{\lambda_2}}{\sqrt{\sigma_{33}}} = 1.$$

The remaining correlations can be neglected, since the third component is unimportant.

Principal components obtained from standardized variables :-

Principal components may also be obtained for the standardized variables.

$$z_1 = \frac{(x_1 - \mu_1)}{\sqrt{\sigma_{11}}}$$

$$z_2 = \frac{(x_2 - \mu_2)}{\sqrt{\sigma_{22}}}$$

$$z_p = \frac{(x_p - \mu_p)}{\sqrt{\sigma_{pp}}}$$

calculate the population principal components  
2) Consider the covariance matrix  $\Sigma = \begin{bmatrix} 1 & 4 \\ 4 & 100 \end{bmatrix}$

and the derived correlation matrix

$$\delta = \begin{bmatrix} 1 & 0.4 \\ 0.4 & 1 \end{bmatrix}$$

Solution:-

The eigenvalue - eigenvector pairs from  $\Sigma$  are

$$\lambda_1 = 100.16, \quad e_1' = [0.040 \quad 0.999]$$

$$\lambda_2 = 0.84, \quad e_2' = [0.999 \quad -0.040]$$

The eigenvalue - eigenvector pairs from  $\delta$  are

$$\lambda_1 = 1.4, \quad e_1' = [0.707 \quad 0.707]$$

$$\lambda_2 = 0.6, \quad e_2' = [0.707 \quad -0.707]$$

The respective principal components become

$$\mathcal{L} : Y_1 = 0.040 X_1 + 0.999 X_2$$

$$Y_2 = 0.999 X_1 - 0.040 X_2$$

and  $\mathcal{S} : Y_1 = 0.707 Z_1 + 0.707 Z_2$

$$= 0.707 \left( \frac{X_1 - \bar{X}_1}{\sqrt{1}} \right) + 0.707 \left( \frac{X_2 - \bar{X}_2}{\sqrt{10}} \right)$$

$$= 0.707 (X_1 - \bar{X}_1) + 0.0707 (X_2 - \bar{X}_2)$$

$$Y_2 = 0.707 Z_1 - 0.707 Z_2$$

$$= 0.707 \left( \frac{X_1 - \bar{X}_1}{\sqrt{1}} \right) - 0.707 \left( \frac{X_2 - \bar{X}_2}{\sqrt{10}} \right)$$

$$= 0.707 (X_1 - \bar{X}_1) - 0.0707 (X_2 - \bar{X}_2)$$

Because of its large variance,  $X_2$  completely dominates the first principal component determined from  $\mathcal{L}$ .

The first principal component explains a proportion

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{100.16}{101} = 0.992$$

of the total population variance.

When the variables  $X_1$  and  $X_2$  are standardized

$$s_{Y_1, Z_1} = e_{11} \sqrt{\lambda_1} = 0.707 \sqrt{1.14} = 0.837$$

$$\text{and } \delta_{4,122} = e_{21} \sqrt{\lambda_1} = 0.707 \sqrt{1.4} = 0.837$$

In this case, the first principal component explains a proportion

$$\frac{\lambda_1}{p} = \frac{1.4}{2} = 0.7$$

of the total (standardized) population variance

Rules to retain number of principal components using sceee plot:

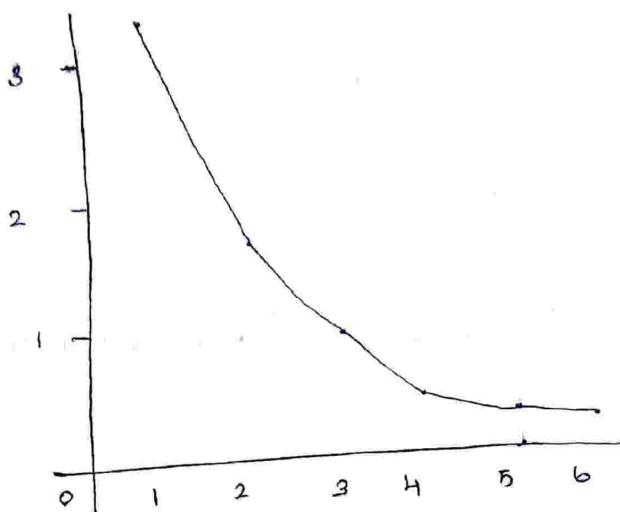
In every application, a decision must be made on how many principal components should be retained in order to effectively summarize the data. The following guidelines have been proposed.

- 1) Retain sufficient components to account for a specified percentage of the total variance, say 80%.
- 2) Retain the components whose eigenvalues are greater than the average of the eigenvalues,

$$\sum_{i=1}^p \lambda_i / p .$$

For a correlation matrix, this average is 1.

- 3) Use the sceee plot, a plot of  $\lambda_i$  versus  $i$ , and look for a natural break between the large eigenvalues and the small eigenvalues.
- 4) Test the significance of the large components, i.e., the components corresponding to the large eigenvalues.



Scree plot.

A useful visual aid to determining an appropriate number of principal components is a scree plot. With the eigenvalues ordered from largest to smallest, a scree plot is a plot of  $\lambda_i$  versus  $i$  - the magnitude of an eigenvalue versus its number. To determine the appropriate number of components, we look for an elbow (bend) in the scree plot.

An elbow occurs in the plot in Figure about  $i=4$ . (i.e.) the eigenvalues after  $\lambda_3$  are all relatively small and about the same size.

### Graphing the principal components:

Plots of the principal components can reveal suspect observations, as well as provide checks on the assumption of normality. Since the ~~linear~~ principal components are linear combinations of the

original variables, it is not unreasonable to expect them to be nearly normal. It is often necessary to verify that the first few principal components are approximately normally distributed when they are to be used as the input data for additional analyses.

The last principal components can help pinpoint suspect observations. Each observation can be expressed as a linear combination

$$\begin{aligned} x_j &= (x_j' \hat{e}_1) \hat{e}_1 + (x_j' \hat{e}_2) \hat{e}_2 + \dots + (x_j' \hat{e}_p) \hat{e}_p \\ &= \hat{y}_{j1} \hat{e}_1 + \hat{y}_{j2} \hat{e}_2 + \dots + \hat{y}_{jp} \hat{e}_p \end{aligned}$$

of the complete set of eigenvectors  $\hat{e}_1, \hat{e}_2, \dots, \hat{e}_p$  of S. Thus, the magnitudes of the last principal components determine how well the first few fit the observations.

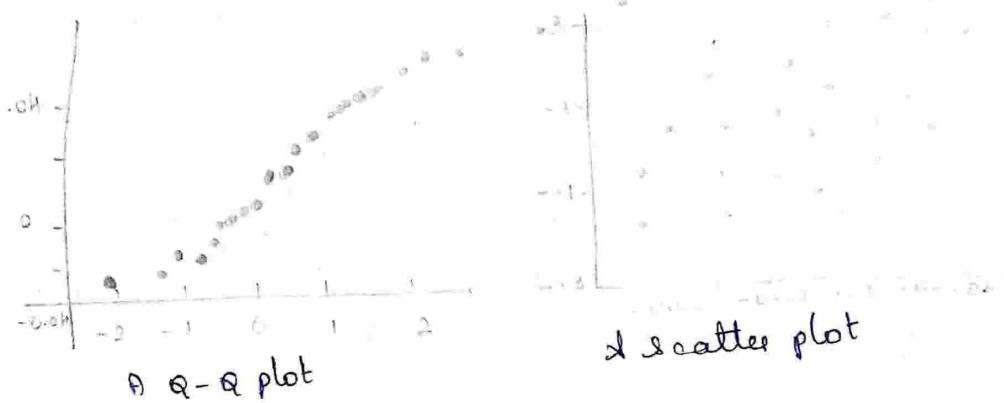
$$(i.e.) \hat{y}_{j1} \hat{e}_1 + \hat{y}_{j2} \hat{e}_2 + \dots + \hat{y}_{jq-1} \hat{e}_{q-1}$$

differs from  $x_j$  by  $\hat{y}_{jq} \hat{e}_q + \dots + \hat{y}_{jp} \hat{e}_p$ , the square of whose length is  $\hat{y}_{jq}^2 + \dots + \hat{y}_{jp}^2$ . Suspect observations will often be such that at least one of the coordinates  $\hat{y}_{jq}, \dots, \hat{y}_{jp}$  contributing to this squared length will be large.

The following statements summarize these ideas.

- 1) To help check the normal assumption, construct scatter diagrams for pairs of the first few principal components. Also, make Q-Q plots from the sample values generated by each principal component.

2) construct scatter diagrams and Q-Q plots for the last few principal components. These help identify suspect observations. A Q-Q plot is a probability plot, which is a graphical method for compactly comparing two probability distributions by plotting quantiles against each other.



### Lage sample Inferences:-

The eigenvalues and eigenvectors of the covariance (or correlation) matrix are the essence of a principal component analysis. The eigenvectors determine the directions of maximum variability, and the eigenvalues specify the variances. The quality of the principal component approximation must be made on the basis of the eigenvalue-eigenvector pairs  $(\hat{\lambda}_i, \hat{e}_i)$  extracted from  $S$  or  $R$ .

The large sample distribution theory for the eigenvalues  $\hat{\lambda}^t = [\hat{\lambda}_1, \dots, \hat{\lambda}_p]$  and eigenvectors  $\hat{e}_1, \hat{e}_2, \dots, \hat{e}_p$  of  $S$ :

1) Let  $\Lambda$  be the diagonal matrix of eigenvalues  $\lambda_1, \lambda_2, \dots, \lambda_p$  of  $S$ , then  $\sqrt{n}(\hat{\Lambda} - \Lambda)$  is approximately  $N_p(0, 2\Lambda^2)$ .

2) Let

$$\hat{e}_i = \lambda_i \sum_{k=1}^p \frac{\lambda_k}{(\lambda_k - \lambda_i)^2} e_k e_k'$$

then  $\sqrt{n}(\hat{e}_i - e_i)$  is approximately  $N_p(0, \Sigma_i)$ .

3) Each  $\hat{\lambda}_i$  is distributed independently of the elements of the associated  $\hat{e}_i$ .

The  $\hat{\lambda}_i$  are independently distributed. Moreover,  $\hat{\lambda}_i$  has an approximate  $N(\lambda_i, 2\lambda_i^2/n)$  distribution.

Using this normal distribution, we obtain

$$P[|\hat{\lambda}_i - \lambda_i| \leq z(\alpha/2) \lambda_i \sqrt{2/n}] = 1 - \alpha$$

$\Rightarrow$  Large sample  $100(1-\alpha)\%$  confidence interval for  $\lambda_i$  is thus provided by

$$\frac{\hat{\lambda}_i}{(1 + z(\alpha/2) \sqrt{2/n})} \leq \lambda_i \leq \frac{\hat{\lambda}_i}{(1 - z(\alpha/2) \sqrt{2/n})}$$

where  $z(\alpha/2)$  is the upper  $100(\alpha/2)^{\text{th}}$  percentile of a standard normal distribution.

Problems:-

1) Obtain 95% confidence interval for  $\lambda_1$ . where

$$\hat{\lambda}_1 = 0.0014, z(0.025) = 1.96 \text{ and } n = 103$$

Solution:-

$$\frac{0.0014}{(1 + z(0.025) \sqrt{2/103})} \leq \lambda_1 \leq \frac{0.0014}{(1 - z(0.025) \sqrt{2/103})}$$

$$\frac{0.0014}{(1 + 1.96 \sqrt{\frac{2}{103}})} \leq \lambda_1 \leq \frac{0.0014}{(1 - 1.96 \sqrt{\frac{2}{103}})}$$

$$0.0011 \leq \lambda_1 \leq 0.0019$$

Testing for the Equal correlation structure

The special correlation structure  $\text{cov}(x_i, x_k) = \rho$ , all  $i \neq k$ , is one important structure in which the eigenvalues of  $S$  are not distinct.

To test for this structure, let

$$H_0: \rho = \rho_0 = \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \dots & 1 \end{bmatrix}$$

$$\text{and } H_1: \rho \neq \rho_0$$

A test of  $H_0$  versus  $H_1$  may be based on a likelihood ratio statistic.

$$\bar{r}_k = \frac{1}{p-1} \sum_{\substack{i=1 \\ i \neq k}}^p r_{ik} \quad k=1, 2, \dots, p$$

$$\bar{r} = \frac{2}{p(p-1)} \sum_{i \neq k} \sum_{l \neq k} r_{ik}$$

$$\gamma_j = \frac{(p-1)^2 [1 - (1-\bar{r})^2]}{p - (p-2)(1-\bar{r})^2} ,$$

It is evident that  $\bar{r}_k$  is the average of the off-diagonal elements in the  $k^{th}$  column (or row) of  $R$  and  $\bar{r}$  is the overall average of the off-diagonal elements.

The large sample approximate  $\alpha$ -level test is to reject  $H_0$  in favour of  $H_1$ , if

$$T = \frac{(n-1)}{(1-\bar{r})^2} \left[ \sum_{i < k} \sum_{j \neq i} (r_{ik} - \bar{r})^2 - \gamma_j \sum_{k=1}^p (\bar{r}_k - \bar{r})^2 \right]$$

$$> \chi^2_{(p+1)(p-2)/2}(\alpha).$$

where  $\chi^2_{(p+1)(p-2)/2}(\alpha)$  is the upper  $(100\alpha)^{th}$  percentile of the chi-square distribution with  $(p+1)(p-2)/2$  d.f.

Problem:-

The sample correlation matrix constructed from the  $n=150$  post-birth weights of female mice is

$$R = \begin{bmatrix} 1.0 & 0.7501 & 0.6329 & 0.6363 \\ 0.7501 & 1.0 & 0.6925 & 0.7386 \\ 0.6329 & 0.6925 & 1.0 & 0.6625 \\ 0.6363 & 0.7386 & 0.6625 & 1.0 \end{bmatrix}$$

Solution:-

Let we use this correlation matrix to illustrate the large sample test.

Here  $p=4$  and we set

$$H_0 : f = f_0 = \begin{bmatrix} 1 & f & f & f \\ f & 1 & f & f \\ f & f & 1 & f \\ f & f & f & 1 \end{bmatrix}$$

$$H_1 : f \neq f_0$$

$$\bar{r}_k = \frac{1}{P-1} \sum_{\substack{i=1 \\ i \neq k}}^P r_{ik}, \quad k = 1, 2, \dots, P$$

$$\bar{r}_1 = \frac{1}{4-1} \sum_{i=1}^4 r_{i1}$$

$$= \frac{1}{3} \cancel{\sum} [r_{11} + r_{21} + r_{31} + r_{41}]$$

$$= \frac{1}{3} [0.7501 + 0.6329 + 0.6363]$$

$$= 0.6731$$

$$\bar{r}_2 = \frac{1}{4-1} \sum_{i=1}^4 r_{i2}$$

$$= \frac{1}{3} [r_{12} + r_{22} + r_{32} + r_{42}]$$

$$= \frac{1}{3} [0.7501 + 0.6925 + 0.7386]$$

$$= 0.7271$$

$$\bar{r}_3 = \frac{1}{4-1} \sum_{i=1}^4 r_{i3}$$

$$= \frac{1}{3} [r_{13} + r_{23} + r_{33} + r_{43}]$$

$$= \frac{1}{3} [0.6329 + 0.6925 + 0.6625]$$

$$= 0.6626$$

$$\bar{r}_H = \frac{1}{H-1} \sum_{i=1}^H r_{iH}$$

$$= \frac{1}{3} [r_{1H} + r_{2H} + r_{3H} + r_{4H}]$$

$$= \frac{1}{3} [0.6363 + 0.7386 + 0.6625] = 0.6791$$

$$\bar{r} = \frac{2}{P(P-1)} \sum_{i < k} \sum_{j < k} r_{ijk}$$

$$= \frac{2}{H(H-1)} \left[ \cancel{r_{21}} + \cancel{r_{31}} + r_{12} + r_{13} + r_{14} + r_{23} + r_{24} + r_{34} \right]$$

$$= \frac{2}{H(3)} \left[ 0.1501 + 0.6329 + 0.6363 + 0.6925 + 0.7386 + 0.6625 \right]$$

$$= 0.6855$$

$$\eta = \frac{(P-1)^2 [1 - (1-\bar{r})^2]}{P - (P-2) (1-\bar{r})^2}$$

$$= \frac{(H-1)^2 [1 - (1 - 0.6855)^2]}{H - (H-2) (1 - 0.6855)^2}$$

$$= \frac{3^2 [1 - (0.3145)^2]}{H - 2 (0.3145)^2}$$

$$= \frac{9 [1 - 0.0989]}{H - 2 (0.0989)} = \frac{8.1099}{3.8022} = 2.1329$$

$$\sum_{i < k} (r_{ik} - \bar{r})^2 = (r_{12} - \bar{r})^2 + (r_{13} - \bar{r})^2 + (r_{14} - \bar{r})^2 + (r_{23} - \bar{r})^2 + (r_{24} - \bar{r})^2 + (r_{34} - \bar{r})^2$$

$$\begin{aligned}
 &= (0.7501 - 0.6855)^2 + (0.6329 - 0.6855)^2 + (0.6363 - 0.6855)^2 \\
 &\quad + (0.6925 - 0.6855)^2 + (0.7386 - 0.6855)^2 + (0.6625 - 0.6855)^2 \\
 &= 0.0042 + 0.0028 + 0.0024 + 0.0004 + 0.0028 \\
 &\quad + 0.00053 \\
 &= 0.01277
 \end{aligned}$$

$$\begin{aligned}
 \sum_{k=1}^n (\bar{r}_k - \bar{r})^2 &= (\bar{r}_1 - \bar{r})^2 + (\bar{r}_2 - \bar{r})^2 + (\bar{r}_3 - \bar{r})^2 \\
 &\quad + (\bar{r}_4 - \bar{r})^2 \\
 &= (0.6731 - 0.6855)^2 + (0.727 - 0.6855)^2 + (0.6626 - 0.6855)^2 \\
 &\quad + (0.6791 - 0.6855)^2 \\
 &= 0.00245
 \end{aligned}$$

$$\begin{aligned}
 T &= \frac{(n-1)}{(1-\bar{r})^2} \left[ \sum_i \sum_{i < k} (\bar{r}_{ik} - \bar{r})^2 - \sum_{k=1}^n (\bar{r}_k - \bar{r})^2 \right] \\
 &= \frac{(150-1)}{(1-0.6855)^2} [0.01277 - (2.1329)(0.00245)] = 11.4
 \end{aligned}$$

$$\text{Since } (P+1)(P-2)/2 = 5(2)/2 = 5$$

$$\chi^2(0.05) = 11.07.$$

$\therefore H_0$  is accepted

**Oblique Rotation** :: (nonorthogonal)

The term oblique rotation refers to a transformation in which the axes do not remain perpendicular. Technically, the term oblique rotation is a misnomer, since rotation implies an orthogonal transformation that preserves distances.

Instead of the orthogonal transformation matrix  $T$ , an oblique rotation uses a general nonsingular transformation matrix  $\Omega$  to obtain  $f^* = \Omega' f$

$$\text{cov}(f^*) = \Omega' I \Omega = \Omega' \Omega \neq I.$$

Thus the new factors are correlated.

Since the distances and angles are not preserved, the communalities of  $f^*$  are different from those of  $f$ .

In other words, oblique Rotation case those allowing the factor to be correlated by

allows

allowing the angle between the axes to be less than  $90^\circ$  degree

The Methods which uses Oblique Rotation are Factor Analysis, namely,

g. Direct Oblimin

Varimax

Promax Methods

oblique rotation is an subset of orthogonal matrix which uses Factor correlation

matrix, structure matrix, pattern matrix and

Factor coefficient matrix for interpretation

For interpretation we usually prefer pattern

matrix rather than the structure matrix.

Uses :-

1. It seeks to express each variable in terms of a minimum number of factors preferably, a single factor.
2. Check on the orthogonality of the factors.

Properties (To Remember) :-

1. Factors are correlated
2. The distance and angles are not preserved.
3. If it produces a correlation matrix that is nearly diagonal then the factors are indeed orthogonal.

oblique loadings:- the oblique loadings give much clearer simple structure than the varimax loadings.

but the interpretation is essentially the same if we neglect loadings below .45 on the varimax rotation

\* Refer Text Book for example and graphical notation \*

### 13.5.3 Oblique Rotation

The term *oblique rotation* refers to a transformation in which the axes do not remain perpendicular. Technically, the term *oblique rotation* is a misnomer, since rotation implies an orthogonal transformation that preserves distances. A more accurate char-

Table 13.8. Varimax Rotated Loadings for the Seishu Data

Variables	Iterated Principal Factor Rotated Loadings				Maximum Likelihood Rotated Loadings			
	$f_1$	$f_2$	$f_3$	$f_4$	$f_1$	$f_2$	$f_3$	$f_4$
Taste	.16	-.01	.99	-.09	.16	-.00	.98	-.10
Odor	-.11	.14	.48	.14	-.07	.14	.49	.17
pH	.88	-.12	.02	-.13	.82	-.10	.08	-.15
Acidity 1	.26	-.09	.09	.54	.29	-.08	.11	.53
Acidity 2	.89	-.06	.10	.43	.91	-.06	.10	.39
Sake-meter	-.43	-.76	.01	.07	-.46	-.80	.04	.10
Reducing sugar	-.37	.76	.18	.03	-.37	.75	.20	.08
Total sugar	-.26	.92	.10	.25	-.27	.91	.11	.26
Alcohol	-.01	.25	.00	.80	-.00	.25	.01	.81
Formyl-nitrogen	.74	-.07	-.08	.20	.76	-.07	-.08	.22
Variance accounted for	2.62	2.12	1.27	1.27	2.61	2.14	1.29	1.28

acterization would be oblique *transformation*, but the term oblique rotation is well established in the literature.

Instead of the orthogonal transformation matrix  $T$  used in (13.16), (13.17), and (13.18), an oblique rotation uses a general nonsingular transformation matrix  $Q$  to obtain  $\mathbf{f}^* = Q'\mathbf{f}$ , and by (3.74),

$$\text{cov}(\mathbf{f}^*) = Q' \mathbf{I} Q = Q' Q \neq \mathbf{I}. \quad (13.50)$$

Thus the new factors are correlated. Since distances and angles are not preserved, the communalities for  $\mathbf{f}^*$  are different from those for  $\mathbf{f}$ . Some program packages report communalities obtained from the original loadings, rather than the oblique loadings.

When the axes are not required to be perpendicular, they can more easily pass through the major clusters of points in the loading space (assuming there are such clusters). For example, in Figure 13.4, we have plotted the varimax rotated loadings for two factors extracted from the sons data of Table 3.7 (see Example 13.5.3 at the end of this section). Oblique axes with an angle of  $38^\circ$  would pass much closer to the points, and the resulting loadings would be very close to 0 and 1. However, the interpretation would not change, since the same points (variables) would be associated with the oblique axes as with the orthogonal axes.

Various analytical methods for achieving oblique rotations have been proposed and are available in program packages. Typically, the output of one of these procedures includes a *pattern matrix*, a *structure matrix*, and a matrix of correlations among the oblique factors. For interpretation, we would usually prefer the pattern

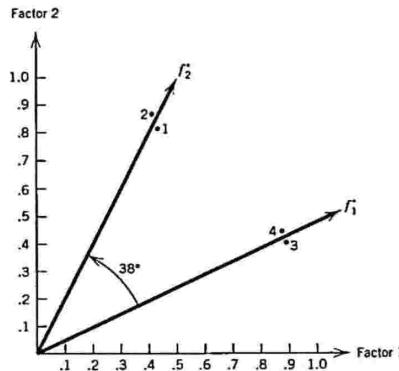


Figure 13.4. Orthogonal and oblique rotations for the sons data.

matrix rather than the structure matrix. The loadings in a row of the pattern matrix are the natural coordinates of the point (variable) on the oblique axes and serve as coefficients in the model relating the variable to the factors.

One use for an oblique rotation is to check on the orthogonality of the factors. The orthogonality in the original factors is imposed by the model and maintained by an orthogonal rotation. If an oblique rotation produces a correlation matrix that is nearly diagonal, we can be more confident that the factors are indeed orthogonal.

**Example 13.5.3.** The correlation matrix for the sons data of Table 3.7 is

$$\mathbf{R} = \begin{pmatrix} 1.000 & .735 & .711 & .704 \\ .735 & 1.000 & .693 & .709 \\ .711 & .693 & 1.000 & .839 \\ .704 & .709 & .839 & 1.000 \end{pmatrix}.$$

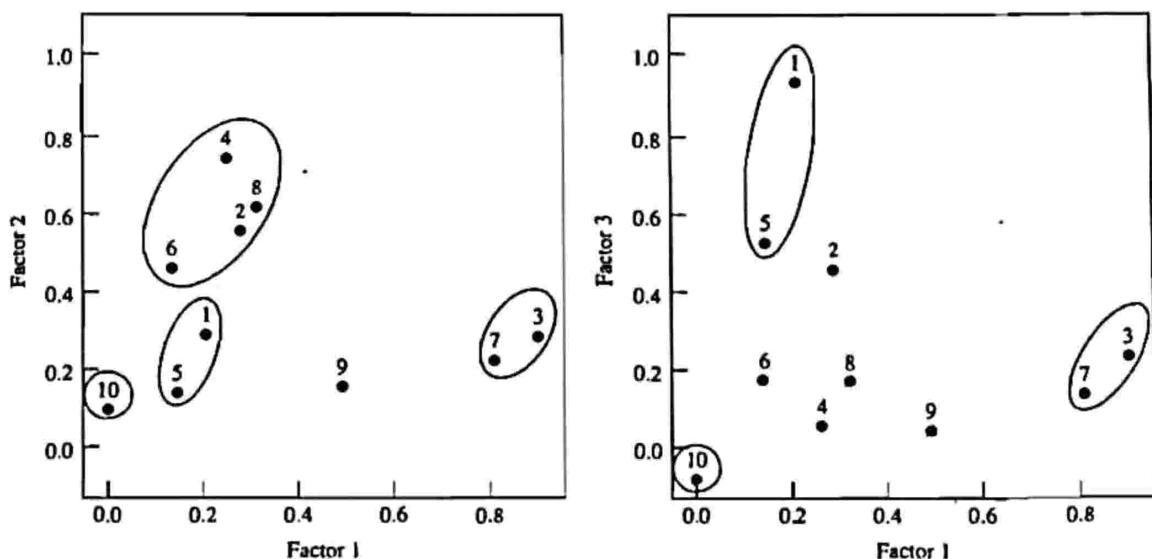
The varimax rotated loadings for two factors obtained by the principal component method are given in Table 13.9 and plotted in Figure 13.4. An analytical oblique rotation (Harris-Kaiser orthoblique method in SAS) produced oblique axes with an angle of  $38^\circ$ , the same as obtained by a graphical approach. The correlation between the two factors is .79 [obtained from  $\mathbf{Q}'\mathbf{Q}$  in (13.50)], which is related to the angle by (3.15),  $.79 = \cos 38^\circ$ . The pattern loadings are given in Table 13.9.

The oblique loadings give a much cleaner simple structure than the varimax loadings, but the interpretation is essentially the same if we neglect loadings below .45 on the varimax rotation.

In Figure 13.4, it is evident that a single factor would be adequate since the angle between axes is less than  $45^\circ$ . The suggestion to let  $m = 1$  is also supported by the first three criteria in Section 13.4: the eigenvalues of  $\mathbf{R}$  are 3.20, .38, .27, and .16. The first accounts for 80%; the second for an additional 9%. The large correlation, .79, between the two oblique factors constitutes additional evidence that a single-factor model would suffice here. In fact, the pattern in  $\mathbf{R}$  itself indicates the presence of only one factor. The four variables form only one cluster, since all are highly correlated. There are no small correlations between groupings of variables.  $\square$

**Table 13.9. Varimax and Orthoblique Loadings for the Sons Data**

Variable	Varimax Loadings		Orthoblique Pattern matrix	
	$f_1$	$f_2$	$f_1$	$f_2$
1	.42	.82	.03	.90
2	.40	.85	-.03	.96
3	.87	.41	.97	-.01
4	.86	.43	.95	.01



**Figure 9.3** Rotated maximum likelihood loadings for factor pairs (1, 2) and (1, 3)—decathlon data. (The numbers in the figures correspond to variables.)

often suggested after one views the estimated factor loadings and do not follow from our postulated model. Nevertheless, an oblique rotation is frequently a useful aid in factor analysis.

If we regard the  $m$  common factors as coordinate axes, the point with the  $m$  coordinates  $(\hat{e}_{i1}, \hat{e}_{i2}, \dots, \hat{e}_{im})$  represents the position of the  $i$ th variable in the *factor space*. Assuming that the variables are grouped into nonoverlapping clusters, an orthogonal rotation to a simple structure corresponds to a *rigid* rotation of the coordinate axes such that the axes, after rotation, pass as closely to the clusters as possible. An oblique rotation to a simple structure corresponds to a *nonrigid* rotation of the coordinate system such that the rotated axes (no longer perpendicular) pass (nearly) through the clusters. An oblique rotation seeks to express each variable in terms of a minimum number of factors—preferably, a single factor. Oblique rotations are discussed in several sources (see, for example, [6] or [10]) and will not be pursued in this book.

### ROTATION: FACTOR ANALYSIS

Factor loadings are unique only up to multiplication by an orthogonal matrix that rotates the loadings. The rotated loadings → preserve the essential properties of the original loadings.

→ Reproduce the covariance matrix

→ satisfy all basic assumptions

### Mathematical representation

$$\hat{\Lambda}^* = \hat{\Lambda}T \rightarrow \Lambda^* = \Lambda T$$

$\hat{\Lambda}$  - estimated loading matrix

T - orthogonal matrix

Since  $T^*T = I$  (Identity matrix), the rotated loadings provide the same estimate of covariance matrix as before.

If we consider covariance matrix

$$\Sigma = \Lambda\Lambda' + \Psi \Rightarrow S = \hat{\Lambda}^*\hat{\Lambda}^{*'} + \tilde{\Psi}$$

In place of loading matrix, we substitute rotated loading matrix  $\Sigma = S \cong \Lambda^*\Lambda^{*'} + \Psi = \Lambda T(\Lambda T')' + \Psi$

$$\begin{aligned} &= \Lambda\Lambda' + T' + \Psi \\ &= \Lambda\Lambda' + \Psi \end{aligned}$$

→ loadings in the  $i^{th}$  row of  $\Lambda$  → points in loading space corresponding to  $y_i$ .

A rotation in which every point is close to an axis, & where each variable loads highly on the factor corresponding to the axis and has small loadings on the remaining factors.

Such a state is called simple structure.

Two basic types of rotations are present  
→ Orthogonal  
→ Obligee

#### ORTHOGONAL ROTATION:

Rotation involving an orthogonal matrix is orthogonal rotation, the perpendicular axes are rotated rapidly and remain perpendicular

Properties are:

- Angles & distances are preserved
- Correlations are unchanged
- Basic configuration of points remain the same.
- Variance for each factor & corresponding proportion will change  
this is because when rows of  $\mathbf{A}$  are rotated the distance to the origin is unchanged

#### Graphical Representation:

If there are only 2 factors ( $m=2$ ) we can use graphical rotation  $\rightarrow$

Rows  $\mathbf{a}_i$  of  $\mathbf{A}$  are pairs of loading  $(\lambda_{i1}, \lambda_{i2})$  where  $i = 1, 2, \dots, p$ .

We choose angle  $\phi \rightarrow$  which axes are rotated  
the new rotated loadings  $(\lambda_{i1}^*, \lambda_{i2}^*)$   
are measured directly in graph & is calculated  
from  $\lambda^* = \lambda T$  using

$$T = \begin{pmatrix} \cos \phi & -\sin \phi \\ \sin \phi & \cos \phi \end{pmatrix}$$

## 13.5 ROTATION

### 13.5.1 Introduction

As noted in Section 13.2.2, the factor loadings (rows of  $\mathbf{A}$ ) in the population model are unique only up to multiplication by an orthogonal matrix that rotates the loadings. The rotated loadings preserve the essential properties of the original loadings; they reproduce the covariance matrix and satisfy all basic assumptions. The estimated loading matrix  $\hat{\mathbf{A}}$  can likewise be rotated to obtain  $\hat{\mathbf{A}}^* = \hat{\mathbf{A}}\mathbf{T}$ , where  $\mathbf{T}$  is orthogonal. Since  $\mathbf{T}\mathbf{T}' = \mathbf{I}$  by (2.102), the rotated loadings provide the same estimate of the covariance matrix as before:

$$\mathbf{S} \cong \hat{\mathbf{A}}^*\hat{\mathbf{A}}^{*'} + \hat{\Psi} = \hat{\mathbf{A}}\mathbf{T}\mathbf{T}'\hat{\mathbf{A}}' + \hat{\Psi} = \hat{\mathbf{A}}\hat{\mathbf{A}}' + \hat{\Psi}. \quad (13.48)$$

Geometrically, the loadings in the  $i$ th row of  $\hat{\mathbf{A}}$  constitute the coordinates of a point in the loading space corresponding to  $y_i$ . Rotation of the  $p$  points gives their coordinates with respect to new axes (factors) but otherwise leaves their basic geometric configuration intact. We hope to find a new frame of reference in which the factors are more interpretable. To this end, the goal of rotation is to place the axes close to as many points as possible. If there are clusters of points (corresponding to groupings of  $y$ 's), we seek to move the axes in order to pass through or near these clusters. This would associate each group of variables with a factor (axis) and make interpretation more objective. The resulting axes then represent the natural factors.

If we can achieve a rotation in which every point is close to an axis, then each variable loads highly on the factor corresponding to the axis and has small loadings on the remaining factors. In this case, there is no ambiguity. Such a happy state of affairs is called *simple structure*, and interpretation is greatly simplified. We merely observe which variables are associated with each factor, and the factor is defined or named accordingly.

In order to identify the natural groupings of variables, we seek a rotation to an interpretable pattern for the loadings, in which the variables load highly on only one factor. The number of factors on which a variable has moderate or high loadings is called the *complexity* of the variable. In the ideal situation referred to previously as simple structure, the variables all have a complexity of 1. In this case, the variables have been clearly clustered into groups corresponding to the factors.

We consider two basic types of rotation: *orthogonal* and *oblique*. The rotation in (13.48) involving an orthogonal matrix is an orthogonal rotation; the original perpendicular axes are rotated rigidly and remain perpendicular. In an orthogonal rotation, angles and distances are preserved, communalities are unchanged, and the basic configuration of the points remains the same. Only the reference axes differ. In an oblique “rotation” (transformation), the axes are not required to remain perpendicular and are thus free to pass closer to clusters of points.

In Sections 13.5.2 and 13.5.3, we discuss orthogonal and oblique rotations, followed by some guidelines for interpretation in Section 13.5.4.

### 13.5.2 Orthogonal Rotation

It was noted above in Section 13.5.1 that orthogonal rotations preserve communalities. This is because the rows of  $\hat{\mathbf{A}}$  are rotated, and the distance to the origin is unchanged, which, by (13.28), is the communality. However, the variance accounted for by each factor as given in (13.31) will change, as will the corresponding proportion in (13.32) or (13.33). The proportions due to the rotated loadings will not necessarily be in descending order.

In Sections 13.5.2a and 13.5.2b, we consider two approaches to orthogonal rotation.

#### 13.5.2a Graphical Approach

If there are only two factors ( $m = 2$ ), we can use a *graphical* rotation based on a visual inspection of a plot of factor loadings. In this case, the rows of  $\hat{\mathbf{A}}$  are pairs of

loadings,  $(\hat{\lambda}_{i1}, \hat{\lambda}_{i2})$ ,  $i = 1, 2, \dots, p$ , corresponding to  $y_1, y_2, \dots, y_p$ . We choose an angle  $\phi$  through which the axes can be rotated to move them closer to groupings of points. The new rotated loadings  $(\hat{\lambda}_{i1}^*, \hat{\lambda}_{i2}^*)$  can be measured directly on the graph as coordinates of the axes or calculated from  $\hat{\Lambda}^* = \hat{\Lambda}\mathbf{T}$  using

$$\mathbf{T} = \begin{pmatrix} \cos \phi & -\sin \phi \\ \sin \phi & \cos \phi \end{pmatrix}. \quad (13.49)$$

**Example 13.5.2a.** In Example 13.3.1, the initial factor loadings for the perception data did not provide an interpretation consistent with the two groupings of variables apparent in the pattern of correlations in  $\mathbf{R}$ . The five pairs of loadings  $(\hat{\lambda}_{i1}, \hat{\lambda}_{i2})$  corresponding to the five variables are plotted in Figure 13.3. An orthogonal rotation through  $-35^\circ$  would bring the axes (factors) closer to the two clusters of points (variables) identified in Example 13.3.1. With the rotation, each cluster of variables corresponds much more closely to a factor. Using  $\hat{\Lambda}$  from Example 13.3.1 and  $-35^\circ$  in  $\mathbf{T}$  as given in (13.49), we obtain the following rotated loadings:

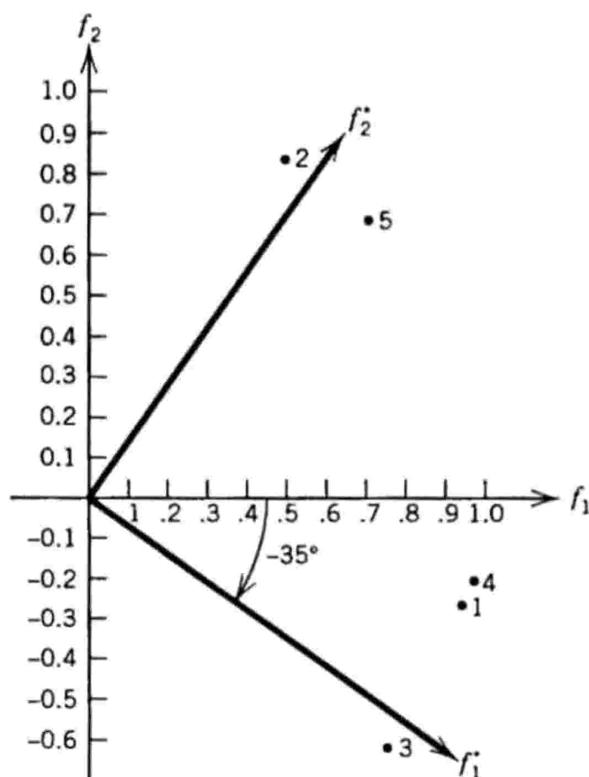


Figure 13.3. Plot of the two loadings for each of the five variables in the perception data of



**Figure 13.3.** Plot of the two loadings for each of the five variables in the perception data of Table 13.1.

$$\begin{aligned}\hat{\Lambda}^* &= \hat{\Lambda}T = \begin{pmatrix} .969 & -.231 \\ .519 & .807 \\ .785 & -.587 \\ .971 & -.210 \\ .704 & .667 \end{pmatrix} \begin{pmatrix} .819 & .574 \\ -.574 & .819 \end{pmatrix} \\ &= \begin{pmatrix} .927 & .367 \\ -.037 & .959 \\ .980 & -.031 \\ .916 & .385 \\ .194 & .950 \end{pmatrix}.\end{aligned}$$

In Table 13.6, we compare the rotated loadings in  $\hat{\Lambda}^*$  with the original loadings in  $\hat{\Lambda}$ .

The interpretation of the rotated loadings is clear. As indicated by the boldface loadings in Table 13.6, the first factor is associated with variables 1, 3, and 4: kind, happy, and likeable. The second factor is associated with the other two variables: intelligent and just. This same grouping of variables is indicated by the pattern in the correlation matrix in (13.35) and can also be seen in the two clusters of points in Figure 13.3. The first factor might be described as representing a person's perceived humanity or amiability, while the second involves more logical or rational practices.

Note that if the angle between the rotated axes is allowed to be less than  $90^\circ$  (an oblique rotation), the lower axis representing  $f_1^*$  could come closer to the points corresponding to variables 1 and 4 so that the coordinates on  $f_2^*$ , .367 and .385, could be reduced. However, the basic interpretation would not change; variables 1 and 4 would still be associated with  $f_1^*$ .  $\square$

**Table 13.6. Graphically Rotated Loadings for the Perception Data of Table 13.1**

Variables	Principal Component Loadings		Graphically Rotated Loadings		Communalities, $\hat{h}_i^2$
	$f_1$	$f_2$	$f_1$	$f_2$	
Kind	.969	-.231	<b>.927</b>	.367	.993
Intelligent	.519	.807	-.037	<b>.959</b>	.921
Happy	.785	-.587	<b>.980</b>	-.031	.960
Likeable	.971	-.210	<b>.916</b>	.385	.987
Just	.704	.667	.194	<b>.950</b>	.940
Variance accounted for	3.263	1.538	2.696	2.106	4.802
Proportion of total variance	.653	.308	.539	.421	.960
Cumulative proportion	.653	.960	.539	.960	.960

## Factor Analysis

The essential purpose of factor analysis is to describe the covariance relationships among many variables in terms of a few underlying, but unobservable random quantities called factors.

Basically, the factor model is motivated by the following argument: suppose variables can be grouped by their correlation, i.e. suppose all variables within a particular group are highly correlated among themselves, but have relatively small correlations with variables in a different group. Then it is conceivable that each group of variables represents a single underlying construct, or factor, that is responsible for the observed correlations.

For Eg:

correlations from the group of test scores in classics, French, English, Mathematics are grouped by underlying factor "intelligence".

Factor analysis can be considered an extension of principal component analysis. Both can be viewed as attempts to approximate the covariance matrix  $\Sigma$ . However, the approximation based on the factor analysis model is more elaborate. The primary question in factor analysis is whether the data are consistent with a prescribed structure.

### The Orthogonal Factor Model:-

The observable random vector  $x$ , with  $P$  components, has mean  $\mu$  and covariance matrix  $\Sigma$ . The factor model postulates that  $x$  is linearly dependent

upon a few ~~and~~ unobservable random variables  $f_1, f_2, \dots, f_m$  called common factors, and  $p$  additional sources of variation  $\epsilon_1, \epsilon_2, \dots, \epsilon_p$  called error or some specific factors. In particular, the factor analysis model is

$$x_1 - \mu_1 = l_{11}f_1 + l_{12}f_2 + \dots + l_{1m}f_m + \epsilon_1$$

$$x_2 - \mu_2 = l_{21}f_1 + l_{22}f_2 + \dots + l_{2m}f_m + \epsilon_2$$

— (1)

$$\vdots$$

$$x_p - \mu_p = l_{p1}f_1 + l_{p2}f_2 + \dots + l_{pm}f_m + \epsilon_p$$

or in matrix notation,

$$x - \mu = L f + \epsilon \quad — (2)$$

The coefficient  $l_{ij}$  is called the loading of the  $i^{\text{th}}$  variable on the  $j^{\text{th}}$  factor, so the matrix  $L$  is the matrix of factor loadings. Note that the  $i^{\text{th}}$  specific factor  $\epsilon_i$  is associated only with the  $i^{\text{th}}$  response  $x_i$ . The  $p$  deviations  $x_1 - \mu_1, x_2 - \mu_2, \dots, x_p - \mu_p$  are expressed in terms of  $p+m$  random variables  $f_1, f_2, \dots, f_m, \epsilon_1, \epsilon_2, \dots, \epsilon_p$  which are unobservable factors.

With so many ~~and~~ unobservable quantities, a direct verification of the factor model from observation on  $x_1, x_2, \dots, x_p$  is hopeless. However, with some additional assumptions about the random vectors  $f$  and  $\epsilon$ , the model in (2) implies certain covariance relationships, which can be checked.

We assume that,

$$E(F) = 0, \quad \text{cov}(F) = E(FF') = I_{(m \times m)}$$

$$E(\epsilon) = 0_{(p \times 1)}, \quad \text{cov}(\epsilon) = E(\epsilon\epsilon') = \Psi = \begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_p \end{bmatrix} \quad (3)$$

and that  $F$  and  $\epsilon$  are independent, so

$$\text{cov}(\epsilon, F) = E(\epsilon F') = 0_{(p \times m)}$$

These assumptions and the relation in (2) constitute the orthogonal factor model.

Orthogonal Factor model with  $m$  common factors :-

$$X_{(p \times 1)} = \mu_{(p \times 1)} + L_{(p \times m)} F_{(m \times 1)} + \epsilon_{(p \times 1)} \quad (4)$$

$\mu_i$  - mean of variable  $i$

$\epsilon_i$  =  $i^{\text{th}}$  specific factor

$f_j$  =  $j^{\text{th}}$  common factor

$l_{ij}$  = loading of the  $i^{\text{th}}$  variable on the  $j^{\text{th}}$  factor.

The two unobservable random vectors  $F$  and  $\epsilon$  satisfy the following conditions:

$F$  and  $\epsilon$  are independent

$$E(F) = 0, \quad \text{cov}(F) = I$$

$$E(\epsilon) = 0, \quad \text{cov}(\epsilon) = \Psi, \quad \text{where } \Psi \text{ is a diagonal matrix.}$$

Covariance structure for the orthogonal factor model :

$$1. \quad \text{cov}(X) = LL' + \Psi$$

(or)

$$\text{Var}(x_i) = l_{ii}^2 + \dots + l_{im}^2 + \psi_i \quad (5)$$

$$\text{Cov}(x_i, x_k) = l_{ii} l_{kk} + \dots + l_{im} l_{km}$$

2)  $\text{Cov}(x_i, F) = L$

(or)  $\text{Cov}(x_i, f_j) = l_{ij}$

The model  $x - u = LF + \epsilon$  is linear in the common factors. If the responses  $x$  are, in fact, related to underlying factors, but the relationship is nonlinear, such as

$$x_1 - u_1 = l_{11} f_1 f_3 + \epsilon_1$$

$$x_2 - u_2 = l_{21} f_2 f_3 + \epsilon_2$$

Then the covariance structure  $L^T + \Psi$  given by (5) may not be adequate. The very important assumption of linearity is inherent in the formulation of the traditional factor model.

That portion of the variance of the  $i$ th variable contributed by the  $m$  common factors is called the  $i$ th communality. That portion of  $\text{Var}(x_i) = \sigma_{ii}$  due to the specific factor is often called the uniqueness, or specific variance. Denoting the  $i$ th communality by  $h_i^2$ ,

$$\sigma_{ii} = h_i^2 + \underbrace{l_{i1}^2 + l_{i2}^2 + \dots + l_{im}^2}_{\text{communality}} + \psi_i$$

$$\sigma_{ii} = \text{Var}(x_i)$$

$\psi_i$  - Specific variance

$$(or) h_i^2 = l_{i1}^2 + l_{i2}^2 + \dots + l_{im}^2$$

$$\text{and } \sigma_{ii} = h_i^2 + \psi_i, i = 1, 2, \dots, p$$

The  $i$ th communality is the sum of squares of the loadings of the  $i$ th variable on the  $m$  common factors.

Problem:-

1) consider the covariance matrix  $\Sigma = \begin{bmatrix} 19 & 30 & 2 & 12 \\ 30 & 57 & 5 & 23 \\ 2 & 5 & 38 & 47 \\ 12 & 23 & 47 & 68 \end{bmatrix}$

and verify the relation  $\Sigma = LL^T + \Psi$

where  $L = \begin{bmatrix} 4 & 1 \\ 1 & 2 \\ -1 & 6 \\ 1 & 8 \end{bmatrix} \quad \Psi = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}$

$$LL^T + \Psi = \begin{bmatrix} l_{11} & l_{12} \\ 4 & 1 \\ 1 & 2 \\ -1 & 6 \\ 1 & 8 \end{bmatrix} \begin{bmatrix} 4 & 1 & -1 & 1 \\ 1 & 2 & 6 & 8 \end{bmatrix} + \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}$$

The communality of  $x_1$  is,

$$h_1^2 = l_{11}^2 + l_{12}^2 = 4^2 + 1^2 = 17.$$

and the variance of  $x_1$  can be decomposed as.

$$\sigma_{11} = (l_{11}^2 + l_{12}^2) + \psi = h_1^2 + \psi, \\ = 17 + 2 = 19$$

2) Let  $p=3$  and  $m=1$ , and suppose the random variables  $x_1, x_2$  and  $x_3$  have the positive definite covariance matrix  $\Sigma = \begin{bmatrix} 1 & 0.9 & 0.7 \\ 0.9 & 1 & 0.4 \\ 0.7 & 0.4 & 1 \end{bmatrix}$ . Check the existence of proper solution.

By factor model, we obtain

$$x_1 - H_1 = l_{11} F_1 + \epsilon_1$$

$$x_2 - H_2 = l_{21} F_1 + \epsilon_2$$

$$x_3 - H_3 = l_{31} F_1 + \epsilon_3$$

The covariance structure,

$$\Sigma = LL' + \Psi$$

$$\Rightarrow \begin{bmatrix} l_{11} & l_{12} & l_{13} \\ l_{21} & l_{22} & l_{23} \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} l_{11} & l_{12} & l_{13} \\ l_{21} & l_{22} & l_{23} \\ l_{31} & l_{32} & l_{33} \end{bmatrix}' + \begin{bmatrix} \psi_1 & 0 & 0 \\ 0 & \psi_2 & 0 \\ 0 & 0 & \psi_3 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0.9 & 0.7 \\ 0.9 & 1 & 0.4 \\ 0.7 & 0.4 & 1 \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} l_{11}^2 & l_{21}l_{11} & l_{31}l_{11} \\ l_{11}l_{21} & l_{21}^2 & l_{31}l_{21} \\ l_{11}l_{31} & l_{21}l_{31} & l_{31}^2 \end{bmatrix} + \begin{bmatrix} \psi_1 & 0 & 0 \\ 0 & \psi_2 & 0 \\ 0 & 0 & \psi_3 \end{bmatrix} =$$

$$l_{11}^2 + \psi_1 = 1 , \quad l_{11}l_{21} = 0.9 , \quad l_{11}l_{31} = 0.7$$

$$l_{21}^2 + \psi_2 = 1 , \quad l_{21}l_{31} = 0.4$$

$$l_{31}^2 + \psi_3 = 1$$

From the pair of equations

$$\text{--- } l_{11}l_{31} = 0.7$$

$$l_{21}l_{31} = 0.4$$

$$l_{21} = \left( \frac{0.4}{0.7} \right) l_{11}$$

$$l_{11} l_{21} = 0.9$$

$$l_{11} \left( \frac{0.4}{0.7} \right) l_{11} = 0.9$$

$$\boxed{l_{11}^2 = 1.575}$$
$$l_{11} = \pm 1.255$$

$$\text{Var}(F_1) = 1 \text{ (by assumption)}$$

$$l_{11}^2 + \psi_1 = 1$$

$$1.575 + \psi_1 = 1$$

$$\boxed{\psi_1 = -0.575}$$

which is unsatisfactory, since it gives a negative value for  $\text{Var}(\epsilon_1) = \psi_1$ .

$\therefore$  The solution is not consistent with the statistical interpretation of the coefficients, so it is not a proper solution.