MODULE 3

CHAPTER 3

Linear Models

Syllabus

The least-squares method, Multivariate Linear Regression, Regularised Regression, Using Least-Squares Regression for classification, Support Vector Machines.

3.1 INTRODUCTION

- In practice very often a relationship is found to exist between two (or more) variables and one wishes to express this relationship in mathematical form by determining an equation connecting the variables.
- A first step is the collection of data showing corresponding values of the variables. For example, suppose x and y denote, respectively the height and weight of boys of engineering colleges in Pune. Then a sample of n individuals would reveal the heights $x_1, x_2, ..., x_n$ and the corresponding weights $y_1, y_2, ..., y_n$.
- A next step is to plot the points $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ on a co-ordinate system. The resulting set of points is called a scatter diagram.
- From the scatter diagram, it is often possible to visualise a smooth curve approximating the data. If the data appear to be approximated well by a straight line, then we say that a linear relationship exists between the variables (Fig. 3.1.1(i)).
- However, although a relationship exists between the variables, it need not be a linear relationship and so we call it a nonlinear relationship (Fig. 3.1.1(ii)). And there may not be any relationship between the variables. (Fig. 3.1.1(iii)).

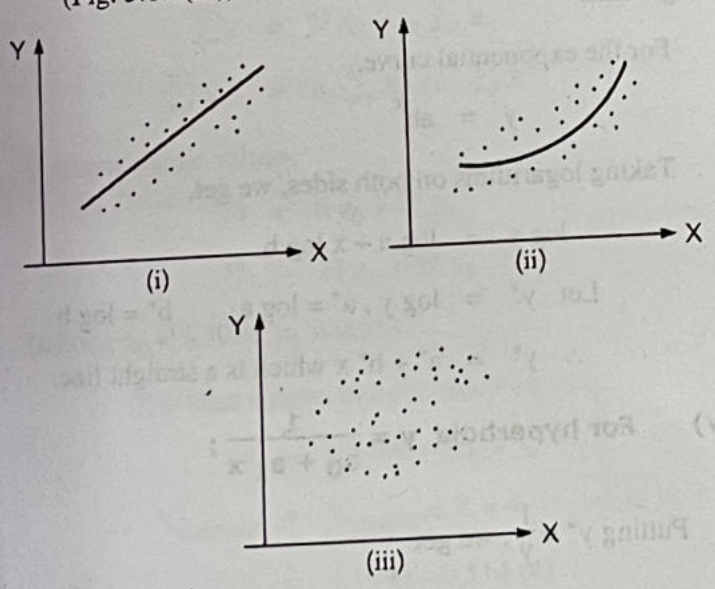

Fig. 3.1.1

3.2 APPROXIMATING CURVE

It is a smooth curve that approximates the given set of data points plotted in the scatter diagram.

(I) Lagrange's Interpolation

Formula is used when $x_i^{'s}$ are equally spaced. Calculations become laborious when the number of data points is large.

(II) Best fitting curve by method of least squares

Let $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ be a given set of n data points.

Let $d_i = y_i - \hat{y}_i$ denote the difference between y_i and the corresponding value

where $\hat{y}_i = f(x_i)$ determined from the curve

$$C : y = f(x)$$

The $d_i^{'s}$ are known deviations, errors or **residuals** which may be positive, negative or zero.

Then the Legendre's principle of least squares (L.S.) or least squares criterion states that, of all the curves approximating a given set of data points, the curve having the least or minimum sum of the squares of the deviations is the 'best fitting' curve.

i.e. $\sum_{i=1}^{n} d_i^2$ the residuals or error sum of squares is minimum.

Such a curve is known as a least squares (L.S.) curve.

Thus the least squares criterion is the measure of "goodness of fit".

If the curve is a straight line fitted according to least squares sense then it is known as least squares straight line, if it is a parabola, it is known as 'least squares parabola' etc.

3.2.1 Some Standard Approximating Curves

1. $y = a_0 + a_1 x$ straight line

2. $y = a_0 + a_1 x + a_2 x^2$, parabola or quadratic curve
3. $y = a_0 + a_1 x + ...+ a_n x^n$, n^{th} degree (polynomial) curve.
4. $y = ab^x$, exponential curve
5. $y = ax^b$, geometric curve
6. $y = \frac{1}{a_0 + a_1 x}$, hyperbola

3.2.2 Curve Fitting by Least Squares

(i) Least square straight line

For a given set of N data points (x_1, y_1), (x_2, y_2), ..., (x_N, y_N); assume that the straight line

$$y = a_0 + a_1 x = f(x) \qquad ...(3.2.1)$$

fits to the data in the least squares sense.

To determine the two unknowns a_0 and a_1 in Equations 3.2.1, we use the L.S. criterion that

$\sum d_i^2$ is minimum, i.e.,

$$\sum_{i=1}^{N} d_i^2 = \sum_{i=1}^{N} (y_i - a_0 - a_i x_i)^2 \qquad ...(3.2.2)$$

where $d_i = y_i - \hat{y}_i$; where y_i is given value and $\hat{y}_i$ is from the equation of line

is minimum. Differentiating Equation (3.2.2) partially with respect to a_0 and a_1 and equating to zero, we get,

$$\frac{\partial}{\partial a_0}(\sum d_i^2) = \frac{\partial}{\partial a_0}[\sum (y_i - a_0 - a_1 x_i)^2]$$

$$= 2\sum (y_i - a_0 - a_1 x_i)(-1) = 0$$

$$\therefore \sum (y_i - a_0 - a_1 x_i) = 0$$

$$\text{i.e. } \sum_{i=1}^{N} y_i = N a_0 - a_1 \sum_{i=1}^{N} x_i \qquad ...(3.2.3)$$

Similarly,

$$\frac{\partial}{\partial a_1}(\sum d_i^2) = \frac{\partial}{\partial q_1}[\sum (y_i - a_0 - a_1 x_i)^2]$$

$$= 2\sum (y_i - a_0 - a_1 x_i)(-x_i) = 0$$

$$\therefore \sum (y_i - a_0 - a_1 x_i) x_i = 0$$

$$\therefore \sum y_i x_i = a_0 \sum x_i + a_1 \sum x_i^2 \qquad ...(3.2.4)$$

Thus the two unknown parameters a_0, a_1 of Equations (3.2.1) are determined from the two equations

$$\sum y_i = N a_0 + a_1 \sum x_i \qquad ...(3.2.5)$$

$$\text{and } \sum x_i y_i = a_0 \sum x_i + a_1 \sum x_i^2 \qquad ...(3.2.6)$$

known as 'Normal equations'.

In such a case Equation (3.2.1) represents a least squares straight line.

(ii) Least Squares quadratic curve (parabola)

We assume that,

$$y = a_0 + a_1 x + a_2 x^2 \qquad ...(3.2.7)$$

Approximates the data according to L.S. principle. Then the unknown three parameters a_0, a_1, a_2 are determined from the following three normal equations obtained in a similar way as above

$$\sum y_i = N a_0 + a_1 \sum x_i + a_2 \sum x_i^2$$

$$\sum x_i y_i = a_0 \sum x_i + a_1 \sum x_i^2 + a_2 \sum x_i^3$$

$$\sum x_i^2 y_i = a_0 \sum x_i^2 + a_1 \sum x_i^3 + a_2 \sum x_i^4$$

(iii) Non-linear curve

Non-linear curves can be transformed to a linear curve straight line.

For the exponential curve,

$$y = ab^x$$

Taking logarithms on both sides, we get,

$$\log y = \log a + x \log b$$

Let $y^* = \log y$, $a^* = \log a$ $\quad b^* = \log b$

$$\therefore y^* = a^* + b^* x \text{ which is a straight line.}$$

(iv) For hyperbola $y = \frac{1}{a_0 + a_1 x}$;

Putting $y^* = \frac{1}{y}$; we get

$$y^* = a_0 + a_1 x \text{ which is linear.}$$

3.2.3 Examples on Fitting Straight Lines using L.S.M.

Ex. 3.2.1 : Find a least square straight line for the following data :

X	1	2	3	4	5	6
Y	6	4	3	5	4	2

and estimate (predict) Y at X = 4 and X at Y = 4.

☑ **Soln. :**

X	Y	X^2	Y^2	XY
1	6	1	36	6
2	4	4	16	8
3	3	9	9	9
4	5	16	25	20
5	4	25	16	20
6	2	36	4	12
Total	**21**	**24**	**91**	**106**	**75**

$\therefore \sum X = 21, \quad \sum y = 24, \quad \sum X^2 = 91$

$\sum Y^2 = 106, \quad \sum XY = 75, \quad N = 6$

Let the least squares straight line of Y on X is

$$Y = a_0 + a_1 X$$

Its normal equations are

$$\sum Y = N a_0 + a_1 \sum X$$

$$\sum XY = a_0 \sum X + a_1 \sum X^2$$

Substituting the values,

$$24 = 6 a_0 + 21 a_1$$

$$75 = 21 a_0 + 91 a_1$$

Solving $a_0 = 5.7999$, $a_1 = 0.514$

$\therefore$ The least square straight line Y on X is,

$$Y = 5.7999 - 0.514 X$$

$$\therefore Y_{estimate} = Y_{predict} \text{ at } X = 4$$

$$= 5.7999 - 0.514 (4)$$

$$= 3.743$$

Similarly, L.S.S.L. of X on Y can be written as,

$$X = b_0 + b_1 Y$$

Normal equations are,

$$\sum X = N b_0 + b_1 \sum Y$$

$$\text{and } \sum XY = b_0 \sum Y + b_1 \sum Y^2$$

$$\therefore 21 = 6 b_0 + 24 b_1$$

$$75 = 24 b_0 + 106 b_1$$

Solving $b_0 = 7.1$, $b_1 = -0.9$

$$\therefore X = 7.1 - 0.9 Y$$

$$\therefore X_{estimate} = X_{predict} \text{ at } Y = 4$$

$$= 7.1 - 0.9 (4) = \mathbf{3.5} \quad \textbf{...Ans.}$$

Ex. 3.2.2 : Fit a least square quadratic curve to the following data :

X	1	2	3	4
Y	1.7	1.8	2.3	3.2

Estimate Y (2.4).

☑ **Soln. :**

Let the quadratic curve be

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

The normal equations are :

$$\sum y = N a_0 + a_1 \sum x + a_2 \sum x^2$$

$$\sum xy = a_0 \sum x + a_1 \sum x^2 + a_2 \sum x^3$$

$$\text{and } \sum x^2 y = a_0 \sum x^2 + a_1 \sum x^3 + a_2 \sum x^4$$

Here N = 4,

x	y	x^2	xy	x^3	x^4	$x^2 y$
1	1.7	1	1.7	1	1	1.7
2	1.8	4	3.6	8	16	7.2
3	2.3	9	6.9	27	81	20.7
4	3.2	16	12.8	64	256	51.2
Total	**10**	**9.0**	**30**	**25.0**	**100**	**354**	**80.8**

Substituting these sums into normal equations,

We have,

$$9.0 = 4a_0 + 10a_1 + 30a_2$$

$$25 = 10a_0 + 30a_1 + 100a_2$$

$$80.8 = 30a_0 + 100a_1 + 354a_2$$

Solving, $a_0 = 2, a_1 = -0.5, a_2 = 0.2$

$\therefore$ L.S. curve is,

$$y = 2 - 0.5x + 0.2x^2$$

$$\therefore y_{estimate} = y(2.4) = 2 - 0.5(2.4) + 0.2(2.4)^2$$

$$= 1.952$$ **...Ans.**

Ex. 3.2.3 : Fit a least square straight line (L.S.S.L.) to the following data :

X	2	7	9	1	5	12
Y	13	21	23	14	15	21

☑ **Soln. :** Here $n = 6, \sum X = 36, \sum Y = 107$

$$\sum X^2 = 304, \quad \sum Y^2 = 2001, \quad \sum XY = 721$$

$$\therefore \overline{X} = \frac{36}{6} = 6, \quad \overline{Y} = \frac{107}{6} = 17.833$$

Now, $S_{XX} = n\sum X^2 - (\sum X)^2$

$$= 6(304) - (36)^2 = 528$$

$$S_{YY} = n\sum Y^2 - (\sum Y)^2$$

$$= 6(2001) - (107)^2 = 557$$

$$S_{XY} = n\sum X_i Y_i - (\sum Y_i)(\sum Y_i)$$

$$= 6(721) - (36)(107) = 474$$

Now, regression coefficient

$$b = \frac{S_{XY}}{S_{XX}} = \frac{474}{528} = 0.8977$$

Intercept $a = \overline{Y} - b\overline{X}$

$$= (17.8333) - (0.8977)(6)$$

$$= 12.447$$

$\therefore$ L.S.S.I. is $Y = a + bX$

$$= 12.45 + 0.8977X$$ **...Ans.**

3.3 LINEAR WEIGHTED LEAST SQUARES APPROXIMATION

- Data are generally not exact. They are subject to measurement errors. Modeling of data aims at summarising a given set of observations by fitting it to a model, a 'merit function' that depends on adjustable parameters.
- The parameters of the model are then adjusted to achieve a minimum in the merit function, yielding "best-fit" parameters.
- Least square fitting is a maximum likelihood estimation of the fitted parameters if the measurement errors are independent and normally distributed with constant standard deviation.
- The least square with constant standard deviation. The least square principle is to minimize the sum of the squares of the errors.
- For a given set of data, it gives a unique solution.
- For discrete data $(x_0, y_0), (x_1, y_1), \ldots, (x_N, y_N)$ weights w_i are positive numbers prescribed according to the relative accuracy of the data points.
- For continuous (data) function, an integrable function $w(x)$ is called a 'Weight function' on (a, b) of $w(x) \geq 0$ for $x \in [a, b]$.

General weighted least squares approximation

Suppose the function $y = f(x)$ is known only at $(N+1)$ tabulated points $(x_0, y_0), \ldots, (x_N, y_N)$ in the form of a discrete data, with weights $w_0, w_1, w_2, \ldots, w_N$ respectively,

x	x_0	x_1	x_2	...	x_N
y	y_0	y_1	y_2	...	y_N
w	w_0	w_1	w_2	...	w_N

Then the function $f(x)$ can be approximated by a function of the form

$$P(x) = a_0\phi_0(x) + a_1\phi_1(x) + \ldots + a_m\phi_m(x) \quad \ldots(3.3.1)$$

Where the set of functions $\{\phi_0, \phi_1, \ldots \phi_N\}$ are linearly independent. These functions $\phi_i(x)$ are known as "basis" or "co-ordinate" functions.

The function p(x) is said to be the weighted least squares approximation of f(x) if the (m + 1) unknown coefficients $a_0, a_1, \ldots a_m$ in Equation (3.3.1) are determined such that the error of approximation.

$$E(a_0, a_1, a_2, \ldots, a_m) = \sum_{k=0}^{N} w_k \left[f(x_k) - \sum_{i=0}^{m} a_i \phi_i(x_k) \right]^2 \quad \ldots(3.3.2)$$

is minimum. The necessary condition for the numbers $a_0, a_1, \ldots a_m$ to minimize E are,

$$\frac{\partial E}{\partial a_j} = 0 \text{ for } j = 0, 1, 2, \ldots, m$$

Differentiating Equation (3.3.2) partially with respect to j, we get (m + 1) linear equations in (m + 1) unknowns $a_0, a_1, \ldots, a_m$ known as normal equations given by,

$$\sum_{k=0}^{N} w_k \left[f(x_k) - \sum_{i=0}^{m} a_i \phi_i(x_k) \right] \phi_j(xk) = 0 \quad \ldots(3.3.3)$$

for j = 0, 1, ..., m

When the function f(x) is known continuous function on [a, b] then the normal equations take the form

$$\int_a^b w(x) \left[f(x) - \sum_{i=0}^{m} a_i \phi_i(x) \right] \phi_j(x) \cdot dx = 0 \quad \ldots(3.3.4)$$

For j = 0, 1, ..., m.

☞ Discrete (Data) case

Suppose $P_1(x) = a_0 + a_1 x$ ...(3.3.5)

be the linear weighted least squares straighted line fitted to the following discrete data

x :	x_0	x_1	x_2	...	x_N
y :	y_0	y_1	y_2	...	y_N
w :	w_0	w_1	w_2	...	w_N

In Equation (3.3.1), we consider m = 1 and $\phi_1(x) = x$, then normal Equation (3.3.3) reduce to,

$$a_0 \sum_{i=0}^{N} w_i + a_1 \sum_{i=0}^{N} w_i x_i = \sum_{i=0}^{N} w_i y_i \quad \ldots(3.3.6)$$

$$a_0 \sum w_i x_i + a_1 \sum w_i x_i^2 = \sum w_i x_i y_i \quad \ldots(3.3.7)$$

Solving Equation (3.3.6) and (3.3.7) we get a_0 and a_1 which when substituted in Equations (3.3.5) gives the required linear weighted least squares approximations.

☞ Continuous function (Case)

Suppose f(x) is a known continuous function defined on the interval [a, b], then the normal Equations (3.3.4) reduce to

$$a_0 \int_a^b w(x)\,dx + a_1 \int_a^b x \cdot w(x)\,dx = \int_a^b w(x) \cdot y(x)\,dx \quad \ldots(3.3.8)$$

$$\text{and } a_0 \int_a^b x\, w(x)\,dx + a_1 \int_a^b x^2 \cdot w(x)\,dx$$

$$= \int_a^b x\, w(x)\, y(x)\,dx \quad \ldots(3.3.9)$$

Solving Equation (3.3.8) and (3.3.9) we get a_0 and a_1. Substituting these values in $y = a_0 + a_1 x$ we get the linear weighted least squares approximation in the continuous case.

3.3.1 Examples : Linear Weighted Least Squares Approximations

Ex. 3.3.1 (Discrete Data)

Fit a linear weighted least squares straight line to the following data :

x	– 2	0	2	4	6
y	1	3	6	8	13
w	2	5	10	1	4

☑ Soln. :

Let $y = a_0 + a_1 x$ be the L.S. line. Then the normal equations are,

$$a_0 \sum_{i=1}^{5} w_i + a_1 \sum_{i=1}^{5} w_i x_i = \sum w_i y_i$$

and $a_0 \sum w_i x_i + a_1 \sum w_i x_i^2 = \sum w_i x_i y_i$

	x	y	w	wx	wx^2	wy	wxy
	-2	1	2	-4	8	2	-4
	0	3	5	0	0	15	0
	2	6	10	20	40	60	120
	4	8	1	4	16	8	32
	6	13	4	24	144	52	312
$\sum_{1}^{5}$			22	44	208	137	460

Thus $N = 5, \sum_{i=1} w_i = 22, \sum w_i x_i = 44$

$\sum w_i x_1^2 = 208, \sum w_i y_i = 137, \sum w_i x_i y_i = 460$

The two normal equations are :

$$22\, a_0 + 44\, a_1 = 137$$

$$44\, a_0 + 208\, a_1 = 460$$

$\therefore a_1 = 1.55, \quad a_0 = 3.127$

$\therefore$ Linear weighted least squares straight line fit is,

$$y = 3.127 + 1.55\, x$$

At $x = 1, \quad y = 4.677$ ...Ans.

Ex. 3.3.2 : (Continuous function)

Fit a linear weighted least square straight line to the function $f(x) = \frac{1}{x}$ on $[1, 3]$ with $w(x) = 1$.

Soln. :

Let $y = a_0 + a_1 x$ be the L.S. straight line.

The normal equations are

$$a_0 \int_1^3 dx + a_1 \int_1^3 x\, dx = \int_1^3 \frac{1}{x}\, dx$$

and $$a_0 \int_1^3 x\, dx + a_1 \int_1^3 x^2\, dx = \int_1^3 x \cdot \frac{1}{x}\, dx$$

$$\therefore 2a_0 + 4\, a_1 = \log 3 - \log 1 = \log 3$$

$$4\, a_0 + \frac{26}{3}\, a_1 = 2$$

Solving $a_1 = -0.2959$ and $a_0 = 1.140$

$\therefore$ The required linear L.S. line is,

$$y = 1.140 - 0.2959\, x$$

At $x = 2, \; y(2) = 0.5484$

At $x = 2, \; f(x) = \frac{1}{x} \quad \therefore \quad f(2) = \frac{1}{2} = 0.50$

3.3.2 Non-Linear Weighted Least Approximation

Given a set of (N + 1) data points, we non-linear m^{th} degree polynomial of the form

$$y = a_0 + a_1 x + a_2 x^2 + \ldots + a_m x^m$$

by minimizing the error function.

$$E(a_0, a_1, a_2 \ldots, a_m) = \sum w_i \left[y_i - \left(a_0 + a_1 x_i + \ldots + a_m x_i^m \right) \right]^2$$

The necessary conditions for minimis Equation (3.3.11) given the following (m + equations :

$$a_0 \sum_{i=0}^{N} w_i + a_1 \sum_i w_i x_i + \ldots + a_m \sum x_i^m w_i$$

$$= \sum_i x_i^m y_i \cdot w_i$$

$$a_0 \sum_i x_i^m w_i + a_0 \sum x_i^{m+1} w_i + \ldots\ldots + a_m \sum x_i^{2m} w_i =$$

w_i

when $m < N + 1$, then the above equations hav solution.

Discrete case

Suppose $y = a_0 + a_1 x + a_2 x^2$ is the no weighted least square approximation to a given (N + 1) data points. Then normal Equations (3.3.12) form

$$a_0 \sum_i w_i + a_1 \sum x_i w_i + a_2 \sum x_i^2 w_i = \sum y_i w_i \quad ...(3.3.13)$$

$$a_0 \sum x_i w_i + a_1 \sum x_i^2 w_i + a_2 \sum x_i^3 w_i = \sum x_i y_i w_i \quad ...(3.3.14)$$

$$a_0 \sum x_i^2 w_i + a_1 \sum x_i^3 w_i + a_2 \sum x_i^4 w_i = \sum x_i^2 y_i w_i \quad ...(3.3.15)$$

Solving these equations, we obtain the values of a_0, a_1 and a_2.

Ex. 3.3.3 : Fit a non-linear weighted least squares (parabola) second degree polynomial

$y = a_0 + a_1 x + a_2 x^2$ to the following data

x	– 3	– 1	1	3
y	15	5	1	5
w	2	5	10	20

☑ **Soln. :** The normal equations are :

$$a_0 \sum_{i=1}^{4} w_i + a_1 \sum w_i x_i + a_2 \sum w_i x_i^2 = \sum w_i y_i$$

$$a_0 \sum_{i=1}^{4} w_i x_i + a_1 \sum w_i x_i^2 + a_2 \sum w_i x_i^3 = \sum w_i x_i y_i$$

$$a_0 \sum w_i x_i^2 + a_1 \sum w_i x_i^3 + a_2 \sum w_i x_i^4 = \sum w_i x_i^2 y_i$$

x	y	w	wx	wx^2	wx^3	wx^4	wy	wxy	wx^2y
– 3	15	2	– 6	18	– 54	162	30	– 90	270
– 1	5	5	– 5	5	– 5	5	25	– 25	25
1	1	10	10	10	10	10	10	10	10
3	5	20	60	45	540	1620	100	300	900
0	26	37	59	78	491	1797	165	195	1205

The data is,

$$N = 4, \sum w_i = 37, \sum w_i x_i = 54, \sum w_i x_i^2 = 78$$

$$\sum w_i x_i^3 = 491, \sum w_i x_i^4 = 1797, \sum w_i y_i = 165$$

$$\sum w_i x_i y_i = 195, \sum w_i x_i^2 y_i = 1205$$

Thus the three normal equations are,

$$37 a_0 + 59 a_1 + 78 a_2 = 165$$

$$59 a_0 + 78 a_1 + 491 a_2 = 195$$

$$78 a_0 + 491 a_1 + 1797 a_2 = 1205$$

Solving $a_0 = 0.38$, $a_1 = 2.65$, $a_2 = -0.07$

Thus the non-linear weighted least square quadratic fit is,

$$y = 0.38 + 2.65 x - 0.07 x^2$$

with $y(1) = 2.96$ **...Ans.**

3.4 MULTIPLE REGRESSION MODEL

- Multiple Linear Regression (MLR), also known s simply 'Multiple Regression'. It is a statistical technique that uses several explanatory variables to predict the outcome of a response variable.
- The aim of multiple linear regression is to **model** the **linear relationship** between the **explanatory** (independent) variables and **response (dependent)** variables.
- In short, multiple regression is the extension of **Ordinary Least-Squares (OLS) regression** because it involves more than one explanatory variable. MLR is used extensively in econometrics and financial inference.

3.4.1 Multiple Regression

It is observed in agriculture that, the crop yield (Y) not only depends on the amount of rainfall (X_1) but also on the amount of fertilizer (X_2) applied, pesticides (X_3) used, quality of seeds (X_4), quality of soil (X_5) etc.

Thus in multiple regression, the dependent variable Y is a function of more than one independent variables, i.e.

$$Y = f(X_1, X_2, ..., X_n)$$

In multiple nonlinear regression, f is non-linear.

In multiple linear regression f is linear.

i.e., $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k.$$

3.4.2 Linear Multiple Regression

Suppose Y depends on two independent variables X_1 and X_2;

$$\text{i.e.,} \quad Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \qquad ...(3.4.1)$$

To estimate the coefficients β_0, β_1, β_2; we apply the least square method to minimise.

$$\sum_{i=1}^{N} \{ Y_i - (b_o + b_1 X_{1i} + b_2 X_{2i}) \}^2$$

This results in three normal equations given by

$$\sum_{i=1}^{N} Y_i = Nb_0 + b_1 \sum_{i=1}^{N} X_{1i} + b_2 \sum_{i=1}^{N} X_{2i}$$

$$\sum_{i=1}^{N} X_{1i} Y_i = b_o \sum X_{1i} + b_1 \sum_{i=1}^{N} X_{1i}^2 + b_2 \sum X_{1i} X_{2i}$$

$$\sum X_{2i} Y_i = b_o \sum X_{2i} + b_1 \sum X_{1i} \cdot X_{2i} + b_2 \sum X_{2i}^2$$

Here b_0, b_1, b_2 are the least squares estimates of β_0, β_1, β_2.

3.4.3 Linear Multiple regression in k-independent Variables

The above analysis can be generalised to fit N(k + 1) tuples (X_{1i}, X_{2i}, ..., X_{ki})

(i = 1 to N), to the equation.

$$Y = \beta_0 + b_1 X_1 + \beta_2 X_2 + ...+ \beta_k X_k.$$

The (k + 1) normal equations are :

$$\sum_{i=1}^{N} Y_i = Nb_0 + b_1 \sum_{i=1}^{N} X_{1i} + b_2 \sum X_{2i} + ... + b_k \sum_{i=1}^{N} X_{ki}$$

Ex. 3.4.1 : Fit a regression equation to estimate β_0, β_1, β_2 to the following data of a transport company on the weights of 6 shipments, the distances they were moved and the damage of the goods that was incurred. Estimate the damage when a shipment of 3700 kg. is moved to a distance of 260 km.

Weight X_1 (1000 kg)	4.0	3.0	1.6	1.2	3.4	4.8
Distance X_2 (100 km)	1.5	2.2	1.0	2.0	0.8	1.6
Damage Y (Rs.)	160	112	69	90	123	186

☑ Soln. :

Let weight X_1 and distance X_2 be independent variables and the damage y be the dependent variable.

Let the equation of regression be,

$$y = b_0 + b_1 x_1 + b_2 x_2$$

Where b_0, b_1, b_2 are estimates of β_0, β_1, β_2. The three normal equations become.

$$\sum_{i=1}^{6} Y_i = nb_0 + b_1 \sum_{i=1}^{6} x_{1i} + b_2 \sum_{i=1}^{6} x_{2i}$$

$$\sum x_{1i}\, y_i = b_0 \sum x_{1i} + b_1 \sum x_{1i}^2 + b_2 \sum x_{1i}\, x_{2i}$$

$$\sum x_{2i}\, y_i = b_0 \sum x_{2i} + b_1 \sum x_{1i}\, x_{2i} + b_2 \sum x_{2i}^2$$

We prepare the table :

	x_1 (weight) (1000 kg)	x_2 distance 100 km	y damage Rs.	x_1^2	x_2^2	$x_1 x_2$	$x_1 y$	$x_2 y$
	4.0	1.5	160	16	2.25	6.0	640	240
	3.0	2.2	112	0	4.84	6.6	336	246.4
	1.6	1.0	69	2.56	1.0	1.6	110.4	69
	1.2	2.0	90	1.44	4.0	2.4	108	180
	3.4	0.8	123	11.56	0.64	2.72	418.2	98.4
	4.8	1.6	186	23.04	2.56	7.68	892.8	297.6
Total	18	9.1	740	63.6	15.29	27	250.54	1131.4

The data is : $n = 6$, $\sum x_{1i} = 18$, $\sum x_{2i} = 9.1$, $\sum y_i = 740$, $\sum x_{1i}^2 = 63.6$, $\sum x_{2i}^2 = 15.29$, $\sum x_{1i}\, x_{2i} = 27$ $\sum x_i\, y_i = 250.54$, $\sum x_{2i}\, y_i = 1131.4$

$\therefore$ Normal equations become

$$740 = 6b_0 + 18\,b_1 + 9.1\,b_2$$

$$250.54 = 18b_0 + 63.6\,b_1 + 27\,b_2$$

$$1131.4 = 9.1\,b_0 + 27\,b_1 + 15.29\,b_2$$

Solving, we get $b_0 = 14.56$, $b_1 = 30.109$,

$b_2 = 12.16$

Thus the required regression equation is

$$y = 14.56 + 30.109\,x_1 + 12.16\,x_2$$

Estimate

For a weight of 3700 kg. ($x_1 = 3.7$) and for a distance of 260 km. ($x_2 = 2.6$) the damage incurred in rupees is

$$y\,(x_1 = 3.7, x_2 = 2.6) = 14.56 + 30.109(3.7) + 16\,(2.6)$$

$$= 714.58 = 715 \text{ Rs.}$$

Exercise

1. Find Y when $X_1 = 10$ and $X_2 = 6$ from the least squares regression equation of Y on X_1 and X_2 for the following data :

Y	90	72	54	42	30	12
X_1	3.	5	6	8	12	14
X_2	16	10	7	4	3	2

Ans. :

$Y = 61.40 - 3.65\,X_1 + 2.54\,X_2$; $Y\,(10.6) = 40.14 \simeq 40$

Hints : $n = 6$, $\sum Y = 300$, $\sum X_{1i} = 48$, $\sum X_{2i} = 42$,

$\sum X_{1i}\,X_{2i} = 236$, $\sum X_{1i}^2 = 474$,

$\sum X_{2i}^2 = 434$, $\sum X_{1i}\,Y_i = 1818$, $\sum X_{2i}\,Y_i = 2820$

3.4.4 Multivariate Linear Regression

- In this section we present multivariate regression model, in which we consider the relationship between more than one dependent variable and one independent variable. This model is similar to the multiple regression model in solving the normal equations and estimate the regression parameters. These parameters are easily estimated using matrix form.
- Suppose that the number of response variables is m, so we have n observations for each y_i, i = 1,2,...,m. The general formula for the multivariate regression model is given by :

$$y_i = \beta_{0i} + \beta_{1i}\,x_1 + \varepsilon_i$$

$$\hat{y}_i = \beta_{0i} + \beta_{1i}\,x_1\,,\ i = 1,2,\ldots,\ m$$

- There are two parameters for each response to be estimated when the linear model includes the intercept β_0.
- Four matrices are needed to express the linear model in matrix notation:
- Y : the n × m matrix of observations on the dependent variable y .
- X : the n × 2 matrix consisting of a column of ones, which is labeled 1, followed by the column vector of the observations on the independent variable.
- B : the 2 × m matrix of parameters to be estimated.

☞ **Example : Multivariate Linear Regression**

Ex. 3.4.2 : Given the following data, find the multivariate regression equations.

X_0	Y_1	Y_2
0	1	-1
1	4	-1
2	3	2
3	8	3
4	9	2

☑ **Soln. :** The regression equation takes the form,

$$\dot{y}_1 = \dot{\beta}_{01} + \dot{\beta}_{11}\,x_1$$

$$\dot{y}_2 = \dot{\beta}_{02} + \dot{\beta}_{12}\,x_2$$

$$Y = \begin{bmatrix} 1 & -1 \\ 4 & -1 \\ 3 & 2 \\ 8 & 3 \\ 9 & 2 \end{bmatrix} \quad X = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}$$

Now, we find the X^TX matrix

$$X^T X = \begin{bmatrix} 5 & 10 \\ 10 & 30 \end{bmatrix}$$

Find the inverse of the matrix $(X^T X)^{-1}$

$$= \begin{bmatrix} 0.6 & -0.25 \\ -0.2 & 0.1 \end{bmatrix}$$

The estimated regression coefficients is given by:

$$\hat{\beta}_{01} = (X^T X)^{-1} X^T Y$$

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

We get the equation :

$$\hat{Y}_1 = 1 + 2x_1 \quad , \quad \hat{Y}_2 = -1 + x_2$$

3.5 REGULARISED REGRESSION

Regularised regression is a type of regression where the **coefficient** estimates are constrained to zero.

The magnitude (size) of coefficient, as well as the magnitude of the **error term** are **penalised**. Complex models are discouraged to avoid **overfitting**.

"Regularisation" is a way to give a penalty to certain models, generally complex ones.

Ridge regression belongs to the set of L2 regularisation tools. L2 regularisation adds penalty called an L2 penalty, and is equal to the square of the magnitude of coefficients.

All coefficients are shrunk by the same factor, hence all coefficients remain in the same model.

The strength of the penalty term is controlled by a tuning parameter. When this tuning parameter (λ) is set to zero, ridge regression equals least square regression.

When $\lambda = \infty$, all coefficients are shrunk to zero.

Hence the ideal penalty is therefore somewhere in between 0 and ∞.

The other type of regularisation, L1 regularisation, limits the size of the coefficients by adding an L1 penalty equal to the absolute value of the magnitude of the coefficients.

Sometimes, this results in the elimination of some coefficients altogether, which can result in sparse models.

3.5.1 Types of Regularised Regression

Two commonly used types of regularised regression methods are :

(1) Ridge Regression

It is a way to create a **parsimonius** model when the number of predictor variables in a set exceeds the number of observations, or when a data set has multicollinearity (correlation between predictor variables).

(2) Lasso Regression

It is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean.

This type is very useful when there is a high level of multicollinearity or when one wants to automate certain parts of model selection, like variable selection/ parameter elimination.

3.5.2 Method of Regularisation Working

We have seen that Regularisation works by adding a penalty or complexity term to the complex model.

Let us consider the simple linear regression equation.

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + b;$

here y represents the value to be predicted;

$x_1, x_2, ..., x_n$ are the features of y;

$\beta_0, \beta_1, ...\beta_n$ are the weights or magnitude attached to the features, respectively' b represents the intercept.

Linear regression models try to optimise the β_0 and b to minimise the cost function.

The equation for the cost function for the linear model is :

$$\sum_{i=1}^{M} \left(y_i - y_i'\right)^2 = \sum_{i=1}^{M} \left(y_i - \sum_{j=0}^{n} \beta_j \cdot x_{ij} \right)^2 \quad ...(i)$$

Now, we add a loss function and optimise parameter to make the model that can predict the accurate value of y.

The loss function for linear regression is called as **RSS** or **Residual Sum of Squares.**

3.5.3 Ridge Regression

(i) Ridge regression is one of the types of linear regression in which a small amount of bias is introduced so that we can get better long-term predictions.

(ii) Ridge regression is a regularisation technique, which is used to reduce the complexity of the model. It is also called as L2 regularisation.

(iii) In this technique, the cost function is altered by adding the penalty term to it. The amount of bias added to the model is called **Ridge Regression Penalty**.

We can calculate it by multiplying with the lambda to the squared weight of each individual feature.

(iv) The equation for the cost function in ridge regression will be

$$\sum_{i=1}^{M} \left(y_i - y_i'\right)^2 = \sum_{i=1}^{M} \left(y_i - \sum_{j=0}^{n} \beta_j \cdot x_{ij} \right)^2 + \lambda \sum_{j=0}^{n} \beta_j^2 \quad \text{...(ii)}$$

(v) In the above equation, the penalty term regularises the coefficients of the model.

Thereby ridge regression reduces the amplitudes of the coefficients that decreases the complexity of the model.

(vi) If the values of λ tend to zero, the equation becomes the cost function of the linear regression model.

Thus, for minimum value of λ, the model will resemble the linear regression model.

(vii) A general linear or polynomial regression fails if there is high collinearity between the independent variables, hence to solve such problems, Ridge regression can be used.

(viii) It helps to solve the problems it we have more parameters than samples.

3.5.4 Lasso Regression

(i) Lasso regression is another regularisation technique to reduce the complexity of the model. It stands for least Absolute and Selection Operator.

(ii) It is similar to the Ridge Regression except that the penalty term contains only the absolute weights instead of a square of weights.

(iii) Since, it takes absolute values, it can shrink the slope to 0, whereas Regression can only shrink it near to 0.

(iv) It is also called as L1 regularisation. The equation for the cost function of Lasso regression is :

$$\sum_{i=1}^{M} \left(y_i - y_i'\right)^2 = \sum_{i=1}^{M} \left(y_i - \sum_{j=0}^{n} \beta_j \cdot x_{ij} \right)^2 + \lambda \sum_{j=0}^{n} |\beta_j|$$

(v) Some of the features in this technique are completely neglected for model evaluation.

(vi) Thus, the Lasso regression can reduce the over fitting in the model as well as the feature selection.

3.5.5 Difference between Ridge Regression and Lasso Regression

(i) **Ridge regression** is mostly used to reduce the over fitting in the model, and it includes all the features present in the model.

It reduces the complexity of the model by shrinking the coefficients.

(ii) **Lasso regression** reduces the over fitting in the model as well as feature selection.

3.5.6 Use of Ridge Regression

Ridge Regression is used for the analysis of prostate-specific antigen and clinical measures among people who were about to have their prostates removed.

The performance of ridge regression is good when there is a subset of true coefficients which are small or even zero.

3.5.7 Advantages and Disadvantages of Ridge Regression

(I) Advantages

(i) Ridge regression is a better predictor than the least squares regression when the predictor variables are more than the observations.

(ii) Ridge Regression is very useful when there is co-linearity in data.

(II) Disadvantages

(i) It includes all the predictors in the final model.

(ii) It is not capable of performing feature selection.

(iii) It shrinks coefficients towards zero.

(iv) It trades variance for bias.

3.6 USING LEAST SQUARE REGRESSION FOR CLASSIFICATION

3.6.1 Extension of MLR to n Variables

Multiple regression analysis is the extension of the two variable regression theory to n variables $X_1, X_2, \ldots\ldots, X_n$.

The basic objectives of multiple linear regression are :

(i) To fit the plane of regression of the dependent variable, say X, on the independent variables $(X_2, X_3, \ldots\ldots X_n)$ by the **principal of least squares** and use this plane to estimate the value of the dependent variable X_1 for a given set of values of the independent variables.

(ii) To estimate and to compute the error $(X_1 - \hat{X}_1)$. This is achieved by calculating the standard error of the estimate. Problem will make this point clear.

(iii) To determine how much variation in the dependent variable is accounted for by the fitted plane of regression.

This is achieved through the **multiple coefficient of** determination.

3.6.2 Yule's Notation

Consider a trivariate distribution with three variables X_1, X_2, X_3. Let X_1 be dependent variable and X_2, X_3 be independent variables.

Then the equation of the plane of regression of X_1 on X_2 and X_3 is :

$$X_1 = a + b_{12.3} X_2 + b_{13.2} X_3 \quad \ldots(i)$$

Since the regression coefficients (and correlation coefficients) are independent of change of origin. We assume that X_1, X_2, X_3 are measured from their respective means, hence

$$E(X_1) = E(X_2) = E(X_3) = 0 \quad \ldots(ii)$$

Taking expection on both sides of Equation (i),

$$E(X_1) = E(a) + b_{1.3}\ E(X_2) + b_{13.2}\ E(X_3)$$

$$\therefore \quad 0 = E(a) + 0 + 0$$

$$\therefore \quad E(a) = 0 \qquad \therefore \ a = 0$$

$\therefore$ Equation (i) becomes, $X_1 = b_{12.3}\ X_2 + b_{13.2}\ X_3 \ldots(iii)$

Where, $b_{12.3}$ = **Partial regression coefficient** of X_1 on X_2 (X_3 being the third variable)

and $b_{13.2}$ = Partial regression coefficient of X_1 on X_3 (X_2 being the third variable) ...(iv)

For given values of X_2 and X_3, the estimate of X_1 given by Equation (iii) is denoted by $e_{1.23}$.

$$\therefore \quad e_{1.23} = \hat{X}_1 = b_{12.3} X_2 + b_{13.2} X_3 \quad \ldots(v)$$

$$\therefore \quad X_{1.23} = X_1 - \hat{X}_1 = X_1 - e_{1.23}$$

$$\therefore \quad X_{1.23} = X_1 - b_{12.3}\, X_2 - b_{13.2}\, X_3 \qquad \ldots(vi)$$

is called the **residual** or **the error of estimate.**

☞ **Remarks**

(1) The subscripts before the dot are known as **primary subscripts** and those the dot are known as **secondary subscripts.**

(2) These notations can be generated to n variables $X_1, X_2, \ldots\ldots X_n$.

The equation of plane of regression of X_1 on $(X_2, X_3, \ldots\ldots, X_n)$ is given by :

$$X_1 = b_{12.34\ldots\ldots n} X_2 + b_{13.24\ldots\ldots n} X_3 + \ldots\ldots + b_{1n.23\ldots\ldots(n-1)} X_n \qquad \ldots(vii)$$

(3) and $X_{1.23\ldots\ldots n} = X_1 - \hat{X}_1$

$$= \left[X_1 - \left(b_{12.3\ldots\ldots n} X_2 + b_{13.24\ldots\ldots n} X_3 + \ldots\ldots + b_{1n.23\ldots(n-1)} X_n \right) \right] \qquad \ldots(viii)$$

Where, $X_i's$; $i = 1, 2, \ldots\ldots, n$ are measured from their respective means i.e.

$$E(X_i) = 0 \; ; \; i = 1, 2, \ldots\ldots\ldots, n \qquad \ldots(ix)$$

3.6.3 Order of Regression Coefficients

The **order of regression coefficients** is given by the **number of secondary subscripts in it.**

For example, '$b_{12.3}$' is a regression coefficient of order 1;

$b_{12.34}$ is a regression coefficient of order 2

$b_{12.34\ldots\ldots n}$ is a regression coefficient of order $(n - 2)$.

Thus, we can say : a regression coefficient with **K secondary subscripts** is called the regression coefficient of **order K.**

☞ **Remarks**

(1) In a regression coefficient, the order of the secondary subscripts is immaterial.

For example,

$$b_{12.345} = b_{12.354} = b_{12.435} = b_{12.453} = b_{12.534} = b_{12.543} \qquad \ldots(x)$$

(2) The ordering of the primary subscripts is important.

Of the two primary subscripts, the first subscript refers to the dependent variable and the second subscript refers to the independent variable.

For example,

In $b_{12.34\ldots\ldots n}$, X_1 refers to the dependent variable and X_2 refers to the independent variable under consideration.

In $b_{21.34\ldots\ldots n}$, X_2 refers to dependent variable and X_1 refers to independent variable

(3) The order of a residual is also determined by the number of secondary subscripts and is independent of the permutations of the secondary subscripts.

For example, $X_{1.2}$, $X_{1.23}$, $X_{1.23\ldots\ldots n}$ are residuals of order 1,2,.....,$(n - 1)$ respectively.

and also,

$$X_{1.234} = X_{1.243} = X_{1.342} = X_{1.324} = X_{1.432} \qquad \ldots(xi)$$

3.6.4 Planes of Regression

Consider the distributions with n variables $X_1, X_2, \ldots\ldots, X_n$.

We assume that the variables $X_1, X_2, \ldots\ldots, X_n$ are measured from their respective means, so that

$$E(X_i) = 0 \; ; \quad \text{for} \quad i = 1, 2, \ldots\ldots, n \qquad \ldots(xii)$$

Let N be the observations on each of the n variables $X_1, X_2, \ldots\ldots X_n$; so that

$$\sigma_i^2 = \frac{1}{N} \sum X_i^2 \; ; \qquad i = 1, 2, \ldots\ldots, n \qquad \ldots(xiii)$$

$$\text{and Cov}\,(X_i, X_j) = \frac{1}{N} \sum X_i X_j \; ; \; i \neq j = 1, 2, \ldots\ldots, n \qquad \ldots(xiv)$$

Summation is taken over N observations for each of the variables X_i, X_j $(i, j = 1, 2, \ldots, n)$

❒ **Definition :** Karl Pearson's correlation coefficient r_{ij} between (X_i, X_j) is given by :

$$r_{ij} = \frac{Cov(X_i, X_j)}{\sigma_i \sigma_j} = \frac{\sum X_i, X_j}{N \sigma_i \sigma_j} \quad \ldots(xv)$$

$$\therefore \sum X_i X_j = N \sigma_i \sigma_j r_{ij}$$

3.6.5 Equations of Planes of Regression

Initially we consider the case of three variables X_1, X_2, X_3 satisfying the conditions (xii), (xiii), (xiv), (xv)

Since X_1, X_2, X_3 are measured from their means, the equation of the plane of regression of X_1 on (X_2 and X_3) is

$$X_1 = b_{12.3} X_2 + b_{13.2} X_3 \quad \ldots(xvi)$$

We determine the constants $b_{12.3}$ and $b_{13.2}$ by the principle of least squares by minimizing the sum of squares of errors of **residuals**.

Thus, we determine $b_{12.3}$ and $b_{13.2}$, so that

$$E = \sum X_{1.23}^2 = \sum (X_1 - b_{12.3} X_2 - b_{13.2} X_3)^2 \text{ is minimum} \quad \ldots(xvii)$$

Using the principle of maxima and minima the **normal equations** for estimating $b_{12.3}$ and $b_{13.2}$ are

$$\frac{\partial E}{\partial b_{12.3}} = -2 \sum X_2 (X_1 - b_{12.3} X_2 - b_{13.2} X_3) = 0$$

and $$\frac{\partial E}{\partial b_{13.2}} = -2 \sum X_3 (X_1 - b_{12.3} X_2 - b_{13.2} X_3) = 0$$

$$\Rightarrow \sum X_2 X_{1.23} = 0 \; ; \quad \sum X_3 X_{1.23} = 0 \quad \ldots(xviii)$$

Also, $$\sum X_1 X_2 - b_{12.3} \sum X_2^2 - b_{13.2} \sum X_2 X_3 = 0$$

and $$\sum X_1 X_3 - b_{12.3} \sum X_2 X_3 - b_{13.2} \sum X_3^2 = 0$$

Using Equations (xiii) and (xv); the equations simplify to :

$$\left.\begin{aligned} r_{12} \sigma_1 \sigma_2 - b_{12.3} \sigma_2^2 - b_{13.2} r_{23} \sigma_2 \sigma_3 = 0 \\ \text{and } r_{13} \sigma_1 \sigma_3 - b_{12.3} r_{23} \sigma_2 \sigma_3 - b_{13.2} \sigma_3^2 = 0 \end{aligned}\right\} \ldots(xix)$$

On Simplification,

$$\left.\begin{aligned} b_{12.3} \sigma_2 + b_{13.2} r_{23} \sigma_3 - r_{12} \sigma_1 = 0 \\ b_{12.3} \sigma_2 r_{23} + b_{13.2} \sigma_3 - r_{13} \sigma_1 = 0 \end{aligned}\right\} \ldots(xx)$$

Solving Equation (xx) by Cramer's rule for $b_{12.3}$ and $b_{13.2}$; we get

$$\frac{b_{12.3}}{r_{12} \sigma_1 \sigma_3 - r_{13} \sigma_1 \sigma_3 r_{23}} = \frac{b_{13.2}}{r_{13} \sigma_1 \sigma_2 - r_{12} r_{23} \sigma_1 \sigma_2} = \frac{1}{\sigma_2 \sigma_3 - r_{23}^2 \sigma_2 \sigma_3}$$

$$\therefore \; b_{12.3} = \frac{\sigma_1 \sigma_3 (r_{12} - r_{13} \cdot r_{23})}{\sigma_2 \sigma_3 (1 - r_{23}^2)}$$

$$= \frac{\sigma_1}{\sigma_2} \cdot \left(\frac{r_{12} - r_{13} \cdot r_{23}}{1 - r_{23}^2} \right) \quad \ldots(x$$

and $$b_{13.2} = \frac{\sigma_1 \sigma_2 (r_{13} - r_{12} \cdot r_{23})}{\sigma_2 \sigma_3 (1 - r_{23}^2)}$$

$$= \frac{\sigma_1}{\sigma_3} \cdot \left(\frac{r_{13} - r_{12} \cdot r_{23}}{1 - r_{23}^2} \right) \quad \ldots(x$$

Substituting these values in (xvi), we get the req equation of the plane of regression of X_1 on (X_2 , X_3)

Remark : Note on Cramer's Rule

Let, $a_1 x + b_1 y + c_1 = 0$; $a_2 x + b_2 y + c_2 = 0$

but $$\frac{x}{b_1 c_2 - b_2 c_1} = \frac{-y}{a_1 c_2 - a_2 c_1} = \frac{1}{a_1 b_2 - a_2 b_1} \; ; \text{ or}$$

$$\frac{x}{b_1 c_2 - b_2 c_1} = \frac{-y}{a_2 c_1 - a_1 c_2} = \frac{1}{a_1 b_2 - a_2 b_1}$$

$$\therefore \; x = \frac{b_1 c_2 - b_2 c_1}{a_1 b_2 - a_2 b_1} \; ;$$

$$y = \frac{a_2 c_1 - a_1 c_2}{a_1 b_2 - a_2 b_1} \quad \ldots(xix)$$

3.6.6 Simpler Form of the Equation of the Plane of Regression

Let us write, $$W = \begin{vmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{vmatrix} \quad \ldots(xx$$

$$= 1. \begin{vmatrix} 1 & r_{23} \\ r_{12} & 1 \end{vmatrix} - r_{12} \begin{vmatrix} r_{21} & r_{23} \\ r_{31} & 1 \end{vmatrix} + r_{13} \begin{vmatrix} r_{21} & 1 \\ r_{31} & r_{32} \end{vmatrix}$$

$$= \left(1 - r_{23}^2\right) - r_{12}\left(r_{12} - r_{13}\, r_{23}\right) + r_{13}\left(r_{12}\, r_{23} - r_{13}\right)$$

$$\therefore \quad W = 1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2\, r_{12}\, r_{13} \cdot r_{23} \qquad \ldots(a)$$

$$[\because r_{ii} = 1 \text{ and } r_{ij} = r_{ji}\ [i \neq j, = 1, 2, 3)] \qquad \ldots(b)$$

Now, Let W_{ij} = Cofactor of element in i^{th} row and j^{th} column of W.

= $(-1)^{1+j}$ [Determinant obtained on deleting the i^{th} row and j^{th} column of w] ...(xxiv)

Now, W_{11} = Cofactor of element in 1^{st} row and 1^{st} column

$$= (-1)^{1+1} \begin{vmatrix} 1 & r_{23} \\ r_{32} & 1 \end{vmatrix} = 1 - r_{23}^2 \qquad \ldots(1)$$

$$(\because r_{23} = r_{32})$$

$$W_{12} = W_{21} = (-1)^{1+2} \begin{vmatrix} r_{21} & r_{23} \\ r_{31} & 1 \end{vmatrix}$$

$$= -\left(r_{21} - r_{31} \cdot r_{23}\right) = \left(r_{13}\ r_{23} - r_{12}\right) \qquad \ldots(2)$$

$$(\because r_{21} = r_{12}, r_{31} = r_{13})$$

$$W_{13} = W_{31} = (-1)^{1+3} \begin{vmatrix} r_{21} & 1 \\ r_{31} & r_{32} \end{vmatrix} = r_{12}\, r_{23} - r_{13} \qquad \ldots(3)$$

$$W_{22} = (-1)^{2+2} \begin{vmatrix} 1 & r_{13} \\ r_{31} & 1 \end{vmatrix} = 1 - r_{13}^2 \qquad \ldots(4)$$

$$W_{33} = (-1)^{3+3} \begin{vmatrix} 1 & r_{12} \\ r_{21} & 1 \end{vmatrix} = 1 - r_{12}^2 \qquad \ldots(5)$$

Substituting Equations (1) and (5) in (xxi) and (xxii), the equation of plane of regression is,

$$X_1 = -\frac{\sigma_1}{\sigma_2} \cdot \frac{W_{12}}{W_{11}} X_2 - \frac{\sigma_1}{\sigma_3} \frac{W_{13}}{W_{11}} X_3$$

$$\therefore \quad \frac{X_1}{\sigma_1} W_{11} + \frac{X_2}{\sigma_2} W_{12} + \frac{X_3}{\sigma_3} W_{13} = 0 \qquad \ldots(xxv)$$

is the Equation of plane of regression.

3.6.7 Remarks

(1) If the variables X_1, X_2, X_3 are not measured from their respective means, then the equation of the plane of regression of X_1 on (X_2, X_3) is given by,

$$X_1 - \bar{X}_1 = b_{12.3}\left(X_2 - \bar{X}_2\right) + b_{13.2}\left(X_3 - \bar{X}_3\right) \qquad \ldots(xxvi)$$

$$\text{i.e.} \left(\frac{X_1 - \bar{X}_1}{\sigma_1}\right) W_{11} + \left(\frac{X_2 - \bar{X}_2}{\sigma_2}\right) W_{12} + \left(\frac{X_3 - \bar{X}_3}{\sigma_3}\right) W_{13} = 0 \qquad \ldots(xxvii)$$

(2) The equation of the plane of regression of X_2 on (X_1, X_3) is given by,

$$X_2 = b_{21.3} X_1 + b_{23.1} X_3 \qquad \ldots(xxviii)$$

and the values $b_{21.3}$ and $b_{23.1}$ are :

$$b_{21.3} = \frac{\sigma_2}{\sigma_1}\left(\frac{r_{12} - r_{13} \cdot r_{23}}{1 - r_{13}^2}\right)$$

$$= -\frac{\sigma_2}{\sigma_1}\left(\frac{W_{21}}{W_{22}}\right)$$

$$b_{23.1} = \frac{\sigma_2}{\sigma_3}\left(\frac{r_{23} - r_{12}\ r_{13}}{1 - r_{13}^2}\right)$$

$$= -\frac{\sigma_2}{\sigma_3}\left(\frac{W_{23}}{W_{22}}\right)$$

Substituting in (xxviii) ; we get

$$X_2 = -\frac{\sigma_2}{\sigma_1}\left(\frac{W_{21}}{W_{22}}\right) X_1 - \frac{\sigma_2}{\sigma_3}\left(\frac{W_{23}}{W_{22}}\right) X_3$$

$$\therefore \quad \frac{X_1}{\sigma_1} W_{21} + \frac{X_2}{\sigma_2} W_{22} + \frac{X_3}{\sigma_3} W_{23} = 0 \qquad \ldots(xxix)$$

Similarly, the equation of the plane of regression of X_3 on (X_1, X_2) is given by,

$$\frac{X_1}{\sigma_1} W_{31} + \frac{X_2}{\sigma_2} W_{32} + \frac{\sigma_3}{\sigma_3} W_{33} = 0$$

(3) By symmetry, the equation of the plane of regression, say X_i, on all other variables X_j; $(j \neq 1, 2, \ldots, n)$ is given by

$$\frac{X_1}{\sigma_1} W_{i1} + \frac{X_2}{\sigma_2} W_{i2} + \ldots + \frac{X_i}{\sigma_i} W_{ii} + \ldots + \frac{X_n}{\sigma_n} W_{in} = 0$$

3.6.8 Interpretation of Partial Regression Coefficients

For a tri-variate distribution with three variables X_1, X_2 and X_3, in the plane of regression of X_1 on X_2 and X_3. We have two partial regression coefficients, i.e. $b_{12.3}$ and $b_{13.2}$.

(i) $b_{12.3}$ represents the change in the value of the variable X_1 for a unit change in the value of the variable X_2, when the variable X_3 is kept constant.

(ii) $b_{13.2}$ represents the change in the value of the variable X_1 for a unit change in the variable X_3, when the variable X_2 is kept constant.

(iii) Similar interpretations can be given to other regression coefficients,

i.e. $b_{ij.k} : i \neq j \neq k = 12.3$

3.6.9 Solved Examples on Regression Equations

Ex. 3.6.1 : Let X_1, X_2 and X_3 be the excess of heights of father, mother and son respectively in 100 samples above their respective mean values in cm. A distribution of these variables gave the following correlation coefficients r_{ij} between X_i and X_j and standard deviations σ_i for $i, j = 1, 2, 3$.

$r_{12} = 0.3, \quad r_{23} = 0.4, \quad r_{31} = 0.5,$

$\sigma_1 = 3, \quad \sigma_2 = 2, \quad \sigma_3 = 4$

Obtain a regression equation of X_1 on X_2 and X_3, and estimate the excess of height of father when excess of heights of mother and son are 0.7 cm and 2.1 cm respectively.

Soln. :

Step I : Given Data is :

$r_{12} = 0.3, \quad r_{23} = 0.4, \quad r_{31} = 0.5,$

$\sigma_1 = 3, \quad \sigma_2 = 2, \quad \sigma_3 = 4$ ...(i)

Since X_1, X_2, X_3 denote the excess of heights of father, mother and son respectively above their respective mean values, they are measured from their means.

Hence, the equation of the plane of regression of X_1 on X_2 and X_3 is given by :

$$X_1 = b_{12.3} X_2 + b_{13.2} X_3 \quad ...(ii)$$

Step II : We have

$$b_{12.3} = \frac{\sigma_1}{\sigma_2}\left(\frac{r_{12} - r_{13}\, r_{23}}{1 - r_{23}^2}\right)$$

and $$b_{13.2} = \frac{\sigma_1}{\sigma_3}\left(\frac{r_{13} - r_{12} \cdot r_{32}}{1 - r_{32}^2}\right)$$

Substituting the given values from Equation (i) and noting that $r_{ij} = r_{ji}$; we get

$$b_{12.3} = \frac{3}{2}\left[\frac{0.3 - 0.5 \times 0.4}{1 - (0.4)^2}\right]$$

$$= \frac{3 \times 0.10}{2 \times 0.84}$$

$$= \frac{10}{56} = 0.1786$$

and $$b_{13.2} = \frac{3}{4}\left[\frac{0.5 - 0.3 \times 0.4}{1 - (0.4)^2}\right]$$

$$= \frac{3 \times 0.38}{4 \times 0.84} = \frac{19}{56}$$

$$= 0.3393$$

Step III : Substituting in Equation (ii), the equation of regression of X_1 on X_2 and X_3 becomes

$$X_1 = 0.1786\, X_2 + 0.3393\, X_3 \quad ...(iii)$$

The estimate of the height of father when the excess of heights of mother and son are 0.7 cm and 2.1 cm respectively, is given by

$$\hat{X}_1 = (0.1786 \times 0.7) + (0.3393 \times 2.1)$$

$$= 0.12502 + 0.71253$$

$$= 0.83755 \text{ cm}$$

$$\therefore \hat{X}_1 = \mathbf{0.83755 \text{ cm}}.$$

Ex. 3.6.2 : From heights (X_1) in inches, weights (X_2) in kg. and ages (X_3) in years of a group of students, the following means, variances and correlation coefficients were obtained :

$\bar{X}_1 = 40, \quad \bar{X}_2 = 50, \quad \bar{X}_3 = 20$;

$S_1 = 3, \quad S_2 = 4, \quad S_3 = 2$;

$r_{12} = 0.4$, $r_{23} = 0.5$, $r_{13} = 0.25$

where $\bar{X}_i$ is the mean of X_i, S_i^2 is the variance of X_i and r_{ij} is correlation coefficient between X_i and X_j for i, j = 1, 2, 3. Find the multiple regressive equation of X_3 (on X_1 and X_2) and estimate the value of X_3 when X_1 = 43 inches, X_2 = 54 kg.

☑ **Soln. :**

▶ **Step I** : The multiple regression equation of X_3 on X_1 and X_2 is given by :

$$X_3 - \bar{X}_3 = b_{31.2}\left(X_1 - \bar{X}_1\right) + b_{32.1}\left(X_2 - \bar{X}_2\right) \quad \ldots(i)$$

Given data : $\bar{X}_1 = 40$, $\bar{X}_2 = 50$, $\bar{X}_3 = 20$;

S_1 = Std. deviation = 3, $S_2 = 4$, $S_3 = 2$

$r_{12} = 0.4, \quad r_{23} = 0.5, \quad r_{13} = 0.25$ ...(ii)

▶ **Step II :**

Now, $$b_{31.2} = \frac{S_3}{S_1}\left(\frac{r_{31} - r_{32}\, r_{12}}{1 - r_{12}^2}\right) \text{ and}$$

$$b_{32.1} = \frac{S_3}{S_2}\left(\frac{r_{32} - r_{31}\, r_{21}}{1 - r_{21}^2}\right) \quad \ldots(iii)$$

Substituting the given values of r_{ij}'s and S_i's in Equation (iii) from Equation (ii) and noting $r_{ij} = r_{ji}$; we get

$$b_{31.2} = \frac{2}{3}\left[\frac{0.25 - 0.5 \times 0.4}{1 - (0.4)^2}\right]$$

$$= \frac{2}{3}\left[\frac{0.25 - 0.20}{1 - 0.16}\right] = \frac{2 \times 0.05}{3 \times 0.84}$$

$$= 0.0397 = 0.04$$

and $$b_{312.1} = \frac{2}{4}\left[\frac{0.5 - 0.25 \times 0.4}{1 - (0.4)^2}\right]$$

$$= \frac{2}{4}\left[\frac{0.5 - 0.1}{1 - 0.16}\right] = \frac{1 \times 0.4}{2 \times 0.84}$$

$$= 0.2381 = 0.24$$

▶ **Step III :** Substituting these values in Equation (i), the required equation of regression of X_3 on X_1 and X_2 becomes :

$$X_3 - 20 = 0.04\,(X_1 - 40) + 0.24\,(X_2 - 50)$$

$$\therefore \quad X_3 = 0.04\,X_1 + 0.24\,X_2 + (20 - 0.04 \times 40 - 0.24 \times 50)$$

$$\therefore \quad X_3 = 0.04\,X_1 + 0.24\,X_2 + 6.4$$

The estimated value of X_3 when X_1 = 43 inches and X_2 = 54 kg is given by

$$\hat{X}_3 = 0.04 \times 43 + 0.24 \times 54 + 6.4 = 17.2 + 12.9 + 6.40 = \mathbf{21.08 \text{ years}}$$

$$= \hat{X}_3 = 21.08 \text{ years}$$

Ex. 3.6.3 : In a three variate (X_1, X_2, X_3) multiple correlation analysis, the following results were found.

X_i = mean , S_i = s.d. of X_i respectively and

r_{ij} = Correlation coefficient between X_i and X_j :

(i, j) = 1, 2, 3 in a sample of size 20.

$r_{12} = 0.6$; $r_{23} = 0.4$, $r_{13} = 0.5$;

$\bar{X}_1 = 60$; $\bar{X}_2 = 70$; $\bar{X}_3 = 80$; $S_1 = 3$, $S_2 = 4$, $S_3 = 5$

Find the regression line of X_1 on X_2 and X_3. Also find X_1 when $X_2 = 74$ and $X_3 = 85$

☑ **Soln. :**

▶ **Step I :** Given data is :

$\bar{X}_1 = 60$; $\bar{X}_2 = 70$; $\bar{X}_3 = 80$; $S_1 = 3$, $S_2 = 4$, $S_3 = 5$

$r_{12} = 0.6$; $r_{23} = 0.4$, $r_{13} = 0.5$; ...(i)

The line of regression of X_1 on X_2 and X_3 is,

$$X_1 - \bar{X}_1 = b_{12.3}\left(X_2 - \bar{X}_2\right) + b_{13.2}\left(X_3 - \bar{X}_3\right) \quad \ldots(ii)$$

▶ **Step II :**

We have, $$b_{12.3} = \frac{S_1}{S_2}\left[\frac{r_{12} - r_{13}\, r_{23}}{1 - r_{23}^2}\right]$$

$$= \frac{3}{4}\left[\frac{0.6 - (0.5)\,(0.4)}{1 - (0.4)^2}\right]$$

$$= \frac{3\,(0.60 - 0.20)}{4\,(0.84)} = \frac{120}{336}$$

$$= 0.3571 \quad \ldots(iii)$$

and $$b_{13.2} = \frac{S_1}{S_3}\left[\frac{r_{13} - r_{12} \cdot r_{32}}{1 - r_{32}^2}\right]$$

$$= \frac{3}{5}\left[\frac{0.5-(0.6)(0.4)}{1-(0.4)^2}\right]$$

$$= \frac{3(0.50-0.24)}{5(1-0.16)}$$

$$= \frac{3\times 0.26}{5\times 0.84} = \frac{78}{420} = 0.1851 \quad \ldots(iv)$$

Step III : Substituting the values from Equations (i), (iii), (iv) in Equation (ii), the line of regression of X_1 on X_2 and X_3 becomes :

$$X_1 - 60 = 0.3571\,(X_2 - 70) + 0.1857\,(X_3 - 80)$$

$$= 0.3571\,X_2 + 0.1857\,X_3 - 24.997 - 14.856$$

$$\therefore \quad X_1 = 0.3571X_2 + 0.1857\,X_3 + 20.147 \quad \ldots(v)$$

Taking $X_2 = 74$ and $X_3 = 85$ in Equation (v), the estimated value of X_1 is given by :

$$\hat{X}_1 = 0.3571 \times 74 + 0.1857 \times 85 + 20.147$$

$$= 26.4254 + 19.7845 + 20.147$$

$$\therefore \quad \hat{X} = 62.3569$$

3.7 SUPERVISED LEARNING : SUPPORT VECTOR MACHINE

UQ. Define Support Vector Machine. Explain how margin is computed and optimal hyper-plane is decided ? **(MU - Dec. 19, 10 Marks)**

3.7.1 Optimal Decision Boundary

- A solution to the classification problem is a rule that partitions the features and assigns each and all the features and partition to the same class.
- The **'boundary'** of this partitioning is the decision boundary of the rule.
- The boundary that this rule produces is the optimal decision boundary.
- A decision boundary is the region of a problem space in which the output label of a **classifier** is **ambiguous**.
- If the decision surface is a hyper plane, then the classification problem is linear, and the classes are **linearly separable**.
- Decision boundaries are not always clear-cut. That is, the transition from one class in the **feature space** to another is not discontinuous, but gradual.

GQ. What is decision boundary in decision tree ?

The first node of the tree called the "root node" contains the number of instances of all the classes respectively.

Basically, we have to draw a line called "decision boundary" that separates the instances of different classes into different regions called "decision regions".

GQ. What is decision boundary ? How do you draw it and what is the advantage of it ?

A decision boundary is a line (in the case of two features), where all (or most) samples of one class are on one side of that line, and all samples of the other class are on one side of that line, and all samples of the other class are on the opposite side of the line. The line separates the two classes.

Maximum Margin Linear Separators

UQ. What is SVM ? Explain the following terms: separating hyper plane, margin and support vectors with suitable example. **(MU - May 15, 4 Marks)**

UQ. What are the key terminologies of Support Vector Machine ? **(MU - May 16, 5 Marks)**

UQ. What is Support Vector Machine ? Illustrate Support Vector machine with neat labeled sketch. **(MU - May 17, Dec. 19, 4 Marks)**

UQ. **(MU - May 19, 4 Marks)**

(1) Support Vector Machine is a type of supervised learning that can be used for classification or regression. Even if the data points are unseen (not from the training dataset), support vector machine classifies the data properly.

2) Let's take an example of dataset that belongs to two different categories, and the distribution of data is proper means the data is separated from each other properly.

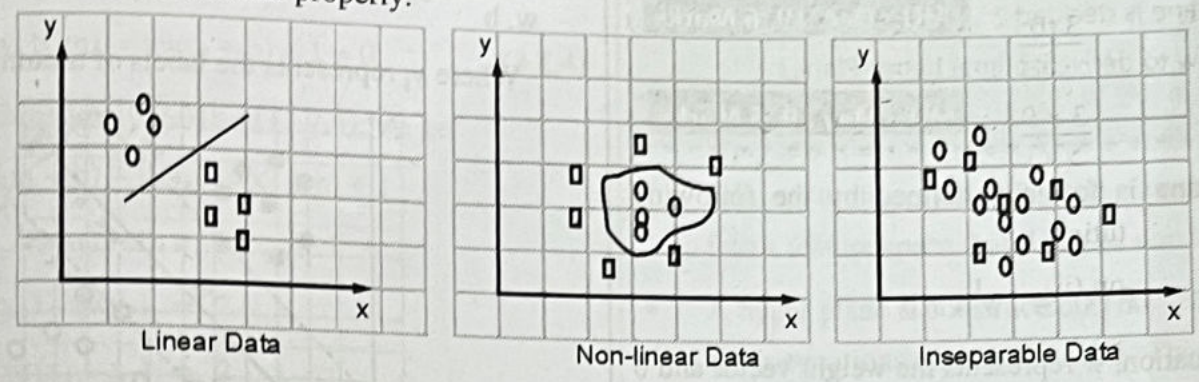

Fig. 3.7.1 : Different types of data

3) In this case we can draw a straight line (decision boundary) on the graph in such a way that the input space is divided in to two regions.

4) Data points that belongs to one category lies on one side of the decision boundary and the data points of other category lies on the opposite side. Such type of data is called as linearly separable data.

5) **Separating hyper plane** is the line which is used to separate the dataset. If we are using simple 2-dimensional plots then it's just a line. We require a plane to separate the data if data is 3 dimensional. So we can say that if data is N dimensional, we require N-1 dimensional hyper plane.

6) We want that our classifier should be designed in a manner that if a data point is far away from the decision boundary then we will be more confident about the prediction we have made.

7) We would like to find the data point near to the separating hyper plane and also make sure that this point should be far away from the separating line as possible. This is called as **margin**. We would like to find the greatest possible margin, because if we trained our classifier on limited data or made a mistake, we would want it to be as robust as possible.

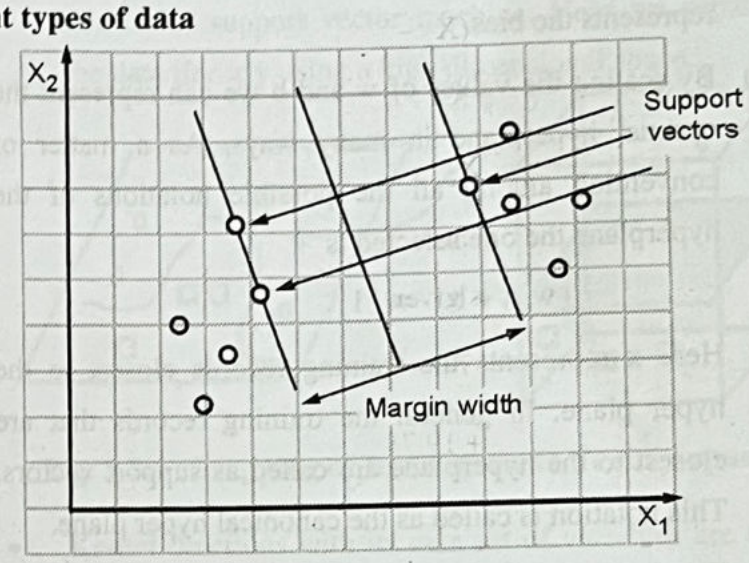

Fig. 3.7.2 : Support vectors and Margin

(8) **Support vectors** are the points which are nearest to the separating hyper plane. We have to maximize the distance between the support vectors and the separating line.

Distance between a hyper plane (w, b) and a point x is calculated as,

$$\text{Distance} = \frac{|w^t x + b|}{\|w\|}$$

3.7.2 Quadratic Programming Solution to Find Maximum Margin Separator

UQ. Explain the term : Hyperplane with suitable example. **(MU - May 15, 1 Marks)**

UQ. Write detail notes on : Quadratic Programming solution for finding maximum margin separation in support vector machine. **(MU - May 16, 10 Mark)**

UQ. How to compute the margin ? **(MU - May 17, 6Marks)**

Machine Learning (MU - Sem 6 - ECS & AIDS)

UQ. Explain how margin is computed and optimal hyper-plane is decided ? (MU - Dec. 19, 6 Marks

UQ. Show how to derive optimal hyper-Plane? MU - May 19, 6 Marks

(1) A **hyperplane** is formally defined by the following notation as,

$$F(x) = w^t x + b$$

In above equation, w represents the weight vector and b represents the bias.

(2) By scaling the values of w and b we can represent the optimal hyperplane in many ways. As a matter of convention among all the possible notations of the hyperplane the one selected is

$$|w^t x + b| = 1$$

(3) Here x represents the training records closest to the hyper plane. In general the training records that are closest to the hyperplane are called as support vectors. This notation is called as the canonical hyper plane.

(4) The distance between a point x and a hyper plane (w, b) is given by the result of geometry as follows,

$$\text{Distance} = \frac{|w^t x + b|}{\|w\|}$$

(5) In general, the numerator is equal to one for the canonical hyperplane and distance to the support vector is given as,

$$\text{Distance}_{sv} = \frac{|w^t x + b|}{\|w\|} = \frac{1}{\|w\|}$$

(6) Margin is twice the distance to nearest samples

$$M = \frac{2}{\|w\|}$$

(7) Ultimately, the task of maximizing M is same as compared to the task of minimizing a function L (w) subject to some conditions. The conditions used to model the requirements for correct classification of all training samples x_i by the hyper plane are formally stated as,

$$\min_{w, b} L(w) = \frac{1}{2}\|w\|^2 \text{ subject to } y_i (w^t x + b) \geq 1 \text{ for all } i.$$

Where y_i represents the labels of training.

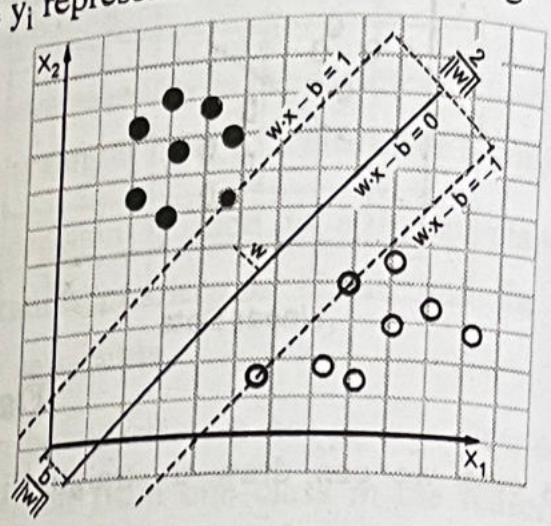

Fig. 3.7.3 : Solution to find maximum margin

(8) This is a problem of Lagrangian optimization that can be solved using Lagrange multiplier to calculate weight vector 'w' and the bias 'b' of the optimal hyper plane.

(9) Let's assume that we have 2 classes of 2 dimensional data to separate. Let's also assume that each class consist of only one point.

These points are,

$$X_1 = A_1 = (3, 3) \; ; \qquad X_2 = B_1 = (6, 6)$$

Find the hyper plane that separates theses 2 classes

$$f(w) = \frac{1}{2}\|w\|^2$$

(10) The constraints are,

$$c_i (w, b) = y_i | wx_i + b | - 1 \geq 0$$

$$c_1 (w, b) = 1 | wx_1 + b | - 1 \geq 0$$

$$c_2 (w, b) = -1 | wx_2 + b | - 1 \geq 0$$

(11) Next, we put equation into form of Lagrangian

$$L(w, b, m) = f(w) - m_1 c_1 (w, b) - m_2 c_2 (w, b)$$

$$= \frac{1}{2}\|w\|^2 - m_1 ((wx_1 + b) - 1) - m_2 (-(wx_2 + b) - 1)$$

$$= \frac{1}{2}\|w\|^2 - m_1 ((wx_1 + b) - 1) + m_2 ((wx_2 + b) + 1)$$

(12) We solve for the gradient of Lagrangian

$$\nabla L (w, b, m) = \nabla f(w) - m_1 \nabla c_1 (w, b) + m_2 \nabla c_2 (w, b) = 0$$

$$\frac{\partial}{\partial w} L (w, b, m) = w - m_1 x_1 + m_2 x_2 = 0 \quad ...(3.7.1)$$

$$\frac{\partial}{\partial b} L (w, b, m) = -m_1 + m_2 = 0 \quad ...(3.7.2)$$

$$\frac{\partial}{\partial\lambda_1} L(w, b, m) = (wx_1 + b) - 1 = 0 \quad ...(3.7.3)$$

$$\frac{\partial}{\partial\lambda_2} L(w, b, m) = (wx_2 + b) + 1 = 0 \quad ...(3.7.4)$$

(13) Equating Equation (3.7.3) and (3.7.4), we get

$$(wx_1 + b) - 1 = (wx_2 + b) + 1$$

$$\therefore (wx_1) - 1 = (wx_2) + 1$$

$$(wx_1) - (wx_2) = 2$$

$$\therefore w(x_1 - x_2) = 2$$

w is divided into parts as,

$$w = (w_1, w_2)$$

$$\therefore W(x_1 - x_2) = 2$$

$$(w_1, w_2)[(3, 3) - (6, 6)] = 2$$

$$\therefore (w_1, w_2)[(-3, -3)] = 2$$

$$-3w_1 - 3w_2 = 2$$

$$\therefore w_1 = -(0.67 + w_2) \quad ...(3.7.5)$$

(14) Adding values to Equation (3.7.1) and combining with Equation (3.7.2)

$$(w_1, w_2) - m_1(1, 1) + m_2(2, 2) = 0$$

From Equation (3.7.2)

$$m_1 = m_2;$$

$$(w_1, w_2) - m_1(3, 3) + m\lambda_1(6, 6) = 0$$

$$(w_1, w_2) + m_1(3, 3) = 0;$$

$$w_1 + 3m_1 = 0 \quad ...(3.7.6)$$

$$w_2 + 3m_1 = 0 \quad ...(3.7.7)$$

(15) Equating these we get, $w_1 = w_2$

Putting this in Equation (3.7.5)

$$w_1 = w_2 = -0.34$$

Putting this in either Equation (3.7.6) or Equation (3.7.7) will give

$$m_1 = m_2 = 0.11$$

(16) And finally, using this in Equation (3.7.3) and Equation (3.7.4)

$$b = 1 - (wx_1) \text{ or } = -1 - (w x_2)$$

$$= 1 - ((-0.34, -0.34), (3, 3)) \text{ or}$$

$$= 1 - ((-0.34, -0.34), (6, 6)) = 3.04$$

3.7.3 Kernels for Learning Non-Linear Functions

- Linear classifiers are able to separate only linearly separable data. Support vector machine provides the solution to this problem by transforming an input space into a feature space that contains non linear features.
- A hyper plane is constructed in the feature space so that other equations remain the same. This is also known as non-linear support vector machine. Here we separate the data linearly using a high dimensional space.

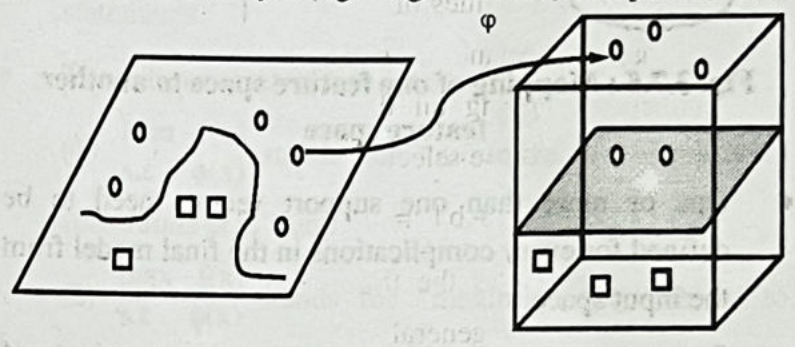

Fig. 3.7.4 : Mapping of Input space to Feature space

- Kernel functions with its own set of variables are used for this purpose. The result is going to be non linear if we convert this back to the original feature space. Particularly, the data is pre-processed with

$$X \rightarrow \Phi(x)$$

And then $\Phi(x)$ is mapped to y $F(x) = w\,\Phi(x) + b$

- Kernel function transforms the data in to an easily understandable form. This is done via mapping input space to another feature space.
- In support vector machine inner product is calculated of two vectors and the result of this is always a single number. When we replace this inner product by a kernel it is called as kernel trick.
- There are different algorithms that use different kinds of kernel functions.
- Among the different types of kernel functions such as nonlinear, linear, radial basis function (RBF), polynomial, and sigmoid, RBF is mostly used. The reason for this is along the X-axis radial basis function gives the localized and finite response.

- The number of support vectors will be determined based on the different criteria's such as what is the complexity of the model, how much slack is allowed.

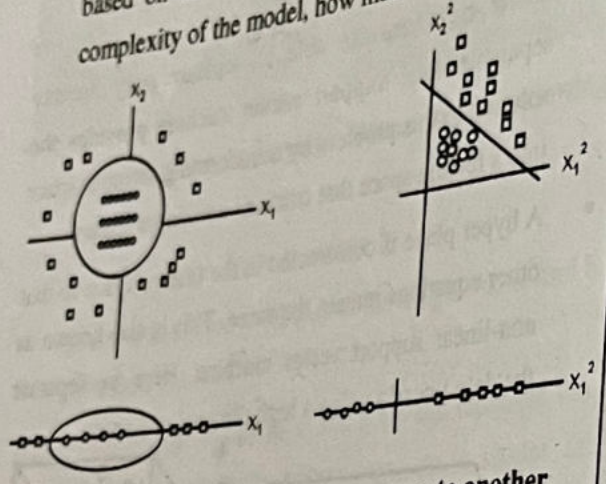

Fig. 3.7.5 : Mapping of one feature space to another feature space

- One or more than one support vectors need to be defined for every complications in the final model from the input space.
- Support vector machines output compromises of support vectors and alpha. This is used to specify the effect of support vectors on the final decision.
- If we select the model with high complexity it will result in to over fitting. For better generalization if large margin is selected then it may lead to incorrect classification.
- And accuracy depends on the trade-off between these two selections criteria. If we over fit the data then the range of support vectors may vary from very less to each single point. This tradeoff is controlled through the selection of kernel and its parameters.
- In support vector machine the data points are tested by taking the dot product of each support vector with the test point.
- Hence the computational complexity increases, if we increase the number of support vectors. Classification of test points will be faster if we have less number of support vectors.

3.7.4 Rules for the Kernel Function

Kernel function or a window is defined as follows:

If $\|\bar{x}\| \leq 1$ then $K(\bar{x}) = 1$ else 0.

This kernel function is shown by the Fig. 3.7.6,

$$K((z - x_i)/h) = 1$$

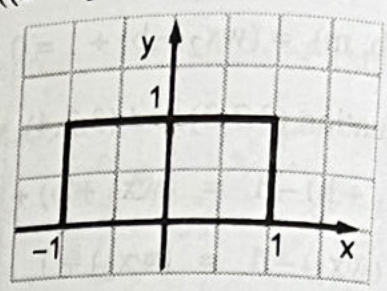

Fig. 3.7.6

For a fixed value of x_i, the function takes the value as 1 as shown in Fig. 3.7.7.

By selecting the argument of K (·), window can be moved to be centred at the point x_i and to be of radius h.

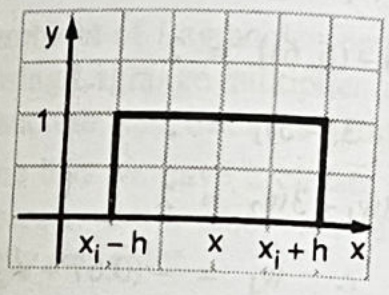

Fig. 3.7.7

3.8 DIFFERENT TYPES OF SVM KERNELS

1. Polynomial kernel

Polynomial kernel is mostly used in image processing methods.

Polynomial Kernel is represented as,

$$K(x_j, x_k) = (x_j \cdot x_k + 1)^p$$

Here p represents the degree of the polynomial.

2. Gaussian kernel

There are some applications where prior knowledge is not available. For this type of applications Gaussian kernel is used.

Gaussian kernel is defined as,

$$K(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right)$$

3. Gaussian radial basis function (RBF)

This is also used for the applications where prior knowledge is not available.

Gaussian radial basis function is defined as,

$$K(x_i, x_j) = \exp(-\gamma \| x - y \|^2) \text{ for } \gamma > 0$$

Sometimes it is parametrized using the value of γ as $1/2\sigma^2$

4. Laplace RBF kernel

Laplace RBF kernel is defined as,

$$K(x, y) = \exp\left(-\frac{\| x - y \|}{\sigma}\right)$$

5. Hyperbolic tangent kernel

Hyperbolic tangent kernel is used in neural networks.

It is defined as,

$$K(x_i, x_j) = \tanh(kx_i * x_j + c), \text{ for some (not every) } k > 0 \text{ and } c < 0.$$

6. Sigmoid kernel

Sigmoid kernel can be used as a proxy for neural networks. It is defined as,

$$K(x, y) = \tanh(\alpha x^d y + c)$$

7. Bessel function of the first kind Kernel

Cross terms in mathematical functions can be removed by using this type of kernel function.

It is defined as,

$$K(x, y) = \frac{J_{v+1}(\sigma \| x - y \|)}{\| x - y \|^{-n(v+1)}}$$

Here j represents the Bessel function of first type.

8. ANOVA radial basis kernel

In regression problems this kernel can be used.

It is defined as,

$$K(x, y) = \sum_{k=1}^{n} \exp(-\sigma (x^k - y^k)^2)^d$$

3.9 SVM as CONSTRAINED OPTIMISATION PROBLEM

- Constrained optimisation problem are problems for which a function f(x) is to be minimized or maximized subject to the given constraints $\phi(x)$.
- Here $f : R^n \rightarrow R$ is called the objective function and $\phi(x)$ is a Boolean valued formula.
- The constraints $\phi(x)$ can be an arbitrary Boolean combination of equations $g(x) = 0$, weak inequalities $g(x) \geq 0$, strict inequalities $g(x) > 0$, and $x \in Z$ statements.
- The following notations are used :

(1) $\begin{matrix} \text{Min} \cdot f(x) \\ \text{s.t} \quad \phi(x) \end{matrix}$ stands for "minimize f(x) subject to constraints $\phi(x)$", and

(2) $\begin{matrix} \max \cdot f(x) \\ \text{s.t} \quad \phi(x) \end{matrix}$ stands for "maximize f(x) subject to constraints $\phi(x)$".

- We say a point $u \in R^n$ satisfies the constraints ϕ it $\phi(u)$ is true.

3.9.1 Global Optimisation

- A point $u \in R^n$ is said to be a global minimum of f subject to constraints ϕ if u satisfies the constraints and for any point v that satisfies the constraints,

$$f(u) \leq f(v)$$

- A value $a \in R$ is said to be the global minimum valve of f subject to constraints and for any point v that satisfies the constraints, $a \leq f(v)$.
- The global minimum valve a exists for any f and ϕ. The global minimum value a is attained if there exists a point u such that $\phi(u)$ is true and $f(u) = a$. Such a point u is necessarily a global minimum.
- If f is a continuous function and the set of points satisfying the constraints ϕ is compact (i.e. closed and bounded) and non empty, then a global minimum exists. Otherwise a global minimum may or may not exist.

3.9.2 Examples

1. Minimise $[\{x, x^2 + y^2 < 1\}, (x, y)]$

   Here minimum value is not attained. The set of points satisfying the constraints is not closed :

   $\because \{-1, \{x \to -1, y \to 0\}\}$

2. Minimise $[\{x^2 + (y-1)^2\ y > x^2\}, (x, y)]$

   $\{0, (x \to 0, y \to 1)\}$

3.9.3 Local Optimisation

- A point $u \in R^n$ is said to be a local minimum of f subject to constraints ϕ if u satisfies the constraints and for some $r > 0$, if v satisfies $|v - u| < r \wedge \phi$ (v), then $f(u) \leq f(v)$.

Remark : *A local minimum may not be a global minimum.*

- *A global minimum is always a local minimum.*

Examples

Find a local minimum that is not a global minimum :

(i) $\{3x^4 - 28x^3 + 84x^2 - 96x + 42, (x, 1\text{-}1)\}$

Ans. : $\{5, (x \to 1)\}$

(ii) Minimise

$[3x^4 - 28x^3 + 84x^2 - 96x + 42, \{x\}]\ \{-22, (x \to 4)\}$

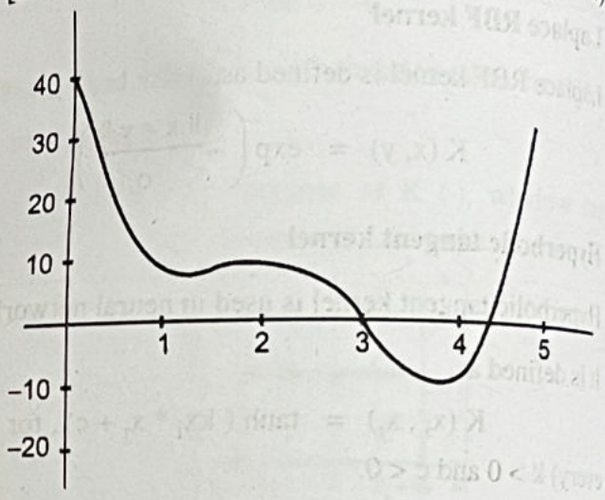

Fig. 3.9.1

Chapter Ends