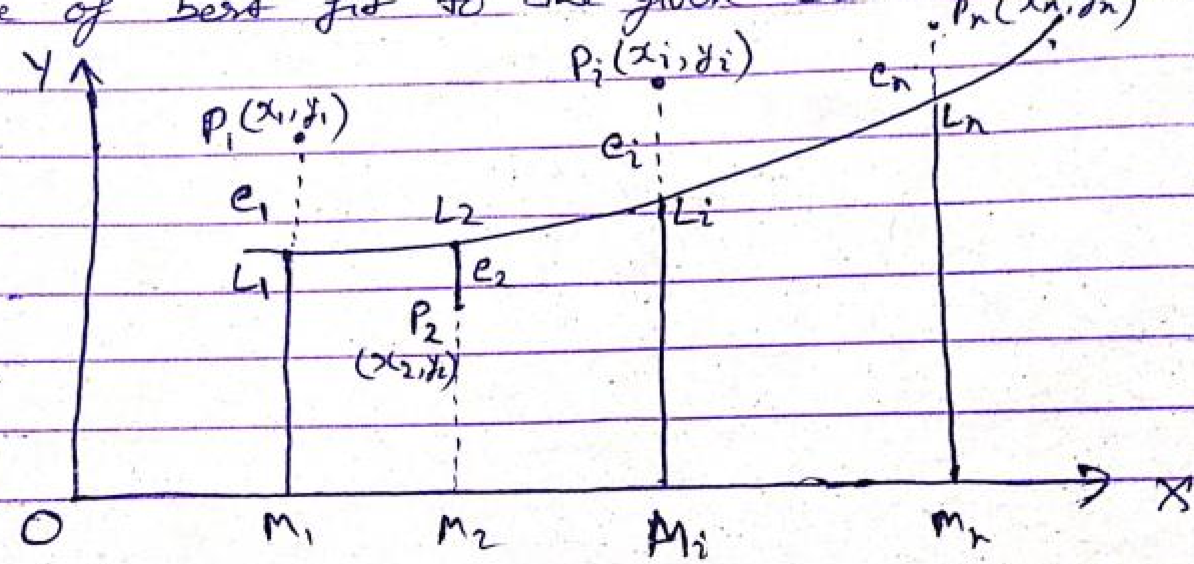
CURVE FITTING

Let there be two variables x and y which give us a set of n pairs of numerical values $(x_1, y_1), (x_2, y_2)$.......(x_n, y_n). In order to have an approximate idea about the relationship of these two variables, we plot these n paired points on a graph, thus we get a diagram showing the simultaneous variation in values of both the variables called *scatter or dot diagram*. From scatter diagram, we get only an approximate non-mathematical relation between two variables. *Curve fitting* means an exact relationship between two variables by algebraic equations. In fact, this relationship is the equation of the curve. Therefore, *curve fitting* means to form an equation of the curve from the given data. Curve fitting is considered of immense importance both from the point of view of theoretical and practical statistics.

Principle of Least Squares :- The method of least square is ~~probably~~ the most systematic procedure to fit a unique curve through the given data points. i.e. Principle of least squares provides a unique set of values to the constants and hence, suggests a curve of best fit to the given data.



Let the curve

$$y = a + bx_i + cx_i^2 + \dots \dots k x_i^n$$

be fitted to the set of n data points $(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_n, y_n)$.

At $(x = x_i)$ the observed (or experimental) value of the ordinate is $y_i = P_iM_i$ and the corresponding

Expected value or calculated value $= L_iM_i = f(x_i)$

The difference of the observed and the expected value is $P_iM_i - L_iM_i = e_i$. This difference is called error at $(x = x_i)$ clearly some of the error $e_1, e_2, e_3, \dots e_i, \dots e_n$ will be positive or negative. To make all errors positive, we square each of the errors, i.e.

$$E = e_1^2 + e_2^2 + e_3^2 + \dots + e_i^2 + \dots + e_n^2$$

$$\boxed{E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - f(x_i)]^2}$$

The curve of best fit is that for which e's are as small as possible i.e. E, the sum of the square of the errors is a minimum this is known as the principle of least square.

Using this principle, we shall fit the following curves —

(i) A straight line, $y = a + bx$

(ii) A parabola, $y = a + bx + cx^2$

(iii) The exponential curve, $y = ae^{bx}$

(iv) The curve, $y = ax^b$

(v) The curve, $y = ab^x$

(vi) " $y = ax + bx^2$

(vii) " $y = pv^n = k$

$(y = a + bx)$

Fitting a Straight Line :-

where Let (x_i, y_i), $i = 1, 2, 4, \ldots n$ be n sets of observations of related data and

$$y = a + bx \quad \text{——— (1)}$$

be the straight line to be fitted. The error at $x = x_i$ is

$$E.V - C.V$$

$$E_i = y_i - f(x_i)$$

$$e_i = y_i - (a + bx_i)$$

$$E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

By the principle of Least Squares, E is minimize

$$\frac{\partial E}{\partial a} = 0 \quad \text{and} \quad \frac{\partial E}{\partial b} = 0$$

$$\therefore 2\sum_{i=1}^{n} (y_i - a - bx_i)(-1) = 0$$

$$\Rightarrow 2\sum_{i=1}^{n} (a + bx_i - y_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i$$

$$\Rightarrow \boxed{\sum y = na + b\sum x} \quad \text{——— (2)}$$

and
$$2\sum_{i=1}^{n} (y_i - a - bx_i)(-x_i) = 0$$

$$\Rightarrow 2\sum_{i=1}^{n} (-x_i y_i + ax_i + bx_i^2) = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2$$

$$\Rightarrow \boxed{\sum xy = a\sum x + b\sum x^2} \quad \text{——— (3)}$$

The eq^n (2) and (3) are known as normal eq^ns. On solving eq^ns (2) and (3), we get the value of a and b. Putting the value of a and b in eq^n (1), we get the eq^n of the line of best fit.

Fitting a Parabola :-

Let a parabola

$y = a + bx + cx^2$ ——(1) which is fitted to a given data $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$

The error at $x = x_i$ is

$$e_i = y_i - f(x_i)$$

$$= y_i - (a + bx_i + cx_i^2)$$

$$\therefore \quad E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[y_i - (a + bx_i + cx_i^2) \right]^2 \quad \text{——(2)}$$

Now E should be minimum for the best values of a, b, and c.

By the principle of least squares, the value of E is minimum, therefore

$$\frac{\partial E}{\partial a} = 0, \quad \frac{\partial E}{\partial b} = 0 \quad \text{and} \quad \frac{\partial E}{\partial c} = 0$$

i.e.

$$\therefore \quad 2\sum_{i=1}^{n} (y_i - a - bx_i - cx_i^2)(-1) = 0$$

$$\Rightarrow \quad \sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2$$

$$\Rightarrow \quad \boxed{\sum y_i = na + b\sum x_i + c\sum x_i^2} \quad \text{--- (3)}$$

and

$$2\sum_{i=1}^{n} (y_i - a - bx_i - cx_i^2)(-x_i) = 0$$

$$2\sum_{i=1}^{n} (-x_i y_i + ax_i + bx_i^2 + cx_i^3) = 0$$

$$\Rightarrow \quad \boxed{\sum x_i y_i = a\sum x_i + b\sum x_i^2 + c\sum x_i^3} \quad \text{--- (4)}$$

and

$$\frac{\partial E}{\partial c} = 0$$

$$\Rightarrow \quad 2\sum_{i=1}^{n} (y_i - a - bx_i - cx_i^2)(-x_i^2) = 0$$

$$\sum_{i=1}^{n} (-x_i^2 y_i + ax_i^2 + bx_i^3 + cx_i^4) = 0$$

$$\Rightarrow \quad \boxed{\sum x_i^2 y_i = a\sum x_i^2 + b\sum x_i^3 + c\sum x_i^4}$$

These three eqns called the normal equations, can be solved for determining a, b, c. Putting the values of a, b, c in eqn (1), we get the eqn of the parabola of best fit for the given data point

Fitting an Exponential Curve :-

Consider the eqn

$$y = ae^{bx}$$

Taking logarithms on both sides, we get

$$\log_{10} y = \log_{10} a + bx \log_{10} e$$

i.e. $\boxed{Y = A + Bx}$ —— ①

where $Y = \log_{10} y$,

$A = \log_{10} a$ and $B = b \log_{10} e$

Thus, we see that fitting of an exponential curve of the form $y = ae^{bx}$ is equivalent to fitting of a straight line $Y = A + Bx$

The normal eqns for (1) are

$$\boxed{\Sigma Y = nA + B\Sigma x}$$

and $$\boxed{\Sigma xY = A\Sigma x + B\Sigma x^2}$$

Solving these, we get A and B

Then $a = \text{antilog } A$ and

$$b = \frac{B}{\log_{10} e}$$

Putting the values of a and b in eqn $y = ae^{bx}$, we get the eqn of the exponential curve of best fit for the given datapoints.

Fitting the curve $y = ax^b$

Taking logarithm on both sides, we get

$$\log_{10} y = \log_{10} a + b\log_{10} x$$

i.e. $\boxed{Y = A + bX}$ ——— (1)

where, $Y = \log y$, $A = \log a$ and $X = \log x$

The normal eq^n of (1) are

$$\boxed{\begin{aligned} \Sigma Y &= nA + b\Sigma X \\ \Sigma XY &= A\Sigma X + b\Sigma X^2 \end{aligned}}$$

which gives A and b on solving

and $a = \text{antilog } A$

Fitting the curve $y = ab^x$

Taking logarithm on both sides

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

$$\Rightarrow \quad Y = A + Bx \quad \text{———} \quad (1)$$

where $Y = \log y$; $A = \log a$, $B = \log b$

This is a linear eqn in Y and x
For estimating A and B, normal eqns are

$$\Sigma Y = nA + B\Sigma x$$

and $$\Sigma xY = A\Sigma x + B\Sigma x^2$$

which gives A and B on solving from which $a = \text{antilog} A$ and
$b = \text{antilog} B$
can be found. Substituting a and b in eqn $y = ab^x$, we have the required curve of the best fit for the given data points.

Fitting of the Curve $y = ax + bx^2$

Error of estimate for i^{th} point (x_i, y_i) is

$$e_i = (y_i - ax_i - bx_i^2)$$

$$E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - ax_i - bx_i^2)$$

By principle of least squares, the value of E is minimum, therefore

Normal equations are given by

$$\frac{\partial E}{\partial a} = 0 \quad \text{and} \quad \frac{\partial E}{\partial b} = 0$$

$$\Rightarrow \quad 2\sum_{i=1}^{n} (y_i - ax_i - bx_i^2)(-x_i) = 0$$

$$\Rightarrow \quad \sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3$$

or $\boxed{\Sigma xy = a\Sigma x^2 + b\Sigma x^3}$

and $$2\sum_{i=1}^{n} (y_i - ax_i - bx_i^2)(-x_i^2) = 0$$

$$\sum_{i=1}^{n} x_i^2 y_i = a\sum_{i=1}^{n} x_i^3 + b\sum_{i=1}^{n} x_i^4$$

$$\boxed{\Sigma x^2 y = a\Sigma x^3 + b\Sigma x^4}$$

Fitting of the curve $y = ax^2 + \frac{b}{x}$

Error of estimate for i^{th} point (x_i, y_i) is

$$e_i = \left(y_i - a x_i^2 - \frac{b}{x_i} \right)$$

$$E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - a x_i^2 - \frac{b}{x_i} \right)^2$$

By principle of least squares, the value of E is minimum, therefore

Normal eqns are given by

$$\frac{\partial E}{\partial a} = 0 \quad \text{and} \quad \frac{\partial E}{\partial b} = 0$$

$$\Rightarrow 2\sum_{i=1}^{n} \left(y_i - ax_i^2 - \frac{b}{x_i}\right)\left(-x_i^2\right) = 0$$

$$\sum_{i=1}^{n} x_i^2 y_i = a\sum_{i=1}^{n} x_i^4 + b\sum_{i=1}^{n} x_i$$

$$\Rightarrow \boxed{\sum x^2 y = a\sum x^4 + b\sum x}$$

$$\text{and} \quad 2\sum_{i=1}^{n} \left(y_i - ax_i^2 - \frac{b}{x_i}\right)\left(-\frac{1}{x_i}\right) = 0$$

$$\sum_{i=1}^{n} \frac{y_i}{x_i} = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} \frac{1}{x_i^2}$$

$$\Rightarrow \boxed{\sum \frac{y}{x} = a\sum x + b\sum \frac{1}{x^2}}$$

Fitting of the curve $pv^r = k$

$$pv^r = k$$

$$\Rightarrow \quad v = k^{1/r} p^{-1/r}$$

Taking logarithm on both sides

$$\log v = \frac{1}{r}\log k - \frac{1}{r}\log p$$

$$\Rightarrow \quad Y = A + BX$$

where $Y = \log v$, $A = \frac{1}{r}\log k$

$B = -\frac{1}{r}$ and $X = \log p$

Normal eqns are obtained as per that of the straight line.

$$\boxed{\begin{array}{l} \Sigma Y = nA + B\Sigma X \\ \hline \Sigma XY = A\Sigma X + B\Sigma X^2 \end{array}}$$

solving these, we ~~got~~ find A and B, then we find

$$r = -\frac{1}{B}$$

and then

$$\log k = rA$$

Example 1. *By the method of least squares, find the straight line that best fits the following data:*

x:	1	2	3	4	5
y:	14	27	40	55	68.

Sol. Let the straight line of best fit be

$$y = a + bx \tag{5}$$

Normal equations are $\Sigma y = ma + b\Sigma x$ (6)

and $\Sigma xy = a\Sigma x + b\Sigma x^2$ (7)

Here $m = 5$

The table is as below:

x	y	xy	x^2
1	14	14	1
2	27	54	4
3	40	120	9
4	55	220	16
5	68	340	25
$\Sigma x = 15$	$\Sigma y = 204$	$\Sigma xy = 748$	$\Sigma x^2 = 55$

Substituting in (6) and (7), we get

$$204 = 5a + 15b$$

$$748 = 15a + 55b$$

Solving, we get $a = 0$, $b = 13.6$

Hence required straight line is $\boxed{y = 13.6x}$

Example 3. *Determine the constants a and b by the Method of Least Squares such that $y = ae^{bx}$ fits the following data:*

x	2	4	6	8	10
y	4.077	11.084	30.128	81.897	222.62

Sol.
$$y = ae^{bx}$$

Taking log on both sides

$$\log y = \log a + bx \log e$$

or
$$Y = A + BX,$$

where
$$Y = \log y$$
$$A = \log a$$
$$B = b \log_{10} e$$
$$X = x.$$

Normal equations are

$$\Sigma Y = mA + B\Sigma X$$

and
$$\Sigma XY = A\Sigma X + B\Sigma X^2.$$

Here $m = 5$.

x	y	X	Y	XY	X^2
2	4.077	2	.61034	1.22068	4
4	11.084	4	1.04469	4.17876	16
6	30.128	6	1.47897	8.87382	36
8	81.897	8	1.91326	15.30608	64
10	222.62	10	2.347564	23.47564	100
		$\Sigma X = 30$	$\Sigma Y = 7.394824$	$\Sigma XY = 53.05498$	$\Sigma X^2 = 220$

Substituting these values in equations (22) and (23), we get

$$7.394824 = 5A + 30B$$

and $$53.05498 = 30A + 220B.$$

Solving, we get $$A = 0.1760594$$

and $$B = 0.2171509$$

$\therefore$ $$a = \text{antilog}\,(A)$$

$$= \text{antilog}\,(0.1760594) = 1.49989$$

and $$b = \frac{B}{\log_{10} e} = \frac{0.2171509}{.4342945} = 0.50001$$

Hence the required equation is

$$\boxed{y = 1.49989\, e^{0.50001x}.}$$

Example 4. *Obtain a relation of the form $y = ab^x$ for the following data by the Method of Least Squares:*

x	2	3	4	5	6
y	8.3	15.4	33.1	65.2	126.4

Sol. The curve to be fitted is $y = ab^x$

or $$Y = A + Bx,$$

where $A = \log_{10} a$, $B = \log_{10} b$ and $Y = \log_{10} y$.

$\therefore$ The normal equations are $\Sigma Y = 5A + B\Sigma x$

and $$\Sigma XY = A\Sigma x + B\Sigma x^2.$$

x	y	$Y = \log_{10} y$	x^2	xY
2	8.3	0.9191	4	1.8382
3	15.4	1.1872	9	3.5616
4	33.1	1.5198	16	6.0792
5	65.2	1.8142	25	9.0710
6	127.4	2.1052	36	12.6312
$\Sigma x = 20$		$\Sigma Y = 7.5455$	$\Sigma x^2 = 90$	$\Sigma xY = 33.1812$

Substituting the values of Σx, etc. from the above table in normal equations, we get

$$7.5455 = 5A + 20B \quad \text{and} \quad 33.1812 = 20A + 90B.$$

On solving A = 0.31 and B = 0.3

$$\therefore \qquad a = \text{antilog } A = 2.04$$

and $$b = \text{antilog } B = 1.995.$$

Hence the required curve is

$$\boxed{y = 2.04(1.995)^x.}$$

GRAPHICAL REPRESENTATION OF A FREQUENCY DISTRIBUTION

Representation of frequency distribution by means of a diagram makes the unwieldy data intelligible and conveys to the eye the general run of the observations. The graphs and diagrams have a more lasting effect on the brain. It is always easier to compare data through graphs and diagrams. Forecasting also becomes easier with the help of graphs. Graphs help us in interpolation of values of the variables.

However there are certain disadvantages as well. Graphs do not give measurements of the variables as accurate as those given by tables. The numerical value can be obtained to any number of decimal places in a table, but from graphs it can not be found to 2nd or 3rd places of decimals. Another disadvantage is that it is very difficult to have a proper selection of scale. The facts may be misrepresented by differences in scale.

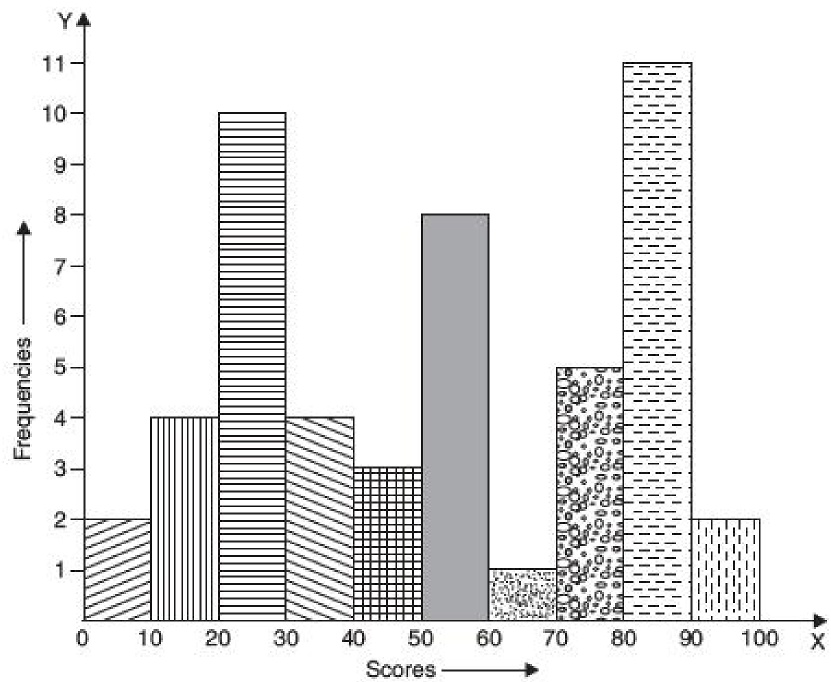
TYPES OF GRAPHS AND DIAGRAMS

Generally the following types of graphs are used in representing frequency distributions:

(1) Histograms, (2) Frequency Polygon, (3) Frequency Curve, (4) Cumulative Frequency Curve or the Ogive, (5) Historigrams, (6) Bar Diagrams, (7) Area

HISTOGRAMS

To draw the histograms of a given grouped frequency distribution, mark off along a horizontal base line all the class-intervals on a suitable scale. With the class-intervals as bases, draw rectangles with the areas proportional to the frequencies of the respective class-intervals. For equal class-intervals, the heights of the rectangles will be proportional to the frequencies. If the class-intervals are not equal, the heights of the rectangles will be proportional to the ratios of the frequencies to the width of the corresponding classes. A diagram with all these rectangles is a **Histogram.**

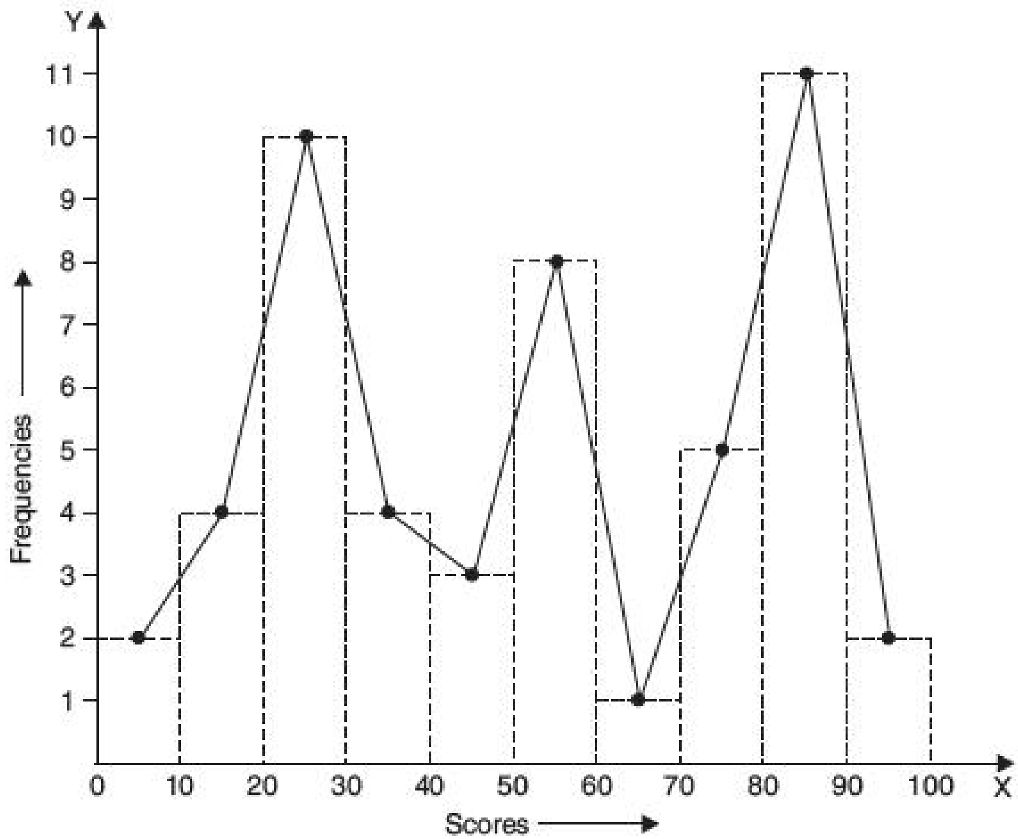


(Histogram for the previous table)

Histograms are also useful when the class-intervals are not of the same width. They are appropriate to cases in which the frequency changes rapidly.

FREQUENCY POLYGON

If the various points are obtained by plotting the central values of the class intervals as x co-ordinates and the respective frequencies as the y co-ordinates, and these points are joined by straight lines taken in order, they form a polygon called **Frequency Polygon.**



In a frequency polygon the variables or individuals of each class are assumed to be concentrated at the mid-point of the class-interval.

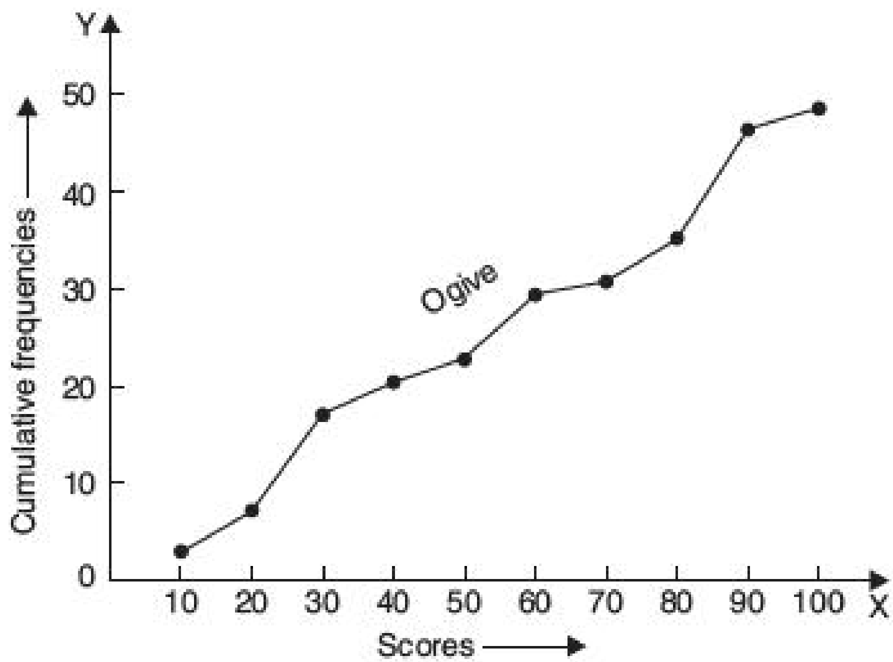
Here in this diagram dotted is the **Histogram** and a polygon with lines as sides is the **Frequency Polygon.**

FREQUENCY CURVE

If through the vertices of a frequency polygon a smooth freehand curve is drawn, we get the **Frequency Curve.** This is done usually when the class-intervals are of small widths.

CUMULATIVE FREQUENCY CURVE OR THE OGIVE

If from a cumulative frequency table, the upper limits of the class taken as x co-ordinates and the cumulative frequencies as the y co-ordinates and the points are plotted, then these points when joined by a freehand smooth curve give the **Cumulative Frequency Curve or the Ogive.**



Regression Analysis :-

Regression analysis is another technique that measures the quantitative relationship existing two variables. The fundamental difference between the problem of curve fitting and regression, if any, is that in regression, any of the variable may be considered as independent or dependent, while in curve fitting one variable cannot be dependent.

(Thus Regression analysis is a technique which refers to the functional relationship between x and y, and estimates the values of dependent variable y for given values of the independent variable x.) For example, relationship between total income of employees and total savings of a particular area, helps to estimate saving at a given values of income.

Linear and Non-Linear Regression :-

(If the given data are plotted on a graph, the points so obtained will concentrate round a curve, called the 'curve of regression' or non-linear regression.

If the regression curve is a straight line, it is called the line of regression and the regression is said to be linear regression.)

Lines of Regression :- The line of regression is the straight line, which gives the 'best fit' in the least square sense to the given frequency.

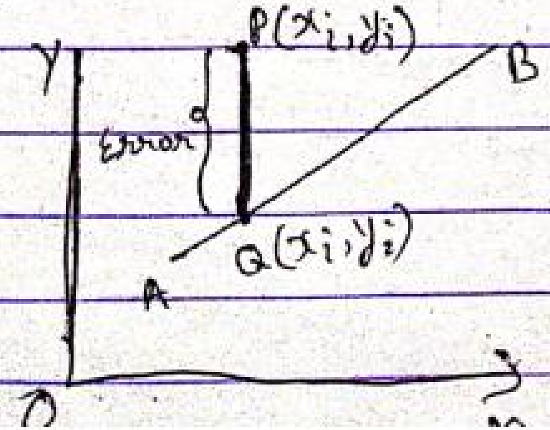
(If we wish to estimate y for given values of x, we shall have the regression equation of the form $y = a + bx$ is called the line of regression of y on x.

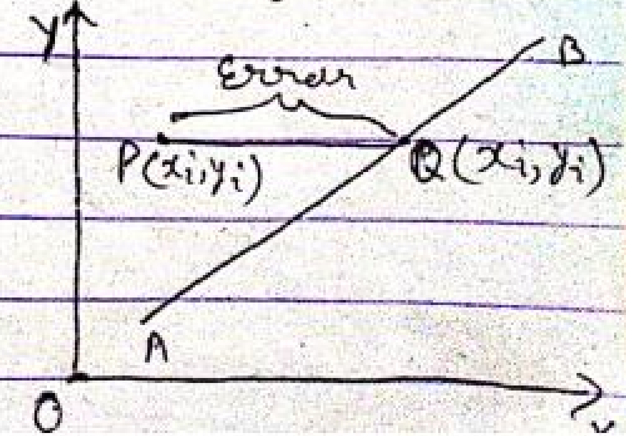
If we wish to estimate x for given values of y, we shall have the regression equation of the form $x = A + By$ is called the line of regression of x on y.

Thus it implies, in general, we always have two lines of regression.)

If the line of regression is so chosen that the sum of squares of deviation parallel to the axis of y is minimised, it is called the line of regression of y on x and it gives the best estimates of y for any given values of x.

If the line of regression is so chosen that the sum of squares of deviations parallel to the axis of x is minimised, it is called the line of regression of x on y and it gives the best estimates of x for any given value of y.





Simple and Multiple Regression :-

When two variables are involved, it is simple regression. There will be one dependent and one independent variables. If more than two variables are involved it is known as multiple regression. There will be one dependent variable and more than two independent variables.]

For example, the sales turnover of a product (a dependent variable) is associated with multiple independent variables such as price of the products, quality of the product, advertisement expenditure etc.

eg ⇒ Linear Regression ⇒ $Y = a + bX + c$

Multiple Regression ⇒ $Y = a + b_1X_1 + b_2X_2 + b_3X_3 + \cdots$

Derivation of Lines of Regression :-

Let $y = a + bx$ ——— ①

be the equation of the regression line of y on x

The error for i^{th} point is

$$e_i = y_i - (a + bx_i)$$

The sum of the squares of deviations of observed value y_i from the expected value is given by

$$E = \sum_{i=1}^{n} e_i^2$$

$$E = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \quad \text{———} ②$$

According to the method of least squares, we have to select 'a' and 'b' so that E is minimum. The conditions of E being minimum are

$$\frac{\partial E}{\partial a} = 0 \quad \text{and} \quad \frac{\partial E}{\partial b} = 0 \qquad \text{———} \ (3)$$

from eqn (2)

$$\frac{\partial E}{\partial a} = 2\sum_{i=1}^{n} (y_i - a - bx_i)(-1)$$

Now $\frac{\partial E}{\partial a} = 0$

$$\therefore \quad 2\sum_{i=1}^{n} (y_i - a - bx_i)(-1) = 0$$

$$\Rightarrow \quad \boxed{\Sigma y = na + b\Sigma x} \qquad \text{———} \ (4)$$

Also,

$$\frac{\partial E}{\partial b} = 2\sum_{i=1}^{n} (y_i - a - bx_i)(-x_i)$$

Now

$$\frac{\partial E}{\partial b} = 0$$

$$\therefore\ 2\sum_{i=1}^{n} (y_i - a - bx_i)(-x_i) = 0$$

$$\Rightarrow \boxed{\Sigma xy = a\Sigma x + b\Sigma x^2} \quad \text{---} (5)$$

eqns (4) & (5) are called normal eqns which on solving for 'a' and 'b', gives we get

$$b_{yx} = \frac{\Sigma xy - \frac{1}{n}\Sigma x \Sigma y}{\Sigma x_i^2 - \frac{1}{n}(\Sigma x)^2}$$

$$b_{yx} = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2} \quad \text{---} (6)$$

and $$a = \frac{\Sigma y}{n} - b\frac{\Sigma x}{n}$$

$$\boxed{a = \bar{y} - b\bar{x}} \quad \text{——— (7)}$$

Where $\bar{x}$ and $\bar{y}$ are the means of x and y values

Hence, $y = a + bx$ passes through point $(\bar{x}, \bar{y})$

Putting $a = \bar{y} - b\bar{x}$ in $y = a + bx$, we get

$$\boxed{y - \bar{y} = b(x - \bar{x})} \quad \text{——— (8)}$$

This eqn is called the regression line of y on x and b is called the regression coefficient of y on x and is denoted by b_{yx}. Hence eqn (8) can be rewritten as

$$\boxed{y - \bar{y} = b_{yx}(x - \bar{x})}$$

In eqn (5), shifting the origin to $(\bar{x}, \bar{y})$, we get

$$\Sigma(x-\bar{x})(y-\bar{y}) = a\Sigma(x-\bar{x}) + b\Sigma(x-\bar{x})^2$$

$$\Rightarrow n r \sigma_x \sigma_y = a(0) + b n \sigma_x^2$$

$\because \Sigma(x-\bar{x}) = 0$

$\frac{1}{n}\Sigma(x-\bar{x})^2 = \sigma_x^2$

and $\frac{\Sigma(x-\bar{x})(y-\bar{y})}{n\sigma_x\sigma_y} = r$

$$\Rightarrow \boxed{b_{yx} = r\frac{\sigma_y}{\sigma_x}} \checkmark$$

from eqn ⑤

$$\Rightarrow \boxed{(y-\bar{y}) = r.\frac{\sigma_y}{\sigma_x}(x-\bar{x})}$$

where r is the coefficient of correlation, σ_x and σ_y are the standard deviation of x and y series respectively. a and b are regression coefficient.

$\checkmark\ \sigma_x = \sqrt{\frac{\Sigma dx^2}{N}}$

$\bar{x} = \frac{\Sigma x}{n}$

Line of Regression of x on y :-

Proceeding in the same way, we can derive the regression line of x on y as

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

where b_{xy} is the regression coefficient of x on y and is given by

$$b_{xy} = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma y^2 - (\Sigma y)^2}$$

or
$$b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

$$\Rightarrow \boxed{(x - \bar{x}) = r \cdot \frac{\sigma_x}{\sigma_y}(y - \bar{y})}$$ ✓

and
$$b_{yx} \cdot b_{xy} = r^2 \times \frac{\sigma_y}{\sigma_x} \times \frac{\sigma_x}{\sigma_y}$$

$$\Rightarrow \boxed{r^2 = b_{yx} \cdot b_{xy}}$$ ✓

Non-Linear Regression :-

Let $y = a + bx + cx^2$ ——— (1)

be a second degree parabolic curve of regression of y on x to be fitted for the data (x_i, y_i) for $i = 1, 2, \ldots n$

Error at $x = x_i$ is

$$e_i = y_i - f(x_i)$$

$$e_i = y_i - a - bx_i - cx_i^2$$

Now let,

$$E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i - cx_i^2)^2$$

By principle of Least squares, E should be minm for the best values of a, b and c.

for this $\frac{\partial E}{\partial a} = 0$, $\frac{\partial E}{\partial b} = 0$ and $\frac{\partial E}{\partial c} = 0$

$$\Rightarrow \frac{\partial E}{\partial a} = 0 \Rightarrow 2\sum_{i=1}^{n}(y_i - a - bx_i - cx_i^2)(-1) = 0$$

$$\Rightarrow \boxed{\Sigma y = na + b\Sigma x + c\Sigma x^2} \text{ ——— } (1)$$

$$\frac{\partial E}{\partial b} = 0 \Rightarrow 2\sum_{i=1}^{n}(y_i - a - bx_i - cx_i^2)(-x_i) = 0$$

$$\Rightarrow \boxed{\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3} \text{ ——— } (2)$$

$$\frac{\partial E}{\partial c} = 0 \Rightarrow 2\sum_{i=1}^{n}(y_i - a - bx_i - cx_i^2)(-x_i^2) = 0$$

$$\Rightarrow \boxed{\Sigma x^2 y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4} \text{ ——— } (3)$$

eqns (1), (2), (3) are the normal eqns for fitting a second degree parabolic curve of regression of y on x. Here n is the no. of pairs of values of x and y.

Multiple Linear Regression :-

In this case, the dependent variable is a function of two or more linear or non-linear independent variables. Consider a linear function as

$$y = a + bx + cz \quad \text{——— (1)}$$

error at $x = x_i$ is

$$e_i = y_i - f(x_i) = y_i - a - bx_i - cz_i$$

The sum of the squares of error is

$$E = \sum_{i=1}^{n} e_i^2$$

$$E = \sum_{i=1}^{n} (y_i - a - bx_i - cz_i)^2 \quad \text{——— (2)}$$

By principle of least squares, E should be min^m for the best values of a, b and c

i.e. $\frac{\partial E}{\partial a} = 0$, $\frac{\partial E}{\partial b} = 0$ and $\frac{\partial E}{\partial c} = 0$

therefore,

$$\frac{\partial E}{\partial a} = 0 \Rightarrow 2 \sum_{i=1}^{n} (y_i - a - bx_i - cz_i)(-1) = 0$$

$$\Rightarrow \boxed{\Sigma y = na + b\Sigma x + c\Sigma z} \quad \text{--- (3)}$$

$$\frac{\partial E}{\partial b} = 0 \Rightarrow 2 \sum_{i=1}^{n} (y_i - a - bx_i - cz_i)(-x_i) = 0$$

$$\Rightarrow \boxed{\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma xz} \quad \text{--- (4)}$$

and $$\frac{\partial E}{\partial c} = 0 \Rightarrow 2 \sum_{i=1}^{n} (y_i - a - bx_i - cz_i)(-z_i) = 0$$

$$\Rightarrow \boxed{\Sigma yz = a\Sigma z + b\Sigma xz + c\Sigma z^2} \quad \text{--- (5)}$$

Solving eq^{n}s (3), (4) and (5), we get value of a, b and c

Example 2. *Calculate linear regression coefficients from the following:*

x	$\rightarrow$	1	2	3	4	5	6	7	8
y	$\rightarrow$	3	7	10	12	14	17	20	24

Sol. Linear regression coefficients are given by

$$b_{yx} = \frac{n\,\Sigma xy - \Sigma x\,\Sigma y}{n\,\Sigma x^2 - (\Sigma x)^2}$$

and

$$b_{xy} = \frac{n\,\Sigma xy - \Sigma x\,\Sigma y}{n\,\Sigma y^2 - (\Sigma y)^2}$$

x	y	x^2	y^2	xy
1	3	1	9	3
2	7	4	49	14
3	10	9	100	30
4	12	16	144	48
5	14	25	196	70
6	17	36	289	102
7	20	49	400	140
8	24	64	576	192
$\Sigma x = 36$	$\Sigma y = 107$	$\Sigma x^2 = 204$	$\Sigma y^2 = 1763$	$\Sigma xy = 599$

Here $n = 8$

$$\therefore \quad b_{yx} = \frac{(8 \times 599) - (36 \times 107)}{(8 \times 204) - (36)^2} = \frac{4792 - 3852}{1632 - 1296} = \frac{940}{336} = 2.7976$$

and

$$b_{xy} = \frac{(8 \times 599) - (36 \times 107)}{(8 \times 1763) - (107)^2} = \frac{940}{2655} = 0.3540$$

Example 4. *Find the regression line of y on x for the following data:*

x	1	3	4	6	8	9	11	14
y	1	2	4	4	5	7	8	9

Estimate the value of y, when $x = 10$.

Sol.

S.No.	x	y	xy	x^2
1	1	1	1	1
2	3	2	6	9
3	4	4	16	16
4	6	4	24	36
5	8	5	40	64
6	9	7	63	81
7	11	8	88	121
8	14	9	126	196
Total	56	40	364	524

Let $y = a + bx$ be the line of regression of y on x. Therefore normal equations are

$$\sum y_i = na + b\sum x_i \quad \Rightarrow \quad 40 = 8a + 56b$$

$$\sum x_i y_i = a\sum x_i + b\sum x_i^2 \quad \Rightarrow \quad 364 = 56a + 524b$$

On solving (1) and (2) we get

$$a = \frac{6}{11} \text{ and } b = \frac{7}{11}$$

The equation of the required line is

$$y = \frac{6}{11} + \frac{7}{11}x \quad \text{or} \quad 7x - 11y + 6 = 0$$

If $\qquad x = 0,\ y = \frac{6}{11} + \frac{7}{11}(10) = \frac{76}{11} = 6\frac{10}{11}$. **Ans.**

Example 2. *Find the line of regression of x on y for the following data:*

x	6	2	10	4	8
y	9	11	5	8	7

Sol. Here $n = 5$. Now, form the table given below :

x_i	y_i	y_i^2	$x_i y_i$
6	9	81	54
2	11	121	22
10	5	25	50
4	8	64	32
8	7	49	56
$\sum x_i = 30$	$\sum y_i = 40$	$\sum y_i^2 = 340$	$\sum x_i y_i = 214$

Let the required line be, $x = a + by$

Then $x_i = a + by_i$ and $x_i y_i = a y_i + b y_i^2$ for each i.

Therefore the normal equations are:

$$\sum x_i = na + b\sum y_i$$

$$\sum x_i y_i = a\sum y_i + b\sum y_i^2$$

Putting the values from the table in (2) and (3), we get

$$30 = 5a + 40b \quad \Rightarrow a + 8b = 6$$

$$214 = 40a + 340b \Rightarrow 20a + 170b = 107$$

On solving these equations we get $a = 16.4$ and $b = -1.3$.

Therefore the requried equation is, $x = 16.4 - 1.3y$. **Ans.**

Example 6. *Find the regression coefficient* b_{yx} *between* x *and* y *for the following data:* $\sum x = 24$, $\sum y = 44$, $\sum xy = 306$, $\sum x^2 = 164$, $\sum y^2 = 574$ *and* $n = 4$.

Sol. The given data may be written as $\sum x_i = 24$, $\sum y_i = 44$, $\sum x_i y_i = 306$, $\sum x_i^2 = 164$, $\sum y_i^2 = 574$ and $n = 4$.

$$b_{yx} = \frac{\sum x_i y_i - \dfrac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}}{\left\{\sum x_i^2 - \dfrac{\left(\sum x_i\right)^2}{n}\right\}} = \frac{306 - \dfrac{24 \times 44}{4}}{164 - \dfrac{(24)^2}{4}}$$

$$= \frac{(306 - 264)}{164 - 144} = \frac{42}{20} = 2.1. \quad \textbf{Ans.}$$

Example 8. *For the following observations (x, y), find the regression coefficient b_{yx} and b_{xy} hence find the correlation coefficient between x and y: (1, 2), (2, 4), (3, 8), (4, 7), (5, 10), (6, 5), (7, 14), (8, 16), (9, 2), (10, 20).*

Sol. Here n = 10. We may prepare the table, given below:

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	2	1	4	2
2	4	4	16	8
3	8	9	64	24
4	7	16	49	28
5	10	25	100	50
6	5	36	25	30
7	14	4	196	98
8	16	64	256	128
9	2	81	4	18
10	20	100	400	200
$\sum x_i = 55$	$\sum y_i = 88$	$\sum x_i^2 = 385$	$\sum y_i^2 = 1114$	$\sum x_i y_i = 586$

$$b_{yx} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\left\{\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right\}} = \frac{586 - \frac{55\times 88}{10}}{385 - \frac{(55)^2}{10}} = \frac{102}{82.5} = 1.24$$

And $$b_{xy} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\left\{\sum y_i^2 - \frac{(\sum y_i)^2}{n}\right\}} = \frac{586 - \frac{(55\times 88)}{10}}{1114 - \frac{(88)^2}{10}} = \frac{102}{339.6} = 0.30$$

Now, $b_{yx}\cdot b_{xy} = \left(r\cdot\frac{\sigma_y}{\sigma_x}\right)\left(r\cdot\frac{\sigma_x}{\sigma_y}\right) = r^2$, where r is the coefficient of correlation.

$\therefore \quad r = \sqrt{b_{yx}\cdot b_{xy}} = \sqrt{1.24\times 0.30} = 0.609.$

Thus, $b_{yx} = 1.24$, $b_{xy} = 0.30$ and $r = 0.609$. **Ans.**

Example 11. *Find the correlation coefficient between x and y, when the lines of regression ar[e] ... $9y + 6 = 0$ and $x - 2y + 1 = 0$.*

Sol. Let the line of regression of x on y be $2x - 9y + 6 = 0$

Then, the line of regression of y on x is $x-2y+1=0$.

Therefore $2x-9y+6=0$ and $x-2y+1=0$

$$\Rightarrow \quad x=\frac{9}{2}y-3 \quad \text{and} \quad y=\frac{1}{2}x+\frac{1}{2}$$

$$\Rightarrow \quad b_{xy}=\frac{9}{2} \quad \text{and} \quad b_{yx}=\frac{1}{2}$$

$$\Rightarrow \quad r=\sqrt{b_{xy}\cdot b_{yx}}=\sqrt{\left(\frac{9}{2}\times\frac{1}{2}\right)}=\frac{3}{2}>1, \text{ which is impossible.}$$

$$\Rightarrow \qquad b_{xy} = \frac{9}{2} \qquad \text{and} \qquad b_{yx} = \frac{1}{2}$$

$$\Rightarrow \qquad r = \sqrt{b_{xy} \cdot b_{yx}} = \sqrt{\left(\frac{9}{2} \times \frac{1}{2}\right)} = \frac{3}{2} > 1, \text{ which is impossible.}$$

So, our choice of regression line is incorrect.

Therefore, the regression line of x on y is $x - 2y + 1 = 0$.

And, the regression line of y on x is $2x - 9y + 6 = 0$.

$$\Rightarrow \qquad x = 2y - 1 \qquad \text{and} \qquad y = \frac{2}{9}x + \frac{2}{3}$$

$$\Rightarrow \qquad b_{xy} = 2 \qquad \text{and} \qquad b_{yx} = \frac{2}{9}$$

$$\Rightarrow \qquad r = \sqrt{b_{xy} . b_{yx}} = \sqrt{\left(2 \times \frac{2}{9}\right)} = \frac{2}{3} \quad \textbf{Ans.}$$

Example 3. *The following table gives age (x) in years of cars and annual maintenance cost (y) in hundred rupees:*

x:	1	3	5	7	9
y:	15	18	21	23	22

Estimate the maintenance cost for a 4 year old car after finding the regression equation.

Sol.

x	y	xy	x^2
1	15	15	1
3	18	54	9
5	21	105	25
7	23	161	49
9	22	198	81
$\Sigma x = 25$	$\Sigma y = 99$	$\Sigma xy = 533$	$\Sigma x^2 = 165$

Here, $n = 5$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{25}{5} = 5$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{99}{5} = 19.8$$

$\therefore$
$$b_{yx} = \frac{n\,\Sigma xy - \Sigma x\,\Sigma y}{n\,\Sigma x^2 - (\Sigma x)^2} = \frac{(5 \times 533) - (25 \times 99)}{(5 \times 165) - (25)^2} = 0.95$$

Regression line of y on x is given by

$$y - \bar{y} = b_{yx}\,(x - \bar{x})$$

$$\Rightarrow \quad y - 19.8 = 0.95\,(x - 5)$$

$$\Rightarrow \quad y = 0.95x + 15.05$$

When $x = 4$ years, $y = (0.95 \times 4) + 15.05$

$= 18.85$ hundred rupees $=$ Rs. 1885.

Example 8. *For 10 observations on price (x) and supply (y), the following data were obtained (in appropriate units):*

$$\Sigma x = 130, \quad \Sigma y = 220, \quad \Sigma x^2 = 2288, \quad \Sigma y^2 = 5506 \text{ and } \Sigma xy = 3467$$

Obtain the two lines of regression and estimate the supply when the price is 16 units.

Sol. Here, $n = 10,\ \bar{x} = \dfrac{\Sigma x}{n} = 13$ and $\bar{y} = \dfrac{\Sigma y}{n} = 22$

Regression coefficient of y on x is

$$b_{yx} = \frac{n\,\Sigma xy - \Sigma x\,\Sigma y}{n\,\Sigma x^2 - (\Sigma x)^2} = \frac{(10 \times 3467) - (130 \times 220)}{(10 \times 2288) - (130)^2}$$

$$= \frac{34670 - 28600}{22880 - 16900} = \frac{6070}{5980} = 1.015$$

$\therefore$ Regression line of y on x is

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 22 = 1.015(x - 13)$$

$$\Rightarrow \qquad y = 1.015x + 8.805$$

Regression coefficient of x on y is

$$b_{xy} = \frac{n \, \Sigma xy - \Sigma x \, \Sigma y}{n \, \Sigma y^2 - (\Sigma y)^2}$$

$$= \frac{(10 \times 3467) - (130 \times 220)}{(10 \times 5506) - (220)^2} = \frac{6070}{6660} = 0.9114$$

Regression line of x on y is

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$x - 13 = 0.9114(y - 22)$$

$$x = 0.9114y - 7.0508$$

Since we are to estimate supply (y) when price (x) is given therefore we are to use regression line of y on x here.

When $x = 16$ units,

$$y = 1.015(16) + 8.805 = 25.045 \text{ units.}$$