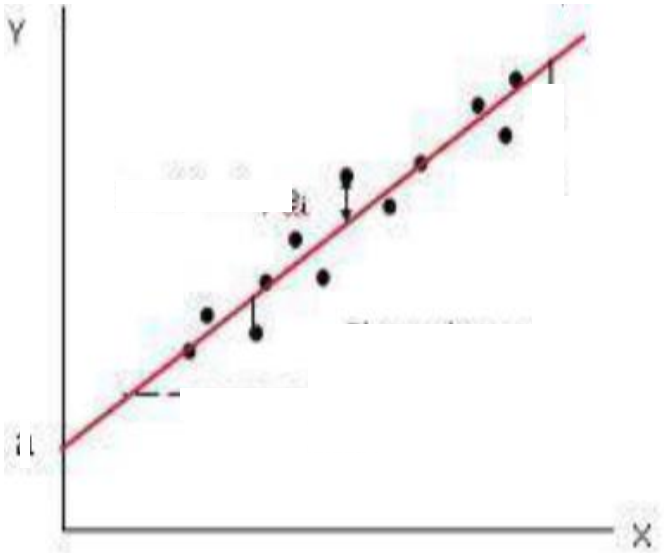
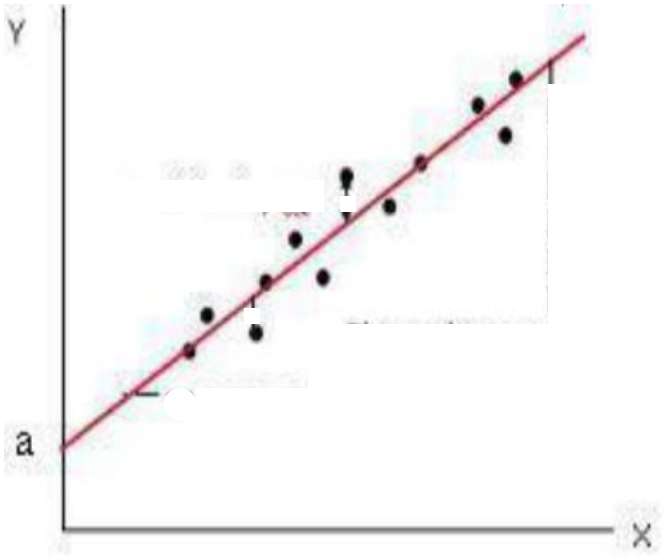
Regression of lines

In the Scatter plot if points cluster around some curve then the curve is called as regression curve. If the regression curve is straight line then it is called as regression line . It is also called as best fitting straight line .

The best fitting straight line is a line such that the sum of square of distances(deviation) of the points from the line is minimum .





The deviation can be vertical or horizontal as shown in diagram , so according to that we get two regression line .
When **vertical** deviation is used for a line then the line is called as regression line of **y on x** and it's equation is written as

$$(y - \bar{y}) = b_{yx}\,(x - \bar{x}) \qquad b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

Here y is dependent variable and x is independent variable
For given bivariate data , if we want to predict value of the variable y for the given value of the variable x then we use this equation

When horizontal deviation is used for a line then the line is called as regression line of x on y and it's equation is written as

$$(x - \bar{x}) = b_{xy}\,(y - \bar{y}) \qquad b_{xy} = r\,\frac{\sigma_x}{\sigma_y}$$

Here x is dependent variable and y is independent variable

For given bivariate data , if we want to predict value of the variable x for the given value of the variable y then we use this equation

Formula of Regression coefficients

(i) When all five sums are given $\sum x, \sum y, \sum x^2, \sum y^2$ and $\sum xy$.

$$b_{yx} = \frac{\sum xy - \frac{\sum x \cdot \sum y}{N}}{\sum x^2 - \frac{(\sum x)^2}{N}} \qquad b_{xy} = \frac{\sum xy - \frac{\sum x \cdot \sum y}{N}}{\sum y^2 - \frac{(\sum y)^2}{N}}$$

(ii) If $\bar{x}$, $\bar{y}$ are integers

$$b_{yx} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = \frac{\sum(x')(y')}{\sum(x')^2} \qquad b_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(y-\bar{y})^2} = \frac{\sum(x')(y')}{\sum(y')^2}$$

(iii) If $\bar{x}$, $\bar{y}$ are not integers, put $d_x = x - A$, $d_y = y - B$

$$b_{yx} = \frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{N}}{\sum d_x^2 - \frac{(\sum d_x)^2}{N}} \text{ and } b_{xy} = \frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{N}}{\sum d_y^2 - \frac{(\sum d_y)^2}{N}}$$

Properties of coefficient of Regression

(I) coefficient of correlation r is the geometric mean between the Regression coefficients

$$b_{yx}.b_{xy} = r\frac{\sigma_y}{\sigma_x}.r\frac{\sigma_x}{\sigma_y} = r^2$$

Note: Since right hand side is positive, values of Regression coefficients both will be either positive or both will be negative

(II) If one coefficient of Regression is greater than 1, the other must be less than 1.

(III) Arithmetic mean of the Regression coefficients is greater than or equal to the coefficient of correlation

(IV) Coefficients of Regressions are independent of change of origin but not of change of scale.

(V)Slopes of regression lines

Consider the regression equation of Y on X;

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

And regression equation of X on Y ;

$$(x - \bar{x}) = b_{xy}(y - \bar{y}) \Rightarrow y = \bar{y} + \frac{1}{b_{xy}}(x - \bar{x})$$

$\therefore$ Slope of regression equation of Y on X is b_{yx}

And slope of regression equation of X on Y is $\frac{1}{b_{xy}}$

(VI) The angle between two regression lines.

The angle between two lines is given by

$\tan\theta = \left|\frac{m_1 - m_2}{1 + m_1 m_2}\right|$ where m_1 and m_2 are slopes of the lines that are intersecting

$$\therefore \tan\theta = \left|\frac{b_{yx} - \frac{1}{b_{xy}}}{1 + b_{yx}\frac{1}{b_{xy}}}\right| = \left|\frac{b_{yx}b_{xy} - 1}{b_{xy} + b_{yx}}\right|$$

$$\therefore \theta = \tan^{-1}\left|\frac{b_{yx}b_{xy} - 1}{b_{xy} + b_{yx}}\right|$$

Ex 1 Find regression coefficients hence equation of regression line of y on x & the coefficient of correlation if

$\Sigma x = 15000,\ \Sigma x^2 = 2272500,\ \Sigma y = 6800,\ \Sigma y^2 = 463025,\ \Sigma xy = 1022250$

Answer

$$b_{yx} = \frac{\Sigma xy - \frac{\Sigma x \cdot \Sigma y}{N}}{\Sigma x^2 - \frac{(\Sigma x)^2}{N}} = \frac{1022250 - \frac{15000 \times 6800}{100}}{2272500 - \frac{15000^2}{100}} = \frac{2250}{22500} = 0{\cdot}1.$$

$$b_{xy} = \frac{\Sigma xy - \frac{\Sigma x \cdot \Sigma y}{N}}{\Sigma y^2 - \frac{(\Sigma y)^2}{N}} = \frac{1022250 - \frac{15000 \times 6800}{100}}{463025 - \frac{6800^2}{100}} = \frac{2250}{625} = 3{\cdot}6.$$

$$r = \sqrt{b_{yx} \times b_{xy}} = \sqrt{0{\cdot}1 \times 3{\cdot}6} = 0{\cdot}6.$$

The equation of the lines of regression of y on x is

$$y - \bar{y} = b_{yx}(x - \bar{x}) \quad \therefore \quad y - 68 = 0{\cdot}1(x - 1500)$$

$$\therefore \quad y = 0{\cdot}1x - 82.$$

Exercise

1. *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.

Ans $7x - 11y + 6 = 0$; $6\frac{10}{11}$

2. Find the equations to the lines of regression and the coefficient of correlation for the following data:

x	2	4	5	6	8	11
y	18	12	10	8	7	5

Ans. $y - 10 = -1.34\,(x - 6)$, $x - 6 = -0.632\,(y - 10)$, $r = -0.92$

3. The following results were obtained from lineups in Applied Mechanics and Engineering Mathematics in an examination :

	Applied Mechanics *(x)*	*Engg. Maths.* *(y)*
Mean	47.5	10.5
Standard deviation	16.8	10.8

$r = 0.95$

Find both the regression equations. Also estimate the value of y for $x = 30$.

Ans. $y = 0.611x + 10.5$, $x = 1.478\,y - 1.143$, $y = 28.83$

4. The following results were obtained from records of age (x) and systolic blood pressure (y) of a group of 10 men :

	x	y
Mean	53	142
Variance	130	165

and $\sum (x - \bar{x})(y - \bar{y}) = 1220$

Find the appropriate regression equation and use it to estimate the blood pressure of a man whose age is 45.

Ans. $y = 0.94\,x + 92.26$, Blood pressure = 134.56

Ex 2

Given $\sigma_x{}^2 = 16, 6y = 5x + 90, 15x = 8y + 130 = 16$

Find $\bar{x}, \bar{y}, \sigma_y{}^2$ and r

Answer

To find $\bar{x}, \bar{y}$

We solve equations of regression lines simultaneously

$$\bar{x} = 30, \bar{y} = 40$$

To find r : Suppose the first equation represents, the line of regression of X on Y.

Writing it as $X = \frac{6}{5}Y - 18$, we find $b_{xy} = \frac{6}{5}$.

Suppose the second equation represents the line of regression of Y on X.

Writing it as $Y = \frac{15}{8}X - \frac{130}{8}$, we find $b_{yx} = \frac{15}{8}$.

$$\therefore \quad r = \sqrt{b_{xy} \times b_{yx}} = \sqrt{\frac{6}{5} \times \frac{15}{8}} = \sqrt{\frac{9}{4}} = \sqrt{2{\cdot}25} = 1{\cdot}5.$$

But the value of r can never be greater than 1 numerically. Hence, our supposition is wrong.

Now treating the first equation as representing the line of regression of Y on X,

$$Y = \frac{5}{6}X + 15 \quad \therefore \quad b_{yx} = \frac{5}{6}.$$

Treating the second equation as representing the line of regression of X on Y,

$$X = \frac{8}{15}Y + \frac{130}{15} \quad \therefore \quad b_{xy} = \frac{8}{15}$$

$$\therefore \quad r = \sqrt{b_{xy} \times b_{yx}} = \sqrt{\frac{8}{15} \times \frac{5}{6}} = \sqrt{\frac{4}{9}} = \frac{2}{3} = 0{\cdot}667.$$

To find $\sigma_y{}^2$

We know that

$b_{yx} = r\frac{\sigma_y}{\sigma_x}$ and $b_{xy} = r\frac{\sigma_x}{\sigma_y}$

$$\frac{b_{yx}}{b_{xy}} = \frac{\sigma_y{}^2}{\sigma_x{}^2}$$

Therefore $\sigma_y{}^2 = \sigma_x{}^2 \frac{b_{yx}}{b_{xy}}$

$= 25$

Exercise

Q.The equations of the two regression lines are $y = 0.516x + 33.73$ and $x = 0.512y + 32.53$ Find r, $\bar{x}$ and $\bar{y}$

Ans r=.514, mean(x)=67.6, mean(y)=68.61

Q.The equations of the two regression lines are

$y = -0.6x + 4.6$ and $x = -0.4y + 6.4$

Find r, $\bar{x}$ and $\bar{y}$

Ans r=-.49, mean(x)=6, mean(y)=1