$$A = \frac{0.6655}{4}$$

taking Anti log: $b = 1.38$ $\quad a = 0.68$.

$\therefore$ Required Curve: $Y = (0.68)(1.38)^X$

Exercises

(1) Fit an equation of the form $Y = ab^X$ to the following date

(a)

X :	2	3	4	5	6
Y :	144	172.8	207.4	248.8	298.6

$Y = (101.3)(1.196)^X$

(b)

X :	2	3	4	5	6
Y :	8.3	15.4	33.1	65.2	127.4

Also estimate: Y when $X = 4.5$, 7 and 3.5

(2) Fit an equation of the form $Y = a e^{bX}$

X :	1	2	3	4	5	6
Y :	1.6	4.5	13.8	40.2	125.0	300

$Y = (0.557) e^{1.05X}$

Correlation And Regression:

(1) Def: Correlation Co-efficient :

Correlation Co-efficient between two r.v 's X and Y usually denoted by $\rho(X, Y)$ <or $r(X, Y)$> is a numerical measure of linear relationship between them and is defined by

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

If (x_i, y_i) $i = 1, 2, \cdots n$ is n pair of values of X and Y then,

$$Cov(X, Y) = E[\{X - E(X)\}\{Y - E(Y)\}]$$
$$= \frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})$$

$$\sigma_X^2 = E[\{X - E(X)\}^2] = \frac{1}{n} \sum (x_i - \bar{x})^2$$

$$\sigma_Y^2 = \frac{1}{n} \sum (y_i - \bar{y})^2 \quad \text{where} \quad \bar{x} = \frac{1}{n} \sum x_i$$

$$\therefore \rho(X, Y) = \frac{\frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})}{\left[\frac{1}{n} \sum (x_i - \bar{x})^2 \cdot \frac{1}{n} \sum (y_i - \bar{y})\right]^{1/2}}$$

$$\therefore \rho^2 = \frac{(\sum a_i b_i)^2}{\sum a_i^2 \sum b_i^2} \quad \text{where} \quad a_i = x_i - \bar{x}, \quad b_i = y_i - \bar{y} \quad \cdots (I)$$

Now we know a well known inequality: If a_i, b_i $i = 1, 2, \ldots n$ are two sets of real quantities then

$$\left(\sum_{i=1}^{n} a_i b_i\right)^2 \leq \left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right) \quad \langle \text{Schwartz inequality} \rangle$$

'=' holds iff $\frac{a_1}{b_1} = \frac{a_2}{b_2} = \cdots = \frac{a_n}{b_n}$.

Using Schwartz Inequality we have from (I)

$$\rho^2 \leq 1$$

$$\Rightarrow |\rho| \leq 1$$

$$\Rightarrow -1 \leq \rho \leq 1$$

Hence Correlation Coefficient Cannot exceed unity numerically. It always lies between +1 and −1

If $r = 1$ then ρ is perfect and positive

$= -1$ " ρ is perfect and negative.

obviously if X and Y are independent random variables then $\rho(X, Y) = 0$ because in such Case $Cov(X, Y) = 0$.

PROBLEM

(1) Calculate the Correlation Coefficient from the following table:

X	Y	X^2	Y^2	XY
65	67	4225	4489	4355
66	68	4356	4624	4488
67	65	4489	4225	4355
67	68	4489	4624	4556
68	72	4624	5184	4896
69	72	4761	5184	4968
70	69	4900	4761	4830
72	71	5184	5041	5112
$\sum$ 544	552	37028	38132	37560

$$\bar{X} = \frac{1}{n}\sum x = 68 \qquad \bar{Y} = 69$$

We have: $Cov(X,Y) = \frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y})$

$$= \frac{1}{n}\sum (x_i y_i - \bar{x}\, y_i - \bar{y}\, x_i + \bar{x}\,\bar{y}$$

$$= \frac{1}{n}\sum x_i y_i - \bar{x}\,\frac{1}{n}\sum y_i - \bar{y}\,\frac{1}{n}\sum x_i + \bar{x}\,\bar{y}$$

$$= \frac{1}{n}\sum x_i y_i - \bar{x}\,\bar{y} - \bar{x}\,\bar{y} + \bar{x}\,\bar{y}$$

$$= \frac{1}{n}\sum x_i y_i - \bar{x}\,\bar{y}$$

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

$$= \frac{1}{n}\sum (x_i^2 - 2x_i\bar{x} + \bar{x}^2)$$

$$= \frac{1}{n}\sum x_i^2 - 2\bar{x}\,\frac{1}{n}\sum x_i + \bar{x}^2$$

$$= \frac{1}{n}\sum x_i^2 - 2\bar{x}^2 + \bar{x}^2$$

$$= \frac{1}{n}\sum x_i^2 - \bar{x}^2$$

$$\sigma_Y^2 = \frac{1}{n}\sum y_i^2 - \bar{y}^2$$

$$\rho(X,Y) = \frac{\frac{1}{n}\sum XY - \bar{X}\,\bar{Y}}{\sqrt{\left(\frac{1}{n}\sum X^2 - \bar{X}^2\right)\left(\frac{1}{n}\sum Y^2 - \bar{Y}^2\right)}}$$

$$= \frac{\frac{1}{8}\times 37560 - 68\times 69}{\sqrt{\left(\frac{37028}{8} - 68^2\right)\left(\frac{38132}{8} - 69^2\right)}}$$

$$= 0.603$$

(2) A Computer while calculating Correlation Coefficient between two Variables X and Y from pairs of observation obtained the following results

$n = 25$, $\sum X = 125$, $\sum X^2 = 650$, $\sum Y = 100$, $\sum Y^2 = 640$

$\sum XY = 508$.

It was, however, later discovered at the time of checking that he was Copied down two pairs as

X	Y
6	14
8	6

while the Correct values were

X	Y
8	12
6	8

obtain the Correct value of the Correlation Coefficient

Ans

Corrected $\sum X = 125 - 6 - 8 + 8 + 6 = 125$

" $\sum Y = 100 - 14 - 6 + 12 + 8 = 100$

" $\sum X^2 = 650 - 6^2 - 8^2 + 8^2 + 6^2 = 650$

" $\sum Y^2 = 640 - 14^2 - 6^2 + 12^2 + 8^2 = 616$

" $\sum XY = 508 - 6\times14 - 8\times6 + 8\times12 + 6\times8 = 520$

$\bar{X} = \frac{1}{n}\sum X = \frac{1}{25}\times125 = 5$, $\bar{Y} = \frac{1}{25}\times100 = 4$

$Cov(X,Y) = \frac{1}{n}\sum XY - \bar{X}\bar{Y} = \frac{1}{25}\times520 - 5\times4 = \frac{4}{5} = 0.8$

$\sigma_X^2 = \frac{1}{n}\sum X^2 - \bar{X}^2 = \frac{1}{25}\times650 - 5^2 = 1$

$\sigma_Y^2 = \frac{1}{n}\sum Y^2 - \bar{Y}^2 = \frac{1}{25}\times616 - 4^2 = \frac{216}{25} = 8.64$

Corrected $\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X\,\sigma_Y} = \frac{4/5}{1\times9/5} = \frac{2}{3} = .67$

$= \frac{0.8}{\sqrt{8.64}} = \frac{0.8}{2.94} = 0.272$

(3) (a) If $Z = aX + bY$ and $\rho \equiv \rho(X,Y)$ Show that $\sigma_Z^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\rho\,\sigma_X\sigma_Y$

(b) Show that the Correlation Coefficient between two r.v's X and Y can be written as:

$$\rho = \frac{\sigma_X^2 + \sigma_Y^2 - \sigma_{X-Y}^2}{2\sigma_X\sigma_Y}.$$

Solution:

(a) $Z = aX + bY$

$E(Z) = aE(X) + bE(Y)$.

$Z - E(Z) = a[X - E(X)] + b[Y - E(Y)]$

Squaring and taking expectation on both sides

$$E[Z - E(Z)]^2 = a^2 E[X - E(X)]^2 + b^2 E[Y - E(Y)]^2 + 2ab\,E\{[X - E(X)][Y - E(Y)]\}$$

$$\sigma_Z^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\,Cov(X,Y)$$

$$\Rightarrow \sigma_Z^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\rho\,\sigma_X\sigma_Y \quad \left[\because \rho = \frac{Cov(X,Y)}{\sigma_X\sigma_Y}\right]$$

(b) In particular if we take $a=1$, $b=-1$ then $Z = X - Y$ and from result of (a)

$$\sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2 - 2\rho\,\sigma_X\sigma_Y$$

$$\Rightarrow \rho = \frac{\sigma_X^2 + \sigma_Y^2 - \sigma_{X-Y}^2}{2\sigma_X\sigma_Y}$$

REGRESSION

1. Def: Regression Analysis is a mathematical measure of the avarage relationship between two or more variables interms of the original units of the given data.

In regression analysis there are two types of variables. The variable whose value is to be predicted is called dependent variable and the variable which is used for prediction is called independent variable.

2. Lines of Regression:

The line of regression is the line which gives the best estimate to the value of one variable for any specific value of other variable. Therefore the line of regression is the line of best fit and is obtained by Principles of least Square

Consider n pair of points (x_i, y_i) $i = 1, 2, \cdots n$
Y is dependent and X is independent variable.
Let the line of regression of Y on X be

$$Y = a + bX$$

By Principle of least square, Normal equation to estimate a, b are

$$\sum y_i = na + b\sum x_i \quad \cdots (1)$$
$$\sum x_i y_i = a\sum x_i + b\sum x_i^2 \quad \cdots (2)$$

From (1) $\bar{y} = a + b\bar{x}$ $\cdots (3)$

$\Rightarrow$ line of reg. of (Y on X) passes through means $(\bar{x}, \bar{y})$

Again, $Cov(X,Y) = \frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}$

$$\Rightarrow \mu_{11} = \frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}$$
$$\Rightarrow \boxed{\frac{1}{n}\sum x_i y_i = \mu_{11} + \bar{x}\bar{y}} \quad \cdots (4)$$

Also, $\sigma_x^2 = \frac{1}{n}\sum x_i^2 - \bar{x}^2$

$$\Rightarrow \boxed{\frac{1}{n}\sum x_i^2 = \sigma_x^2 + \bar{x}^2} \quad \cdots (5)$$

From (2) using (4) and (5)

$$\frac{1}{n}\sum x_i y_i = a\cdot\frac{1}{n}\sum x_i + b\,\frac{1}{n}\sum x_i^2$$
$$\mu_{11} + \bar{x}\bar{y} = a\bar{x} + b(\sigma_x^2 + \bar{x}^2) \quad \cdots (6)$$

(3) $\times \bar{x} \Rightarrow$ $\bar{x}\bar{y} = a\bar{x} + b\bar{x}^2$ $\cdots (7)$

(7) $-$ (6) $\Rightarrow$ $\mu_{11} = b\sigma_x^2$

$$\Rightarrow \boxed{b = \frac{\mu_{11}}{\sigma_x^2}} \quad \cdots (8)$$

Since b is the slope of the line of regression of Y on X and since the line passing through $(\bar{x}, \bar{y})$ its equation is

$$Y - \bar{y} = \frac{\mu_{11}}{\sigma_x^2}(x - \bar{x})$$

$$\Rightarrow \boxed{Y - \bar{y} = \rho \frac{\sigma_Y}{\sigma_x}(x - \bar{x}) \quad \left[\because \rho = \frac{\mu_{11}}{\sigma_x \sigma_Y}\right]} \quad \cdots (9)$$

Starting with the equation $X = a + bY$ and proceeding similarly the equation of the line of regression of X on Y becomes:

$$X - \bar{x} = \frac{\mu_{11}}{\sigma_Y^2}(Y - \bar{y})$$

$$\Rightarrow \boxed{X - \bar{x} = \rho \frac{\sigma_x}{\sigma_Y}(Y - \bar{y})} \quad \cdots (10)$$

Hence there are always two lines of regression one of Y on X and other X on Y. In fact Y on X is used to estimate the value of Y for any given value of X. While X on Y is used to estimate the value of X for any given value of Y. The estimate so obtained will be best in the sense that it will have the minimum possible error as defined by the Principle of least square.

Important Properties:

1. b the slope of the line of regression of Y on X is also called Coefficient of regression of Y on X and is denoted by b_{YX}

$$\therefore b_{YX} = \frac{\mu_{11}}{\sigma_x^2} = \rho \frac{\sigma_Y}{\sigma_X} \quad \cdots (11)$$

b_{YX} represents the increment in the value of the dependent variable Y corresponding to a unit change in the value of the independent variable X.

Similarly

$$b_{XY} = \text{Coefficient of regression of X on Y}$$

$$= \frac{\mu_{11}}{\sigma_Y^2} = \rho \frac{\sigma_x}{\sigma_Y} \quad \cdots (12)$$

$$(11) \times (12) \Rightarrow b_{YX}\, b_{XY} = \rho^2$$

$$\Rightarrow \rho = \pm \sqrt{b_{YX}\, b_{XY}} \quad \cdots (13)$$

$\Rightarrow$ Correlation Coefficient is the G.M of regression Coefficient.

Note that if the regression Coefficients are positive ρ is +ve and if they are negative ρ is -ve (13) So that the sign to be taken before the square root is that of the regression Coefficients.

If one regression coefficient is >1 then other must <1

$b_{YX} > 1$ then we have to show that

2. ~~If possible that~~ $\wedge$ $b_{XY} < 1$

Now: $b_{YX} > 1 \Rightarrow \frac{1}{b_{YX}} < 1$

Since $\rho^2 \leq 1 \Rightarrow b_{YX}\, b_{XY} \leq 1$

$\Rightarrow b_{XY} \leq \frac{1}{b_{YX}} < 1$ <PROVED>

$\Rightarrow$ If one reg. Coefficient is >1 then other must be <1

3. A.M of regression Co-efficients is greater than the Correlation Co-efficient

We have to prove that, $\frac{1}{2}(b_{YX} + b_{XY}) \geq \rho$

$\Rightarrow \frac{1}{2}\left(\rho \frac{\sigma_Y}{\sigma_X} + \rho \frac{\sigma_X}{\sigma_Y}\right) \geq \rho$

$\Rightarrow \sigma_X^2 + \sigma_Y^2 - 2\sigma_X\sigma_Y \geq 0$.

$\Rightarrow (\sigma_X - \sigma_Y)^2 \geq 0$

Which is always true. Hence the result.

4. Regression line of:

Y on X : $Y - \bar{y} = \rho \frac{\sigma_Y}{\sigma_X}(x - \bar{x})$ ....(1)

X on Y : $x - \bar{x} = \rho \frac{\sigma_X}{\sigma_Y}(Y - \bar{y})$ ....(2)

$\therefore$ Slope of (1) $m_1 = \rho \frac{\sigma_Y}{\sigma_X}$

" " (2) $m_2 = \frac{1}{\rho} \frac{\sigma_Y}{\sigma_X}$.

If θ be the angle between (1) & (2)

$$\tan\theta = \frac{\rho \frac{\sigma_Y}{\sigma_X} - \frac{1}{\rho}\frac{\sigma_Y}{\sigma_X}}{1 + \frac{\sigma_Y^2}{\sigma_X^2}} = \frac{\rho^2 - 1}{\rho}\left(\frac{\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2}\right)$$

$$= \frac{1-\rho^2}{\rho}\left(\frac{\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2}\right) \left[\because \rho^2 \leq 1\right]$$

$$\therefore \theta = \tan^{-1}\left\{\frac{1-\rho^2}{\rho}\left(\frac{\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2}\right)\right\} \quad \ldots(3)$$

(i) if $\rho = 0$, $\tan\theta = \infty \Rightarrow \theta = \frac{\pi}{2}$ $\Rightarrow$ if two variables are uncorrelated then lines of regression becames perpendicular to each other

(ii) if $\rho = \pm 1$ then $\tan\theta = 0 \Rightarrow \theta = 0$ or π In this case the lines of regression either Coincide or they are parallel to each other. But since the lines of regression passing through $(\bar{x}, \bar{y})$ they can not be parallel. Hence in the case of perfect Correlation +ve or -ve the regression lines are Coincide

(iii) Whenever two lines of intersect, there are two angles between them, one angle acute and other obtuse.

$\tan\theta > 0$ if $0 < \theta < \pi/2$ i.e θ is an acute angle.

$\tan\theta < 0$ if $\frac{\pi}{2} < \theta < \pi$ i.e θ is obtuse angle

and since $0 < \rho^2 < 1$ an acute angle θ_1 and obtuse angle θ_2 are given by

$$\theta_1 = \text{Acute angle} = \tan^{-1}\left\{\frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2} \; \frac{1-\rho^2}{\rho}\right\}$$

$$\theta_2 = \text{Obtuse angle} = \tan^{-1}\left\{\frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2} \; \frac{\rho^2-1}{\rho}\right\}$$

PROBLEMS

#① In partially destroyed Laboratory record of an analysis of correlation data, the following results only are legible.

$\text{Var}(x) = 9$.

Regression equations: $8x - 10y + 66 = 0$, $40x - 18y = 214$

What are (i) the mean values of x and y

(ii) $\rho(x, y)$

(iii) σ_y.

Solution

(i) Since both lines passes through $(\bar{x}, \bar{y})$

$\therefore$ $8\bar{x} - 10\bar{y} + 66 = 0$

$40\bar{x} - 18\bar{y} - 214 = 0$

Solving: $\bar{x} = 13$, $\bar{y} = 17$

(ii) Let $8x - 10y + 66 = 0$ be line of y on x

$40x - 18y - 214 = 0$ " " " x on y

$$y = \frac{8}{10}x + \frac{66}{10}, \quad x = \frac{18}{40}y + \frac{214}{40}.$$

$$b_{yx} = \frac{8}{10} = \frac{4}{5}, \quad b_{xy} = \frac{18}{40} = \frac{9}{20}$$

$$\rho^2 = \frac{8}{10} \times \frac{9}{20} = \frac{9}{25} \Rightarrow \rho = \pm 0.6$$

Since both regression co-efficient are positive

$\rho = 0.6$

iii) We have $b_{YX} = \rho \frac{\sigma_Y}{\sigma_X} \Rightarrow \frac{4}{5} = \frac{3}{5} \times \frac{\sigma_Y}{3}$

$\Rightarrow \sigma_Y = 4$.

Note that if we assume that $8X - 10Y + 66 = 0$: X on Y
$40X - 18Y - 214 = 0$: Y on X

$X = \frac{10}{8}Y - \frac{66}{8}$ $\quad b_{XY} = \frac{10}{8}$

$Y = \frac{40}{18}X - \frac{214}{18}$ $\quad b_{YX} = \frac{40}{18}$

$\rho^2 = 2.78$ Since $0 < \rho^2 \leq 1$ this assumption is not valid here.

#② Find the most likely price in Mumbai corresponding to the price of Rs 70 at Calcutta from the following

	Calcutta	Mumbai
Avarage Price ——	65	67
Standard deviation ——	2.5	3.5

Correlation Coefficient between prices of Commodities in the two Cities is 0.8.

Solution Price of Mumbai = Y
" " Calcutta = X

$\bar{X} = 65$ $\quad \bar{Y} = 67$ $\quad \sigma_X = 2.5$ $\quad \sigma_Y = 3.5$

$\rho(X,Y) = 0.8$ We want Y for $X = 70$

Line of regression of Y on X

$Y - \bar{Y} = \rho \frac{\sigma_Y}{\sigma_X}(X - \bar{X})$

$\Rightarrow Y = 67 + 0.8 \frac{3.5}{2.5}(X - 65)$

$\Rightarrow Y\Big|_{X=70} = 67 + 0.8 \cdot \frac{3.5}{2.5}(70 - 65) = 72.6$

#③ Given $X = 4Y + 5$ regression line of X on Y
$Y = KX + 4$ " " " Y on X.

Then $b_{XY} = 4$, $b_{YX} = K$

$\therefore r^2 = b_{XY}\, b_{YX} = 4K$

Since $0 < r^2 < 1 \Rightarrow 0 < 4K \leq 1$

In particular if: $K = \frac{1}{16}$ then $r^2 = 4 \cdot \frac{1}{16} = \frac{1}{4}$

$r = +\frac{1}{2}$ since both b_{XY} & $b_{YX} > 0$ $(\because K > 0)$

(i) Prove that $0 < K \leq 1/4$

(ii) If $K = \frac{1}{16}$, Find the value of ρ (Correlation Coefficient) $\bar{X}$ and $\bar{Y}$.