Regression

After establishing the fact that two variable are closely related, using ~~coef~~ correlation, we may be interested in estimating (predicting) the value of one variable, given the value of other variable.
For ex. if we know that yield of rice and rainfall are closely related, we may need to find out amount of rain required to achieve a certain production.

Regression analysis reveals average relationship between two or more variables and this makes possible estimation or prediction.

correlation is merely a tool of ascertaining the degree of relationship b/w two varible and therefor, we can'nt say that one variable is the cause and other the effect. for examplee, a high degree of correlation b/w price and demand for a certain commodity at a particular point of time may not suggest which is the cause and which is effect. While in regression analysis one variable is taken as dependent variable while other as independent variable.

Regression line :- ~~The~~ If we take two variables X and Y, we have two regression lines as: regression of X on Y and the regression of Y on X.

The regression line of Y on X gives the most probable values of Y for given values of X and the regression line of X on Y gives the most probable values of X on Y, for given values of Y.

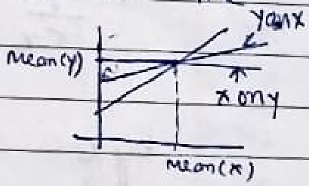
Note

1. when there is perfect positive or perfect negative correlation b/w two variables ($r = \pm 1$), the two regression lines with coincide ie we will have only one line.

2. The farther the two regression line from each other, the lesser is the degree of Correlation

3. If X & Y are independent, ie r is zero the lines of regression are at right angles ie no correlation b/w X and Y.

4. The regression lines cuts each other at the point of average of X & Y. ie



Regression eq^n. of Y on X :-

$$Y = a + b\,X$$

Y → dependent var.
X → Independent var.

using algebra, the equations which gives the values of a & b is given as:-

$$\left.\begin{aligned} \Sigma y_i &= na + b\,\Sigma x_i \\ \Sigma x_i y_i &= a\,\Sigma x_i + b\,\Sigma x_i^2 \end{aligned}\right] \text{Normal eq}^n\text{s} \quad (1)$$

these are called normal equation. ~~but this~~ solving these two, a & b can be computed, but this method of finding regression equation is tedious as Σx_i, Σy_i, $\Sigma x_i y_i$, Σx_i^2 will be difficult to compute for large no. data.

Instead calculation can very much simplified instead of dealing with ~~when~~ actual values of x and y, we take the deviation of x & y from their respective mean.

In this case the two regression eq^{n}s are written as follows:-

(i) Regression line of Y on X

$$Y - \bar{Y} = r \frac{\sigma_y}{\sigma_x} (X - \bar{X})$$

$\bar{X}$ = mean of X
$\bar{Y}$ = mean of Y

$r \frac{\sigma_y}{\sigma_x}$ is called regression coeff. of Y on X. denoted by b_{yx}

ie $\boxed{b_{yx} = r \frac{\sigma_y}{\sigma_x}}$ $\left[\text{since } r = \frac{\mathrm{cov}(x,y)}{\sigma_x \sigma_y} \text{ ie } b_{yx} = \frac{\mathrm{cov}(x,y)}{\sigma_x^2} \right]$

(2) Regression line of X on Y :-

$$X - \bar{X} = r \frac{\sigma_x}{\sigma_y} (Y - \bar{Y})$$ $\quad b_{xy} = \frac{\mathrm{cov}(x,y)}{\sigma_y^2}$

where $r \frac{\sigma_x}{\sigma_y} = b_{xy}$ called regression coeff of X on Y.

$\theta = \tan^{-1}\left\{\frac{1-r^2}{|r|}\left(\frac{\sigma x\, \sigma y}{\sigma x^2+\sigma y^2}\right)\right\}$ If $r=0$, $\theta = \pi/2$ $\perp r$
If $r = \pm 1$, $\theta = 0$ $\parallel$el

also $b_{xy} = r\frac{\sigma x}{\sigma y}$, $b_{yx} = r\frac{\sigma y}{\sigma x}$

$b_{xy} \times b_{yx} = r^2$

$$\boxed{r = \sqrt{b_{yx} \times b_{xy}}}$$

Also

$$\left[\frac{b_{xy}}{b_{yx}} = \frac{\sigma x^2}{\sigma y^2}\right]$$

if $\boxed{b_{xy}\ \&\ b_{yx} > 0 \text{ then } r > 0}$

* **if both b_{yx} & b_{xy} are +ive, r is +ive**
If both are neg, r is negative

Q:- From the data given below, calculate the regression equation, taking deviation of items from mean of X and Y series.

X: 6 2 10 4 8
Y: 9 11 5 8 7

Sol^n

$$\bar{X} = \frac{\Sigma X}{N} = \frac{30}{5} = 6$$

$$\bar{Y} = \frac{\Sigma Y}{N} = \frac{40}{5} = 8$$

X	Y	$X-\bar{X}$ $X-6$	$(X-\bar{X})^2$ $(X-6)^2$	$Y-\bar{Y}$ $(Y-8)$	$(Y-\bar{Y})^2$ $(Y-8)^2$	$(X-\bar{X})(Y-\bar{Y})$
6	9	0	0	1	1	0
2	11	-4	16	3	9	-12
10	5	4	16	-3	9	-12
4	8	-2	4	0	0	0
8	7	2	4	-1	1	-2
			$\Sigma = 40$		$\Sigma = 20$	$\Sigma = -26$

$$\sigma x = \sqrt{\frac{\Sigma(X-\bar{X})^2}{N}} = \sqrt{\frac{40}{5}} = \sqrt{8}$$

$$\sigma y = \sqrt{\frac{\Sigma(Y-\bar{Y})^2}{N}} = \sqrt{\frac{20}{5}} = 2$$

$\cancel{\text{Cov } xy,\ r = \frac{\frac{1}{n}\Sigma XY - \frac{\Sigma x}{n}\frac{\Sigma y}{n}}{\sigma x \times \sigma y}} =$

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

$$r = \frac{cov(x, y)}{\sigma_x \times \sigma_y}$$

$$\left[b_{yx} = \frac{cov(x, y)}{\sigma_x^2} \right]$$

$$\left[cov(x, y) = \frac{1}{n} \Sigma xy - \frac{\Sigma x}{n} \frac{\Sigma y}{n} \right]$$

$$cov(x, y) = \frac{1}{5} \times -26 = -5.2$$

Regression line of Y on X is

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

$$y - 8 = \frac{-5.2}{8} (x - 6)$$

$$\boxed{Y = 81.9 - 0.65X}$$

Ans

Regression line of x on y :-

$$b_{xy} = \frac{cov(x,y)}{\sigma_y^2} = \frac{-5.2}{4}$$

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

$$x - 6 = -\frac{5.2}{4}(y - 8)$$

$$\boxed{x = 16.4 - 1.3y}$$ Ans

Q. In a partially destroyed laboratory record of an analysis of correlation data, the following results are legible: Variance of x is equal to 1. The regression equations are $3x + 2y = 26$ and $6x + y = 31$. What were
(i) The mean values of x and y
(ii) the standard deviation of y
(iii) the correlation coeff b/w x & y.

Soln since line of regression intersect at mean ie $(\bar{x}, \bar{y})$
ie we have $6\bar{x} + \bar{y} = 31$, $3\bar{x} + 2\bar{y} = 26$
solving we get

$$\boxed{\bar{x} = 4, \ \bar{y} = 7}$$

Let us assume $3x + 2y = 26$ is regression line of x on y and $6x + y = 31$ is the regression eqn of y on x

ie

$3x + 2y = 26$ & $6x + y = 31$

$x = -\frac{2}{3}y + \frac{26}{3}$ $\qquad$ $y = -6x + 31$

ie $b_{xy} = -2/3$ $\quad b_{yx} = -6$

as $\quad r = \sqrt{b_{xy} \times b_{yx}}$

$r = \sqrt{4}$ $\quad r = \pm 2$ $\qquad$ as $-1 \le r \le 1$

ie our assumption is wrong

ie first eqn is regression line of y on x

ie $y = -\frac{3}{2}x + 13$ ——— ①

and Second eqn is x on y

$x = -\frac{1}{6}y + \frac{31}{6}$ ——— ②

$r = \sqrt{b_{xy} \times b_{yx}}$

$r = \sqrt{\frac{-3}{2} \times \frac{-1}{6}}$ $\qquad$ $b_{xy} = r\sigma_x$

$r = \sqrt{\frac{1}{4}}$

$r = -\frac{1}{2}$ as b_{xy} and b_{yx} are negative

Now

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

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

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

$\boxed{\sigma_x = 1}$

$b_{xy} = -1/6$ $\quad b_{yx} = -\frac{3}{2}$

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

$\sigma_y^2 = 1 \times \frac{-3/2}{-1/6}$

$\sigma_y^2 = 9$ $\quad \sigma_y = \sqrt{9}$

$\boxed{\sigma_y = 3}$

(using assumed mean)

Q Obtain the equation of the lines of regression from the following data.

X:	1	2	3	4	5	6	7
Y:	9	8	10	12	11	13	14

Sol^n

X	Y	$d_x = X - A_x$	$d_y = Y - A_y$	d_x^2	d_y^2	$d_x d_y$
1	9	-3	-2	9	4	6
2	8	-2	-3	4	9	6
3	10	-1	-1	1	1	1
4	12	0	1	0	1	0
5	11	1	0	1	0	0
6	13	2	2	4	4	4
7	14	3	3	9	9	9
N = 7		0	0	28	28	26

Let $A_x = 4$ $A_y = 11$

$$\bar{X} = A + \frac{\Sigma d_x}{N}$$

$$\bar{X} = 4 + 0 \qquad \boxed{\bar{X} = 4}$$

$$\boxed{\bar{Y} = 11}$$

Now $$\sigma_x^2 = \frac{\Sigma d_x^2}{N} - \left(\frac{\Sigma d_x}{N}\right)^2$$

$$\sigma_x^2 = \frac{28}{7} - (0) \qquad \sigma_x^2 = 4$$

$$\boxed{\sigma_x = 2}$$

Slly $$\sigma_y^2 = \frac{\Sigma d_y^2}{N} - \left(\frac{\Sigma d_y}{N}\right)^2 = \frac{28}{7}$$

$$\sigma_y^2 = 4 \qquad \boxed{\sigma_y = 2}$$

$\gamma = cov(x,y) = E(XY) - E(X)E(Y)$

$= \frac{\sum xy}{n}$

$= \frac{\sum dx\,dy}{N} - \frac{\sum dx}{N} \cdot \frac{\sum dy}{N}$ (from assumed mean)

$= \frac{26}{7} = 3.7$

Now

regression line of Y on X

$y - \bar{y} = r \frac{\sigma y}{\sigma x} (x - \bar{x})$

$y - \bar{y} = \frac{cov(x,y)}{\sigma x^2} (x - \bar{x})$

$y - 11 = \frac{3.7}{4} (x - 4)$

$\boxed{3.7x - 4y + 29.2 = 0}$

regression line of X on Y :-

$x - \bar{x} = \frac{cov(x,y)}{\sigma y^2} (y - \bar{y})$

$x - 4 = \frac{3.7}{4} (y - 11)$

$\boxed{4x - 3.7y + 24.7 = 0}$

classmate

Date

Page

Q. Two lines of regressions are $x+2y-5=0$ and $2x+3y-8=0$, and the variance of x is 12. Find the variance of y. and ~~regress~~ coefficient of correlation.

Soln. Let $y = -\frac{1}{2}x + \frac{5}{2}$ be regression line of y on x.

& $x = -\frac{3}{2}y + \frac{8}{2}$ be " " " x on y.

$byx = -1/2 \qquad bxy = -3/2$

$r = \sqrt{byx \times bxy}$

$r = \sqrt{\frac{3}{4}} = \frac{\sqrt{3}}{-2} < 1$ as both bxy & byx are -ive

$\sigma x = \sqrt{12} = 2\sqrt{3}$

$\sigma y = ?$

$bxy = r \frac{\sigma x}{\sigma y}$

$-\frac{\cancel{3}}{2} = \frac{\cancel{\sqrt{3}}}{-2} \frac{2\cancel{\sqrt{3}}}{\sigma y}$

$\sigma y = 2$

$\boxed{\sigma y^2 = 4}$

ie $\boxed{\text{variance of } y = 4}$

$\boxed{r = -\frac{\sqrt{3}}{2}}$

Multiple Regression :-

If no of independent variables in a regression model is more than one, then the model is called a multiple regression model.

The multiple regression eqn describe the average relationship between these variables and this relationship is used to predict or control the dependent variable.

It is given by

$$Y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4$$

Y is dependent variable

x_1, x_2, x_3, x_4 are independent variables.

* For two independent variables x_1, x_2 multiple regression eqn is given by

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

where b_0, b_1, b_2 are determined by solving following normal equations.

$$\Sigma Y = N b_0 + b_1 \Sigma X_1 + b_2 \Sigma X_2$$

$$\Sigma X_1 Y = b_0 \Sigma X_1 + b_1 \Sigma X_1^2 + b_2 \Sigma X_1 X_2$$

$$\Sigma X_2 Y = b_0 \Sigma X_2 + b_1 \Sigma X_1 X_2 + b_2 \Sigma X_2^2$$

Slly Normal eqs can be written for more than two variables.

Q. Find the multiple regression eqⁿ of x_1 on x_2 and x_3 from the following data

x_1	4	6	7	9	13	15
x_2	15	12	8	6	4	3
x_3	30	24	20	14	10	4

Solⁿ ~~Normal eqn.~~

Regression eqⁿ of x_1 on x_2 & x_3

$$x_1 = a + b x_2 + c x_3$$

where a, b, c are calculated by following normal equations

$$\Sigma x_1 = Na + b\Sigma x_2 + c\Sigma x_3$$
$$\Sigma x_1 x_2 = a\Sigma x_1 + b\Sigma x_2^2 + c\Sigma x_1 x_3$$
$$\Sigma x_1 x_3 = a\Sigma x_3 + b\Sigma x_2 x_3 + c\Sigma x_3^2 .$$

x_1	x_2	x_3	$x_1 x_2$	$x_2 x_3$	$x_3 x_1$	x_1^2	x_2^2	x_3^2
4	15	30	60	450	120	16	225	900
6	12	24	72	288	144	36	144	576
7	8	20	56	160	140	49	64	400
9	6	14	54	84	126	81	36	196
13	4	10	52	40	130	169	16	100
15	3	4	45	12	60	225	9	16
Σx_1 $= 54$	Σx_2 $= 48$	Σx_3 $= 102$	$\Sigma x_1 x_2$ $= 339$	$\Sigma x_2 x_3$ $= 1034$	$\Sigma x_1 x_3$ $= 720$	Σx_1^2 $= 576$	Σx_2^2 $= 494$	Σx_3^2 $= 2188$

$N = 6$

Substituting values in Normal equations

$$54 = 6a + 48b + 102c \quad \text{--- (1)}$$
$$339 = 48a + 494b + 103c$$
$$720 = 102a + 1034b + 2188c$$

On solving these system of equations

$$a = 16.479, \quad b = 0.389 \quad c = -0.623$$

$$\boxed{X_1 \doteq 16.479 + 0.389\, X_2 - 0.623\, X_3}$$

Q The following data represent the chemistry grades for a random sample of 12 freshmen at a certain college along with their scores on an intelligence test administered while they were still seniors in high school. The no. of class periods missed is also given.

(a) Fit a multiple linear regression equation of the form

$$\hat{y} = b_0 + b_1 x + b_2 x_2$$

(b) Estimate the chemistry grade for a student who has an intelligence test score of 60 and misses 4 classes.

Q:- The annual sales revenue of a product as a function of sales force and annual advertising expenditure for the past 10 years are summarised.

Annual sales revenue (Y):	20	23	25	27	21	29	22	24	27	35
Sales force (X_1):	8	13	8	18	23	16	10	12	14	20
Annual expenditure (X_2):	28	23	38	16	20	28	23	30	36	32

Solⁿ: Let $y = b_0 + b_1 x_1 + b_2 x_2$

$N = 10$ $\quad \Sigma x_1 = 142, \quad \Sigma x_2 = 264, \quad \Sigma y = 253$

$\Sigma x_1^2 = 2246 \quad \Sigma x_2^2 = 7326$

$\Sigma y x_1 = 3678 \quad \Sigma y x_2 = 6751$

$\Sigma x_1 x_2 = 3617$

for Normal equations

ie $\Sigma y = N b_0 + b_1 \Sigma x_1 + b_2 \Sigma x_2$

$\Sigma x_1 y = b_0 \Sigma x_1 + b_1 \Sigma x_1^2 + b_2 \Sigma x_2 x_1$

$\Sigma x_2 y = b_0 \Sigma x_2 + b_1 \Sigma x_1 x_2 + b_2 \Sigma x_2^2$

$253 = 10 b_0 + 142 b_1 + 264 b_2$ — (1)

$3678 = 142 b_0 + 2246 b_1 + 3617 b_2$ — (2)

$6751 = 264 b_0 + 3617 b_1 + 7326 b_2$ — (3)

$$\boxed{y = 5.1483 + 0.6190 x_1 + 0.4304 x_2}$$

Solution