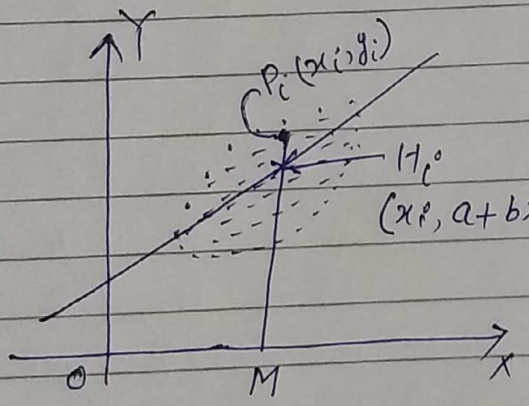
Linear and curvilinear regression

Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of original units of data.

Linear Regression

If the variables in a bivariate distribution are related we will find that the points in the scatter diagram will cluster round some curve called the "curve of regression". If the curve is straight line, it is called the line of regression and there is said to be linear regression between variables, otherwise regression is said to be curvilinear.



The line of regression is the line of "best fit" and is obtained by the principle of least square

Let the line of regression of Y on X be

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

Here our aim is to determine the constants a and b so that the line given by (1) is best fit

Let $P_i(x_i, y_i)$ be any general point on the scatter diagram.

Draw P_iM bar to x-axis, ~~not~~ which touches the x-axis at point M

The bar line P_iM meets the best fit line at point H_i. Since H_i is at the line $y = a+bx$ so the co-ordinate of H_i is $(x_i, a+bx_i)$

$P_iH_i = P_iM - H_iM$

$= y_i - (a+bx_i)$

According to the principal of least square we have to determine 'a' and 'b' so that

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

is minimum.

From the principle of maxima and minima, the partial derivatives of E; w.r.t a and b should be zero.

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

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

$$\Rightarrow \begin{cases} \sum_{i=1}^{n} y_i = na + b \sum_{i=1}^{n} x_i \quad [\text{from A}] \quad \text{---(2)} \\ \sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2 \quad [\text{from B}] \quad \text{---(3)} \end{cases}$$

These two equations are known as normal equations for estimating a and b.

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

[Dividing both sides by n]

$$\Rightarrow \bar{y} = a + b\bar{x}$$

Thus the line of regression passes through the point $(\bar{x}, \bar{y})$

Remark

How to write the normal equations for the line of regression of y on x

$$y = a + bx \quad \text{—— (1)}$$

Take summation on both sides

$$\Sigma y = an + b\Sigma x \quad \text{—— (2)}$$

Multiply (1) by x and take summation

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

Ex (1) Find the line of regression of y on x

x:	1	2	3	4	5
y:	8	7	5	9	11

x	y	x^2	xy
1	8	1	8
2	7	4	14
3	5	9	15
4	9	16	36
5	11	25	55
Total 15	40	55	128

The normal eqns are

$\Sigma y = na + b\Sigma x$

$\Sigma xy = a\Sigma x + b\Sigma x^2$

$40 = 5a + 15b$

$128 = 15a + 55b$

We need to solve these two equations to find a and b

$5a + 15b = 40 \Rightarrow 15a + 45b = 120$

$15a + 55b = 128$

$-10b = -8$

$b = \frac{-8}{-10} = \frac{4}{5}$

also, $5a = 40 - 15b$

$= 40 - 15 \cdot \frac{4}{5} = 40 - 12 = 28$

$a = \frac{28}{5}$

The line of regression of y on x is $\boxed{y = \frac{28}{5} + \frac{4}{5}x}$ ∴ (*)

Observation

$$\bar{x} = \frac{15}{5} = 3, \quad \bar{y} = \frac{40}{5} = 8$$

Substituting $\bar{x} = 3$ and $\bar{y} = 8$ in (*)

$$8 = \frac{28}{5} + \frac{4}{5} \cdot 3 = \frac{28}{5} + \frac{12}{5} = \frac{40}{5} = 8$$

So we note that $(\bar{x}, \bar{y})$ i.e. $(3, 8)$ lies on the regression line.

In the above discussion we have seen that when we are writing the regression line of y on x, we are considering x as the independent variable and y as the dependent variable, means for any given value of x, we can estimate the corresponding value of y through the regression line.

Now the question is, can we estimate the value of x from a given value of y from the same regression line, the answer is 'no'.

~~For estim~~ For estimating the value of x for a given value of y, we need another regression line, ie the regression line of x on y [The reason of this is that to find the regression line of y on x, we have minimised the sum of squares of distance parallel to y-axis, while to find the regression line of x on y, we have to minimise the sum of square of distances parallel to x-axis.

The normal equation for the regression line of x on y is $x = a + by$

$$x = a + by \quad \text{——} (1)$$

Take summation on both sides

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

multiply (1) by y on both sides and then take the summation

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

Another way to write lines of Regressions

The line of regression of y on x

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

and

The line of regression of x on y

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

'r' is the correlation coefficient between x and y.

Regression coefficients

Regression coefficient of y on x $= r\frac{\sigma_y}{\sigma_x}$

Regression coefficient of x on y $= r\frac{\sigma_x}{\sigma_y}$

Properties of Regression Coefficient

(a) Correlation coefficient is the geometric mean between regression coefficients.

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

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

The sign of the regression coefficients and the sign of correlation coefficient is same.

(b) If one of the regression coefficients is greater than unity, the other must be less than unity.

(c) The modulus value of the arithmetic mean of the regression coefficients is not less than the modulus value of the correlation coefficient r.

ie $\left|\frac{1}{2}(b_{yx} + b_{xy})\right| \geq |r|$

(d) Regression coefficients are independent of the change of origin but not of scale.

Angle Between two lines of Regression:

The equations of the lines of regression of y on x and x on y are

$$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) \quad \text{---} (1)$$

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

$$\Rightarrow y - \bar{y} = \frac{\sigma_y}{r\sigma_x}(x - \bar{x}) \quad \text{---} (2)$$

Slope of the lines are $r\frac{\sigma_y}{\sigma_x}$ and $\frac{\sigma_y}{r\sigma_x}$ respectively.

If θ is the acute angle between them

$$\tan\theta = \left|\frac{r\frac{\sigma_y}{\sigma_x} - \frac{\sigma_y}{r\sigma_x}}{1 + r^2\frac{\sigma_y}{\sigma_x}\cdot\frac{\sigma_y}{r^2\sigma_x}}\right|$$

$$= \left| \frac{\frac{r^2\sigma_y - \sigma_y}{r\sigma_x}}{\frac{\sigma_x^2 + r\sigma_y^2}{r\sigma_x^2}} \right| = \frac{(r^2-1)\sigma_y}{r\sigma_x} \times \frac{r\sigma_x^2}{\sigma_x^2+\sigma_y^2}$$

$$= \left|\frac{(r^2-1)}{r}\right| \left(\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}\right)$$

$$= \frac{1-r^2}{|r|} \frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2} \qquad (\because r^2 \le 1 \Rightarrow r^2-1 \le 0, \; |r^2-1| = 1-r^2)$$

$$\therefore \theta = \tan^{-1}\left\{\frac{1-r^2}{|r|} \frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}\right\}$$

Case I : $r = 0$, If $r=0$, $\tan\theta = \infty \Rightarrow \theta = \frac{\pi}{2}$

Thus if the two variables are uncorrelated, the lines of regression become perpendicular to each other.

Case II : $r = \pm 1$, If $r = \pm 1$

$$\tan\theta = 0 \Rightarrow \theta = 0 \text{ or } \pi$$

In this case two lines of regression either coincide or they are parallel to each other. But since both lines of regression pass through the point $(\bar{x}, \bar{y})$, they cannot be parallel. Hence in case of perfect correlation, positive or negative, the two lines of regression

Example

① Obtain the equations for two lines of regression for the following data. Also obtain the estimate of x for Y = 70

X:	65	66	67	67	68	69	70	72
Y:	67	68	65	68	72	72	69	71

Sol

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
544	552	37028	38132	37560

Let $y = a+bx$ be the line of regression of y on x

$y = a+bx$

Normal equations are given by

$\Sigma y = na + b\Sigma x$ i.e. $552 = 8a + 544b$ — (1)

$\Sigma xy = a\Sigma x + b\Sigma x^2$ $37560 = 544a + 37028b$ — (2)

Multiply (1) by 68, $37536 = 544a + 36992b$

$37560 = 544a + 37028b$

$\Rightarrow -24 = -36b$

$\Rightarrow b = \frac{-24}{-36} = \frac{2}{3} = 0.66$

From (1), $8a = 552 - 544 \times 2/3 = 552 - 362.66$

$\Rightarrow 8a = 189.34 \Rightarrow a = \frac{189.34}{8} = 23.66$

The line of eqn of y on x, is $\boxed{y = 23.66 + 0.66x}$

Similarly The line of regression of x on y is

$x = a + by$

The Normal equations are $\Sigma x = na + b\Sigma y$

$\Sigma xy = a\Sigma y + b\Sigma y^2$

ie the normal equations are

$544 = 8a + 552b$ ——— (1)

$37560 = 552a + 38132b$ ——— (2)

Multiplying (1) by 69, we have

$37536 = 552a + 38088b$

$37560 = 552a + 38132b$

$\Rightarrow -24 = -44b \Rightarrow b = \frac{-24}{-44} = 0.545$

From (1), $8a = 544 - 552 \times \frac{24}{44}$

$= 544 - 301.09 = 242.91$

$\Rightarrow a = \frac{242.91}{8} = 30.36$

The line of regression $x = 30.36 + 0.545y$

For $y = 70$, $x = 30.36 + 0.545 \times 70$

$x = 68.51$

Q(2) In a partially destroyed laboratory, record of an analysis of correlation data. the following results only are legible:

Variance of $X = 9$. Regression equations: $8X - 10Y + 66 = 0$

$40X - 18Y = 214$.

What are (i) the mean values X and Y

(ii) the correlation coefficient between X and Y, and

(iii) Standard deviation of Y ?

Sol: (i) Since both lines of regression pass through the point $(\bar{X}, \bar{Y})$, we have

$8\bar{X} - 10\bar{Y} = -66$ — (1)

$40\bar{X} - 18\bar{Y} = 214$ — (2)

$40\bar{X} - 50\bar{Y} = -330$

$40\bar{X} - 18\bar{Y} = 214$

$\Rightarrow -32\bar{Y} = -544$

$\bar{Y} = \frac{-544}{-32} = 17$

From (1) $8\bar{X} = -66 + 10\bar{Y} = -66 + 170 = 104$

$\bar{X} = \frac{104}{8} = 13$

(ii) let $8X - 10Y + 66 = 0$ and $40X - 18Y = 214$ be the regression lines of Y on X and X on Y respectively

$Y = \frac{8}{10}X + \frac{66}{10}$, $X = \frac{18}{40}Y + \frac{214}{40}$

$b_{YX} = \frac{8}{10} = \frac{4}{5}$ $\qquad$ $b_{XY} = \frac{18}{40} = \frac{9}{20}$

$r^2 = b_{YX} \times b_{XY} = \frac{8}{10_5} \times \frac{9\,18}{40_5} = \frac{9}{25}$

$r = \pm \frac{3}{5} = \pm 0.6$

(iii) $b_{YX} = r\frac{\sigma_Y}{\sigma_X} \Rightarrow \frac{4}{5} = r \times \frac{\sigma_Y}{3}$

$\Rightarrow \frac{4}{5} = \frac{3}{5} \times \frac{\sigma_Y}{3} \Rightarrow \boxed{\sigma_Y = 4}$

Remarks **1.** It can be verified that the values of $\bar{X} = 13$ and $\bar{Y} = 17$ as obtained in part (*i*) satisfy both the regression equations. In numerical problems of this type, this check should invariably be applied to ascertain the correctness of the answer.

2. If we had assumed that $8X - 10Y + 66 = 0$, is the equation of the line of regression of X on Y and $40X - 18Y = 214$ is the equation of line of regression of Y on X, then we get respectively :

$$8X = 10Y - 66 \quad \text{and} \quad 18Y = 40X - 214$$

$$\Rightarrow \quad X = \frac{10}{8}Y - \frac{66}{8} \quad \text{and} \quad Y = \frac{40}{18}X - \frac{214}{18}$$

$$\Rightarrow \quad b_{XY} = \frac{10}{8} \quad \text{and} \quad b_{YX} = \frac{40}{18}$$

$$\therefore \quad r^2 = b_{XY} \cdot b_{YX} = \frac{10}{8} \times \frac{40}{18} = 2{\cdot}78$$

But since r^2 always lies between 0 and 1, our supposition is wrong.

Example 11·3. *Find the most likely price in Mumbai corresponding to the price of Rs. 70 at Kolkata from the following :*

	Kolkata	*Mumbai*
Average price	65	67
Standard deviation	2·5	3·5

Correlation coefficient between the prices of commodities in the two cities is 0·8.

Solution. Let the prices (in Rupees) in Kalkata and Mumbai be denoted by X and Y respectively. Then we are given :

$\bar{X} = 65$, $\bar{Y} = 67$, $\sigma_X = 2{\cdot}5$, $\sigma_Y = 3{\cdot}5$ and $r = r(X, Y) = 0{\cdot}8$. We want Y for $X = 70$.

Line of regression of Y on X is :

$$Y - \bar{Y} = r \frac{\sigma_Y}{\sigma_X} (X - \bar{X}) \quad \Rightarrow \quad Y = 67 + 0{\cdot}8 \times \frac{3{\cdot}5}{2{\cdot}5} (X - 65)$$

When $X = 70$, $\hat{Y} = 67 + 0{\cdot}8 \times \frac{3{\cdot}5}{2{\cdot}5} ((70 - 65) = 72{\cdot}6$

Hence, the most likely price in Mumbai corresponding to the price of Rs. 70 at Kolkata is Rs. 72·60.

Example 11·4. *Can $Y = 5 + 2{\cdot}8\,X$ and $X = 3 - 0{\cdot}5Y$ be the estimated regression equations of Y on X and X on Y respectively ? Explain your answer with suitable theoretical arguments.*

Solution. Line of regression of Y on X is : $Y = 5 + 2{\cdot}8\,X \quad \Rightarrow \quad b_{YX} = 2{\cdot}8 \quad \ldots (*)$

Line of regression of X on Y is : $X = 3 - 0{\cdot}5Y \quad \Rightarrow \quad b_{XY} = -0{\cdot}5 \quad \ldots (**)$

This is not possible, since each of the regression coefficients b_{YX} and b_{XY} must have the same sign, which is same as that of Cov (X, Y). If Cov (X, Y) is positive, then both the regression coefficients are positive and if Cov (X, Y) is negative, then both the regression coefficients are negative. Hence (*) and (**) cannot be the estimated regression equations of Y on X and X on Y respectively.

Curvilinear Regression

As we have seen earlier that linear regression was based on the fact that, linear relation exist between x and y. But in certain cases the scatter diagram gives us an idea that curvilinear regression explains the relationship better.

① Fitting the curve $y = a + b_1 x + b_2 x^2$ — (A)

The normal equations are given by

Multiply (A) by x and take summation on both sides to get (2)

$$\Sigma y = na + b_1 \Sigma x + b_2 \Sigma x^2 \quad \text{—(1)}$$

$$\Sigma xy = a\Sigma x + b_1 \Sigma x^2 + b_2 \Sigma x^3 \quad \text{—(2)}$$

$$\Sigma x^2 y = a\Sigma x^2 + b_1 \Sigma x^3 + b_2 \Sigma x^4 \quad \text{—(3)}$$

To get (3), multiply (A) by x^2 and take summation both sides.

From these we find A and b and consequently $\log$

Example 11·6. *For 10 randomly selected observations, the following data were recorded :*

Observation No.	:	1	2	3	4	5	6	7	8	9	10
Overtime hrs. (X)	:	1	1	2	2	3	3	4	5	6	7
Additional units (Y)	:	2	7	7	10	8	12	10	14	11	14

Determine the coefficients of regression and regression equation using the non-linear form : $Y = a + b_1 X + b_2 X^2$.

Solution.

S. No	X	Y	X^2	X^3	X^4	XY	$X^2 Y$
1	1	2	1	1	1	2	2
2	1	7	1	1	1	7	7
3	2	7	4	8	16	14	28
4	2	10	4	8	16	20	40
5	3	8	9	27	81	24	72
6	3	12	9	27	81	36	108
7	4	10	16	64	256	40	160
8	5	14	25	125	625	70	350
9	6	11	36	216	1296	66	396
10	7	14	49	343	2401	98	686
Total	34	95	154	820	4774	377	1849

Using normal equations (11·15*a*), we get

$10a + 34\, b_1 + 154\, b_2 = 95$, $34a + 154\, b_1 + 820\, b_2 = 377$, and $154\, a + 820\, b_1 + 4774\, b_2 = 1849$.

The solutions to these three simultaneous equations are :

$$a = 1{\cdot}80, \quad b_1 = 3{\cdot}48 \quad \text{and} \quad b_2 = -0{\cdot}27$$

The regression equations, therefore, is :

$$Y = 1{\cdot}80 + 3{\cdot}48\, X - 0{\cdot}27 X^2.$$

(2) Fitting of a Power curve $y = ax^b$

Taking log of each sides

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

$$\Rightarrow U = A + bV \quad \text{——— (1)}$$

where $U = \log y$, $A = \log a$, $V = \log x$

This is a linear linear in V and U

Normal equations for estimating A and B are

$$\Sigma U = nA + b\Sigma V \quad \text{——— (2)}$$

$$\Sigma UV = A\Sigma V + b\Sigma V^2 \quad \text{——— (3)}$$

Solve (2) and (3) to get A and b, and consequently $a = \underline{\text{antilog}(A)}$

Example 11·7. *Fit an exponential curve of the form* $Y = ab^X$ *to the following data :*

X :	1	2	3	4	5	6	7	8
Y :	1·0	1·2	1·8	2·5	3·6	4·7	6·6	9·1

Solution.

S. No.	X	Y	$U = \log Y$	XU	X^2
1	1	1·0	0·0000	0·0000	1
2	2	1·2	0·0792	0·1584	4
3	3	1·8	0·2553	0·7659	9
4	4	2·5	0·3979	1·5916	16
5	5	3·6	0·5563	2·7815	25
6	6	4·7	0·6721	4·0326	36
7	7	6·6	0·8195	5·7365	49
8	8	9·1	0·9590	7·6720	64
Totals	36	30·5	3·7393	22·7385	204

(11·18*a*) gives the normal equation as :

$$3·7393 = 8A + 36B \quad \text{and} \quad 22·7385 = 36A + 204B$$

Solving, we get

$$B = 0·1408 \quad \text{and} \quad A = -0·1662 = \bar{1}·8338$$

$\therefore$ $b = \text{Antilog } B = 1·383$ and $d = \text{Antilog } A = 0·6821$

Hence the equation of the required curve is : $Y = 0·6821\ (1·38)^X$.

11·4. REGR

③ Fitting of Exponential Curves

ⓐ $y = ab^x$ —— ①

Taking log on both sides

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

Taking $U = \log y$, $A = \log a$, $B = \log b$

We have, $U = A + BX$

The normal equations are given by

$\sum U = nA + B\sum X$ —— ②

$\sum UX = A\sum X + B\sum x^2$ —— ③

Solve ② and ③ to get A and B

Finally $a = \text{antilog}(A)$

$b = \text{antilog}(B)$

ⓑ $y = ae^{bx}$

Taking log on both sides

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

~~Take $U = \log y$, $A = \log a$,~~

ⓑ $y = ae^{bx}$

Taking log both sides

$\log y = \log a + bx \log e = \log a + (b \log e)X$

$U = A + BX$

where $U = \log y$, $A = \log a$, $B = b \log e$

The normal equations are given by

~~$\log U$~~ — $\sum U = nA + B\sum X$ —— ②

$\sum UX = A\sum X + B\sum x^2$ —— ③

Solving ② and ③, we get A and B

and finally $a = \text{antilog}(A)$

$b = \dfrac{B}{\log e}$