Correlation and Regressions

Main topic here is to fit a linear curve (or parabolic curve) corresponding to Two Random Variables X and Y using sampling i.e.; using statistics

Sense of relation between X and Y

- Husband's age and wife's age.
- Price and demand.
- Income and expenditure.
- Height and Weight (recall BMI).
- Height of Fathers and Sons.

Are they related ?

- Runs scored by a batsman and consumption of fertilizer in the local market.
- Number of flights in space and population of tigers.

Meaning of correlation

If changing of X affects a change in the other variable Y,

then X and Y are said to be correlated.

Positive Correlation

- If X and Y both increase (or decrease) together then it is said to be direct or positive correlation.
- Example (1) Height and weight (2) Income and expenditure

Negative Correlation

- If when X increase then Y decrease (or Y increase then X decrease) then it is said to be diverse or negative correlation.
- Example: (1) Price and demand. (2) Volume and pressure.

Scatter Diagram

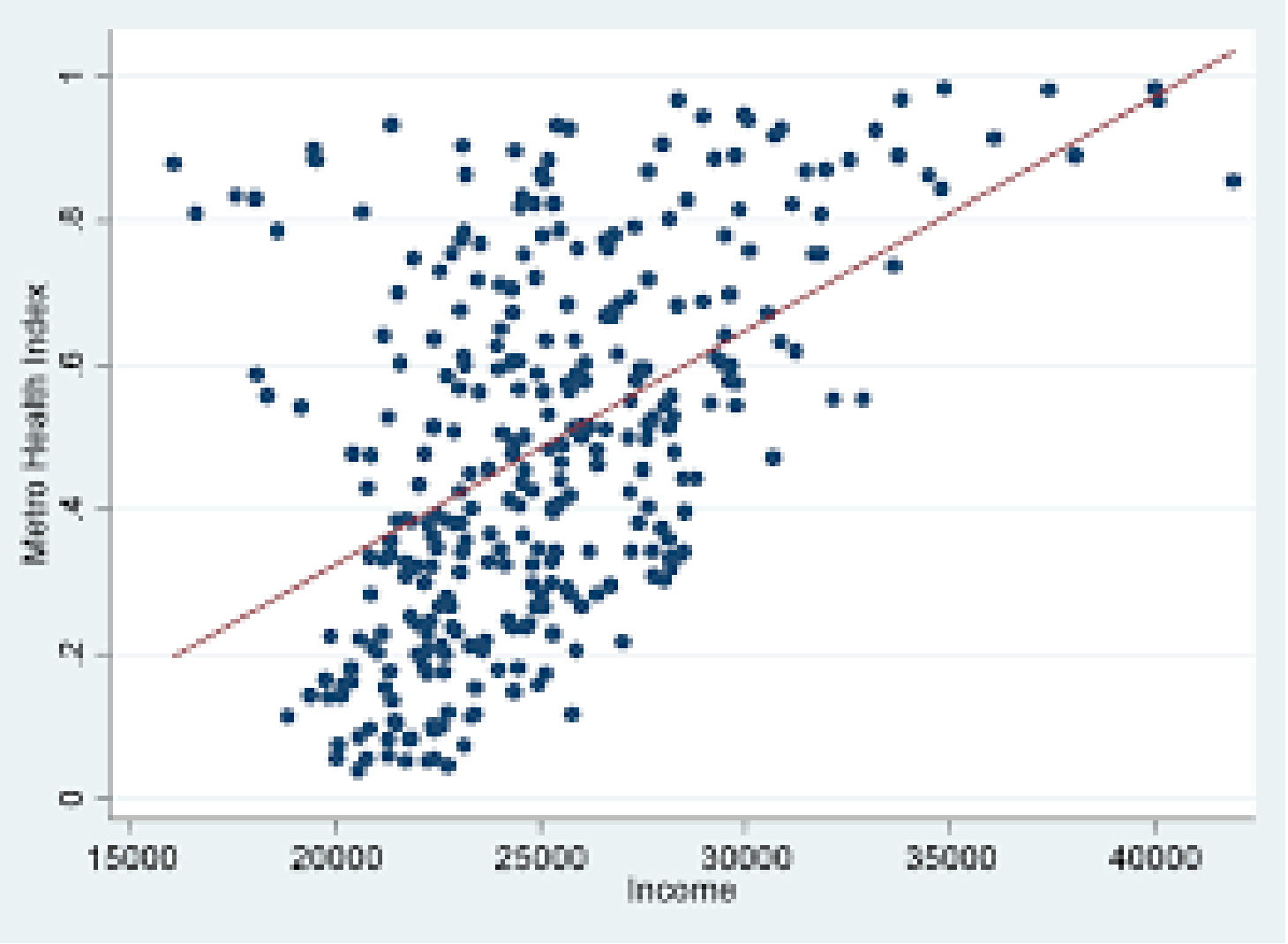
Take a bivariate sample $(x_i, y_i), i = 1, 2, 3, \ldots, n$.

If we plot this date on x-y plane taking the values

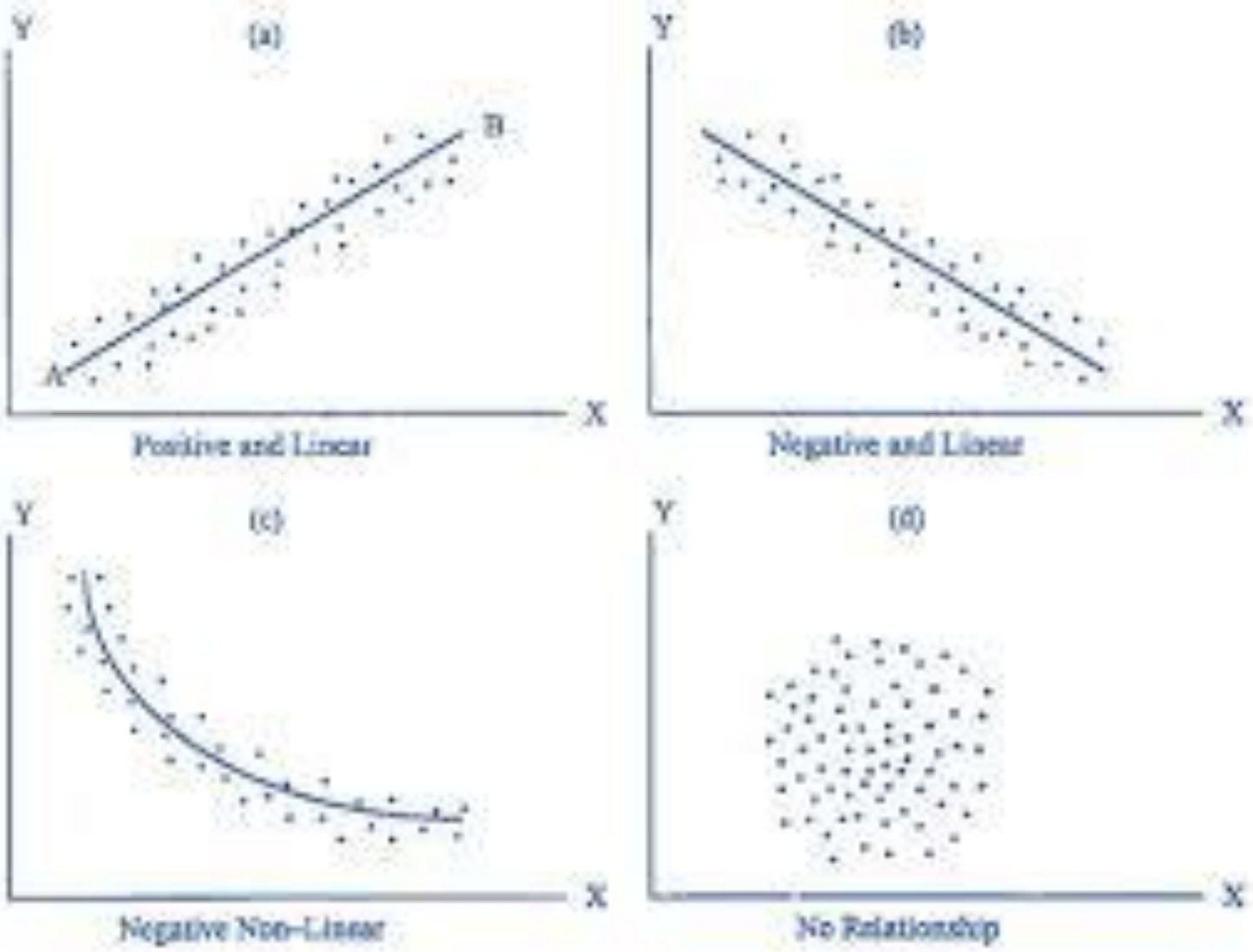
X on x-axis and the value of Y on y-axis, the diagram of

dots so obtained is known as scatter diagram.

Scatter Diagram



Scatter Diagram



Karl Pearson's Coefficient of Correlation .

Correlation coefficient between two random variables X and Y, denoted by r(X,Y) or r_{XY}, is a numerical measure Of linear relationship between them and is defined by

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

Karl Pearson's Coefficient of Correlation .

Take a bivariate sample $(x_i, y_i), i = 1, 2, 3, \ldots, n$.

Then $Cov(X, Y) = E[\{X - E(X)\}\{Y - E(Y)\}]$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \mu_{11}) .$$

$$\sigma_X^2 = E[\ \{X - E(X)\}^2] \quad \sigma_Y^2 = E[\ \{Y - E(Y)\}^2]$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Limit of Correlation coefficient

$$-1 \leq r(X,Y) \leq 1$$

Theorem 1

Correlation coefficient is independent of change of Origin and scale i.e.; if $U = \frac{X-a}{h}, V = \frac{Y-b}{k}$, $h, k > 0$ then $r(X,Y) = r(U,V)$

Theorem 2

$$r(aX + b, cY + d) = \frac{ac}{|ac|} r(X, Y), a \neq 0, c \neq 0$$

Theorem 3

- If X and Y are independent then r(X,Y)=0

NOTE

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1n}^{n} x_i y_i - \overline{x}\,.\,\overline{y}$$

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1`}^{n} x_i^2 - \overline{x}^2$$

$$\sigma_Y^2 = \frac{1}{n}\sum_{i=1`}^{n} y_i^2 - \overline{y}^2$$

Example 1

Calculate the correlation coefficient for the following heights of fathers(X)

And their sons (Y):

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

X	Y	XY	X^2	Y^2	U=X-68	V=Y-69	U^2	V^2	UV
65	67	4355	4225	4489	-3	-2	9	4	6
66	68	4488	4356	4624	-2	-1	4	1	2
67	65	4355	4489	4225	-1	-4	1	16	4
67	68	4556	4489	4624	-1	-1	1	1	1
68	72	4896	4624	5184	0	3	0	9	0
69	72	4968	4761	5184	1	3	1	9	3
70	69	4830	4900	4761	2	0	4	0	0
72	71	5112	5184	5041	4	2	16	4	8
544	552	37560	37028	38132	0	0	36	44	24
68	69	4695	4629	4767					
		3	4.5	5.5					

Example 2

A computer while calculating correlation coefficient between two Variables X and Y from 25 observations obtained the following results n=25, $\sum X = 125, \sum X^2 = 650, \sum Y = 100, \sum Y^2 = 460, \sum XY = 508.$ If was, however, later discovered at the time of checking that he had Copied down two pairs as

X	Y
6	14
8	6

While the correct values were

X	Y
8	12
6	8

Find the correct value of the Correlation of coefficient.

Example 3

- The variables X and Y are connected by the equation $aX+bY+c=0$. Show that the correlation coefficient between them is -1 if the signs of a and b are alike and +1 if they are different.

Rank Correlation

Let $(x_i, y_i), i = 1,2, \dots, n$ be the ranks of n persons on two characteristics (say, intelligence and beauty, or Math and Physics etc.). we want to calculate the Correlation coefficient.

Spearman's Rank Correlation coefficient

For no two individual are bracketed equal

in either classification

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \qquad d_i = (x_i - y_i)$$

Example 1

Ten participants in a contest are ranked by two judges as

X	1	6	5	10	3	2	4	9	7	8
Y	6	4	9	8	1	2	3	10	5	7

Calculate the rank correlation coefficient $\rho(X, Y)$.

X	Y	d=X-Y	d^2
1	6	-5	25
6	4	2	4
5	9	-4	16
10	8	2	4
3	1	2	4
2	2	0	0
4	3	1	1
9	10	-1	1
7	5	2	4
8	7	1	1
			60
			0.636364

Example 2

The rank of same 16 students in Mathematics and Physics Are as follows: (1,1), (2,10), (3,3), (4,4), (5,5), (6,7), (7,2), (8,6), (9,8), (10,11), (11,15), (12,9), (13,14), (14,12), (15,16), (16,13). Calculate the rank correlation coefficient.

A	B	C	D
X	Y	d=X-Y	d^2
1	1	0	0
2	10	-8	64
3	3	0	0
4	4	0	0
5	5	0	0
6	7	-1	1
7	2	5	25
8	6	2	4
9	8	1	1
10	11	-1	1
11	15	-4	16
12	9	3	9
13	14	-1	1
14	12	2	4
15	16	-1	1
16	13	3	9
			136
			0.8

Repeated Rank

If any two or more individuals are bracketed equal in any classification with respect to characteristics A and B then Spearman's formula for the rank correlation coefficient breaks down, since in this case each of the variables X and Y does not assume $1,2,3,\ldots,n$ and $\bar{x} \neq \bar{y}$.

In this case, common rank are given to the repeated items, This common rank is the average of the ranks which these items would have assumed if they were slightly different from each other and the next item will get the rank next to the ranks already assumed.

In the formula, we add the factor $\frac{m(m^2-1)}{12}$ to $\sum d^2$, where m is the number of times an item is repeated.This correction factor is to be added for each repeated value in both X and Y series.

Example 1

Q: Obtain the rank correlation coefficient for the following data:

X	68	64	75	50	64	80	75	40	55	64
Y	62	58	68	45	81	60	68	48	50	70

Example 1

X	Y	rank X	rank Y	d	d^2
64	81	6	1	5	25
64	70	6	2	4	16
75	68	2.5	3.5	-1	1
75	68	2.5	3.5	-1	1
68	62	4	5	-1	1
80	60	1	6	-5	25
64	58	6	7	-1	1
55	50	8	8	0	0
40	48	10	9	1	1
50	45	9	10	-1	1
				0	72

Regression

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

Linear Regression

If there exists a relation between X an Y as $Y=a+bX$ (or $X=a+bY$).

Otherwise it is called **Curvilinear** regression.

Principle of least squares

Let $(x_i, y_i), i = 1,2,\dots,n,$ be a bivariate sample.

Let the line of regression of Yon X be Y=a+bX.

We want to minimize $E = \sum_{i=1}^{n} (y_i - a - bx_i)^2$

With respect to a and b.

Principle of least squares

So $\frac{\delta E}{\delta a} = 0$ $\boldsymbol{and}$ $\frac{\delta E}{\delta b} = 0$. These produce the equations $\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i$ and

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

These equations are known as **Normal** equations.

Line of regression of Y on X

Solving the normal equations we get the line of regression of Y on X as $y - \bar{y} = \frac{cov(X,Y)}{\sigma_X^2}(x - \bar{x})$.

This can by written as $y - \bar{y} = \frac{r(X,Y)\sigma_Y}{\sigma_X}(x - \bar{x})$

Or $y - \bar{y} = b_{YX}(x - \bar{x})$.

Line of regression of X on Y

$$x - \overline{x} = \frac{r(X,Y)\sigma_x}{\sigma_y}(y - \overline{y})$$

Or $x - \overline{x} = b_{XY}(y - \overline{y})$.

Regression Coefficient

b_{YX} = *Regression* Coefficient of Y on X

$$= \frac{r(X,Y)\sigma_y}{\sigma_x}.$$

b_{XY} = *Regression* Coefficient of X on Y

$$= \frac{r(X,Y)\sigma_x}{\sigma_y}.$$

Example 1

Obtain the equations of two lines of regression for the following data.

Also estimate of X for Y=70.

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

$$r(X,Y) = 0.603, \sigma_X^2 = 4.5, \sigma_Y^2 = 5.5, \overline{X} = 68, \overline{Y} = 69$$

In a partially destroyed laboratory, record of an analysis of correlation data, the following results are legible:

Variance of X=9. Regression equations 8X-10Y+66=0, 40X-18Y=214. What are:

(i) the mean values of X and Y,

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

(iii) the standard deviation of Y?

Properties of Regression coefficients property-1

Correlation coefficient is the geometric mean

of the regression coefficients.

$$b_{YX}.b_{XY} = \left(r.\frac{\sigma_Y}{\sigma_X}\right)(r.\frac{\sigma_X}{\sigma_Y})$$

$$so, r^2 = b_{YX}.b_{XY}$$

$$or, r = \pm\sqrt{b_{YX}.b_{XY}}$$

Note $r(X,Y), b_{YX}$ *and* b_{XY} *must have same sign.*

Property 2

Regression coefficients are independent of the Change of origin but not of scale. So if $U = \frac{X-a}{h}, V = \frac{Y-b}{k}, h, k > 0$. Then $r(U,V) = r(X,Y)$,but $b_{VU} \neq b_{YX}$. The result is $b_{YX} = \left(\frac{k}{h}\right) b_{VU}$.

And $b_{XY} = \left(\frac{h}{k}\right) b_{UV}$

Property 3
Angle between regression lines

If θ is an acute angle between the two lines $\tan\theta = \left|\frac{m_1 - m_2}{1 - m_1 m_2}\right|$, where m_1 and m_2 are the slopes of line.

Here $m_1 = b_{yx}$ and $m_2 = \frac{1}{b_{xy}}$, so $\tan\theta = \left|\frac{\left(b_{yx} - \left(\frac{1}{b_{xy}}\right)\right)}{1 - b_{yx}\left(\frac{1}{b_{xy}}\right)}\right|$

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

Angle between regression lines

If θ_1 is an obtuse angle then $\theta_1 = \pi - \theta$

Note: if $r = 0$ $then$ $\theta = \frac{\pi}{2}$ $i.e., lines$ are $perpendicular.$

If $r = \pm 1$ $then$ $\theta = 0$ or π $, so$ $they$ $coincide.$