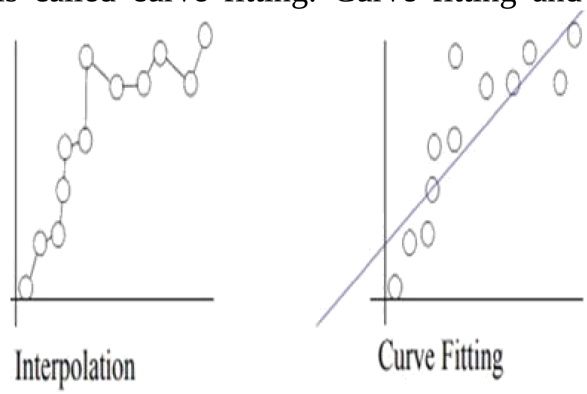
Chapter 4

Curve Fitting & Correlation

4.1 Introduction

The process of constructing an approximate curve $y = f(x)$, which fit best to a given discrete set of points $(x_i, y_i),\ i = 1, 2,3, \dots, n$ is called curve fitting. Curve fitting and interpolation are closely associated procedures. In interpolation, the fitted function should pass through all given data points; whereas curve fitting methodologically fits a unique curve to the data points, which may or may not lie on the fitted curve. The difference between interpolation and curve fitting; while attempting to fit a linear function; is illustrated in the adjoining figure.

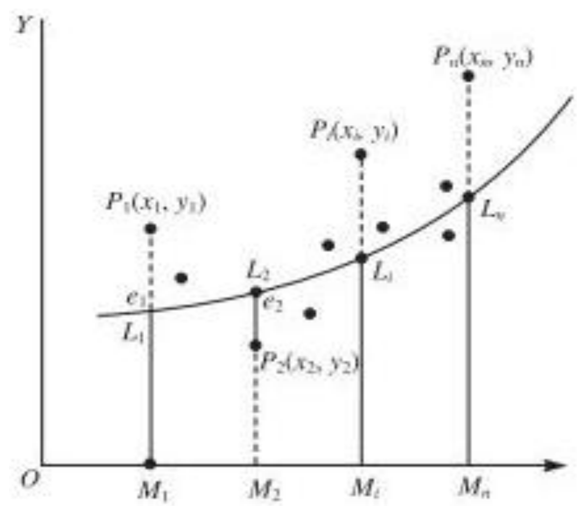


4.2 Principle of Least Squares

The principle of least squares is one of the most popular methods for finding the curve of best fit to a given data set $(x_i, y_i),\ i = 1, 2,3, \dots, n$.

Let $y = f(x)$ be the equation of the curve to be fitted to the given set of points $P_1(x_1, y_1)$, $P_2(x_2, y_2)$,$\cdots$,$P_n(x_n, y_n)$.

Then at a point $x = x_i$, the observed value of the ordinate is $P_iM_i = y_i$ say and let the expected (theoretical) value be $f(x_i)$, shown by L_iM_i in the adjoining figure. The difference between the observed and expected values is the error $e_i = P_iL_i$



Then $e_1 = y_1 - f(x_1)$

$$e_2 = y_2 - f(x_2)$$

$$\vdots$$

$$e_n = y_n - f(x_n)$$

Squaring each error e_i (to take care of negative errors) and adding, we get

$E = e_1^2 + e_2^2 + \cdots . + e_n^2 = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - f(x_i))^2$. The curve of best fit is that for which E is minimum. This is called the Principle of least squares.

4.2.1 Fitting a straight line $y = ax + b$.

Let $y = ax + b$ be the straight line to be fitted to the given set of data points (x_1, y_1), (x_2, y_2),……, (x_n, y_n).

Then $\quad e_i = y_i - f(x_i) = y_i - (ax_i + b), \qquad i = 1, 2, \dots, n$

$$\therefore E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \big(y_i - (ax_i + b)\big)^2$$

Now by the principle of least square, for the curve of best fit, E is minimum

$$\therefore \frac{\partial E}{\partial a} = 0 \text{ and } \frac{\partial E}{\partial b} = 0$$

$\therefore \dfrac{\partial E}{\partial a} = 0$ and $\dfrac{\partial E}{\partial b} = 0$

$$\Rightarrow \sum_{i=1}^{n} 2\left(y_i - (ax_i + b)\right)(-x_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} (y_i x_i - a x_i^2 - b x_i) = 0$$

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

$$\Rightarrow \sum_{i=1}^{n} 2(y_i - ax_i - b)(-1) = 0$$

$$\Rightarrow \sum_{i=1}^{n} (y_i - ax_i - b) = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i = a \sum_{i=1}^{n} x_i + nb \qquad \dots ②$$

Since x_i, y_i are known set of values, equation ① and ② are two equation with variables a and b, ignoring the suffices, ② and ① can be rewritten as

$$\sum y = a \sum x + nb \qquad \dots ③$$

and $\quad \sum xy = a \sum x^2 + b \sum x \qquad \dots ④$

③and ④are known as Normal equation for fitting a straight line $y = ax + b$

- If the equation of line is taken as $y = a + bx$
  we get normal equations as: $\sum y = na + b \sum x$ and $\sum xy = a \sum x + b \sum x^2$

Example 1 By the method of least squares, find a straight line that best fits the following data points.

x	0	1	2	3	4
y	1.0	2.9	4.8	6.7	8.6

Solution: Let line of best fit be given by $y = ax + b \qquad \dots ①$

Normal equations are given by:

$$\sum y = a \sum x + nb \qquad \dots ②$$

and $\quad \sum xy = a \sum x^2 + b \sum x \qquad \dots ③$

Calculating $\sum x$, $\sum y$, $\sum xy$ and $\sum x^2$

x	y	xy	x^2
0	1.0	0	0
1	2.9	2.9	1
2	4.8	9.6	4
3	6.7	20.1	9
4	8.6	34.4	16
$\sum x = 10$	$\sum y = 24$	$\sum xy = 67.0$	$\sum x^2 = 30$

Substituting values of $\sum x, \sum y, \sum xy$ and $\sum x^2$ in ② and ③

$$\Rightarrow 24 = 10a + 5b \qquad \text{... ④}$$

and $$67 = 30a + 10b \qquad \text{... ⑤}$$

Solving ④and ⑤, we get $a = 1.9$ and $b = 1$

Substituting in ①, line of best fit is $y = 1.9x + 1$

Example 2 Fit a straight line to following data

x	0	1	2	3	4
y	1.0	1.8	3.3	4.5	6.3

Solution: Let line of best fit be given by $y = ax + b$ ... ①

Normal equations are given by:

$$\sum y = a\sum x + nb \qquad \text{... ②}$$

and $$\sum xy = a\sum x^2 + b\sum x \qquad \text{... ③}$$

Calculating $\sum x, \sum y, \sum xy$ and $\sum x^2$

x	y	xy	x^2
0	1.0	0	0
1	1.8	1.8	1
2	3.3	6.6	4
3	4.5	13.5	9
4	6.3	25.2	16
$\sum x = 10$	$\sum y = 16.9$	$\sum xy = 47.1$	$\sum x^2 = 30$

Substituting values of $\sum x, \sum y, \sum xy$ and $\sum x^2$ in ② and ③

$$\Rightarrow 16.9 = 10a + 5b \qquad \text{... ④}$$

and $\quad 47.1 = 30a + 10b \quad$ ...⑤

Solving ④and ⑤, we get $a = \frac{133}{100} = 1.33$ and $b = \frac{18}{25} = 0.72$

Substituting in ①, line of best fit is $y = 1.33x + 0.72$

Example 3 If F is the force required to lift a load W, by means of a pulley, fit a linear expression $F = a + bW$ against the following data:

W	50	70	100	120
F	12	15	21	25

Solution: Line for best fit is given as $F = a + bW$ ...①

Normal equations are given by:

$$\sum F = na + b\sum W \quad \text{...②}$$

and
$$\sum WF = a\sum W + b\sum W^2 \quad \text{...③}$$

W	F	WF	W^2
50	12	600	2500
70	15	1050	4900
100	21	2100	10000
120	25	3000	14400
$\sum W = 340$	$\sum F = 73$	$\sum WF = 6750$	$\sum W^2 = 31800$

Substituting values of $\sum W, \sum F, \sum WF$ and $\sum W^2$ in ② and ③

$$\Rightarrow 73 = 4a + 340b \quad \text{...④}$$

and
$$6750 = 340a + 31800b \quad \text{...⑤}$$

Solving ④and ⑤, we get $a = 2.2759$ and $b = 0.1879$

Substituting in ①, line of best fit is $F = 2.2759 + 0.1879W$

4.2.2 Fitting a parabola $y = ax^2 + bx + c$

Let $y = ax^2 + bx + c$ be the parabola to be fitted to the given set of data points (x_1, y_1), (x_2, y_2),……, (x_n, y_n).

Then $\quad e_i = y_i - f(x_i) = y_i - (ax_i^2 + bx_i + c), \quad i = 1, 2, \dots, n$

$$\therefore E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}\left(y_i - (ax_i^2 + bx_i + c)\right)^2$$

Now by the principle of least square, for the curve of best fit, E is minimum

$$\therefore \frac{\partial E}{\partial a} = 0, \quad \frac{\partial E}{\partial b} = 0 \text{ and } \frac{\partial E}{\partial c} = 0$$

Solving we get normal equations as:

$$\sum y \quad = a\sum x^2 + b\sum x + nc$$
$$\sum xy \ = a\sum x^3 + b\sum x^2 + c\sum x$$
$$\sum x^2 y = a\sum x^4 + b\sum x^3 + c\sum x^2$$

Example 4 Fit a parabola $y = ax^2 + bx + c$ to the given data

x	10	12	15	23	20
y	14	17	23	25	21

Solution: Let the parabola of best fit be given by $y = ax^2 + bx + c$ ... ①

Normal equations are given by:

$$\sum y \quad = a\sum x^2 + b\sum x + nc \quad ...②$$
$$\sum xy \ = a\sum x^3 + b\sum x^2 + c\sum x \quad ...③$$
$$\sum x^2 y = a\sum x^4 + b\sum x^3 + c\sum x^2 \quad ...④$$

$\boldsymbol{x}$	$\boldsymbol{y}$	$\boldsymbol{xy}$	$\boldsymbol{x^2}$	$\boldsymbol{x^2y}$	$\boldsymbol{x^3}$	$\boldsymbol{x^4}$
10	14	140	100	1400	1000	10000
12	17	204	144	2448	1728	20736
15	23	345	225	5175	3375	50625
23	25	575	529	13225	12167	279841
20	21	420	400	8400	8000	160000
$\sum \boldsymbol{x} =$ **80**	$\sum y =$ 100	$\sum xy =$ 1684	$\sum x^2 =$ 1398	$\sum x^2 y =$ 30684	$\sum x^3 =$ 26270	$\sum x^4 =$ 521202

Substituting values of $\sum x, \sum y, \sum xy$ and $\sum x^2$ in ② and ③ and ④

$$\Rightarrow 100 = 1398a + 80b + 5c \quad ...⑤$$
$$1684 = 26270a + 1398b + 80c \quad ...⑥$$
$$30648 = 521202a + 26270b + 1398c \quad ...⑦$$

Solving ⑤ ⑥ and ⑦ , we get $a = -0.07$, $b = 3.01$, $c = -8.73$

Substituting in ①, parabola of best fit is $y = -0.07\,x^2 + 3.01x - 8.73$

Example 5 Fit a *2nd* parabola to the given data

x	1	3	4	6	8	9	11	14
y	1	2	4	4	5	7	8	9

Solution: Let the parabola of best fit be given by $y = ax^2 + bx + c$ ... ①

x	y	xy	x^2	x^2y	x^3	x^4
1	1	1	1	1	1	1
3	2	6	9	18	27	81
4	4	16	16	64	64	256
6	4	24	36	144	216	1296
8	5	40	64	320	512	4096
9	7	63	81	567	729	6561
11	8	88	121	968	1331	14641
14	9	126	196	1764	2744	38416
$\sum x = 56$	$\sum y = 40$	$\sum xy = 364$	$\sum x^2 = 524$	$\sum x^2y = 3846$	$\sum x^3 = 5624$	$\sum x^4 = 65348$

Normal equations are given by:

$$\sum y \quad = a\sum x^2 + b\sum x + nc \qquad \ldots ②$$

$$\sum xy \quad = a\sum x^3 + b\sum x^2 + c\sum x \qquad \ldots ③$$

$$\sum x^2y = a\sum x^4 + b\sum x^3 + c\sum x^2 \qquad \ldots ④$$

Substituting values of $\sum x, \sum y, \sum xy$ and $\sum x^2$ in ② and ③ and ④

$$\Rightarrow 40 \;=\; 524a + 56b + 8c \qquad \ldots ⑤$$

$$364 \;=\; 5624a + 524b + 56c \qquad \ldots ⑥$$

$$3846 \;=\; 65348a \;+\; 5624b \;+\; 524c \qquad \ldots ⑦$$

Solving ⑤ ⑥ and ⑦ , we get

$$a = \frac{103}{11229} = 0.009\,,\; b = \frac{8672}{11229} = 0.77\,,\; c = \frac{4375}{22458} = \; 0.195$$

Substituting in ①, parabola of best fit is $y = 0.009\,x^2 + 0.77x + 0.195$

4.2.3 Change of Scale

If the data values are equispaced (with height (h)) and quite large for computation, simplification may be done by origin shifting as given below:

- When number of observations (n) is odd, take the origin at middle value of the table; say (x_0) and substitute $u = \frac{x-x_0}{h}$
- y values if small; may be left unchanged; or we can shift them at average value of y data $v = \frac{y-y_0}{h}$

- When number of observations (n) is even, take the origin as mean of two middle values, with new height $\frac{h}{2}$ and substitute $u = \frac{x-x_0}{h/2}$

Example6 Fit a 2*nd* degree parabola for the following data:

x	1929	1930	1931	1932	1933	1934	1935	1936	1937
y	352	356	357	358	360	361	361	360	359

Solution: Since number of observations is odd and $h = 1$,

taking $x_0 = 1933, y_0 = 357, u = x - 1933, v = y - 357$

The equation $y = ax^2 + bx + c$ is transformed to $v = Au^2 + Bu + C$ ...①

Normal equations are

$$\sum v = A\sum u^2 + B\sum u + 9c \quad \text{...②}$$

$$\sum uv = A\sum u^3 + B\sum u^2 + c\sum u \quad \text{...③}$$

$$\sum u^2 v = A\sum u^4 + B\sum u^3 + c\sum u^2 \quad \text{...④}$$

Calculating $\sum u, \sum u^2, \sum u^3, \sum u^4, \sum v, \sum uv$ and $\sum u^2 v$

x	u	y	v	uv	u^2	u^2v	u^3	u^4
1929	-4	352	-5	20	16	-80	-64	256
1930	-3	356	-1	3	9	-9	-27	81
1931	-2	357	0	0	4	0	-8	16
1932	-1	358	1	-1	1	1	-1	1
1933	0	360	3	0	0	0	0	0
1934	1	361	4	4	1	4	1	1
1935	2	361	4	8	4	16	8	16
1936	3	360	3	9	9	27	27	81
1937	4	359	2	8	16	32	64	256
	$\sum u = 0$		$\sum v = 11$	$\sum uv = 51$	$\sum u^2 = 60$	$\sum u^2 v = -9$	$\sum u^3 = 0$	$\sum u^4 = 708$

Substituting $\sum u, \sum u^2, \sum u^3, \sum u^4, \sum v, \sum uv$ and $\sum u^2 v$ in ② and ③ and ④

$$\Rightarrow 11 = 60A + 9C \quad \text{...⑤}$$

$$51 = 60B \quad \text{...⑥}$$

$$-9 = 708A + 60C \quad \text{...⑦}$$

Solving ⑤ ⑥ and ⑦ , we get $A = \frac{-247}{924}$ and $B = \frac{17}{20}$, $C = \frac{694}{231}$

Substituting in ①, parabola of best fit is $v = \frac{-247}{924}x^2 + \frac{17}{20}x + \frac{694}{231}$

$$\Rightarrow y - 357 = \frac{-247}{924}(x-1933)^2 + \frac{17}{20}(x-1933) + \frac{694}{231}$$

$$\Rightarrow y = -0.267x^2 + 1034.29x - 1000106.41$$

Example 7 The weight of a calf taken at end of every month is given below. Fit a straight line using the method of least squares. Also compute monthly growth rate.

x	1	2	3	4	5	6	7	8	9	10
y	52.5	58.7	65.0	70.2	75.4	81.1	87.2	95.5	102.2	108.4

Solution: Here $n = 10$ is even, $\therefore$ taking origin at $\frac{5+6}{2} = 5.5$ and new height as

$\frac{h}{2} = 0.5 \;\therefore u = \frac{x-5.5}{0.5}$ and let $v = y$

Let line of best fit $y = ax + b$ be transformed to $v = Au + B$ ...①

Normal equations are given by $\sum v = a\sum u + nb$ ...②

and $\sum uv = a\sum u^2 + b\sum u$ ...③

x	$u = \frac{x-5.5}{0.5}$	$v = y$	uv	u^2
1	-9	52.5	-472.5	81
2	-7	58.7	-410.9	49
3	-5	65.0	-325.0	25
4	-3	70.2	-210.6	9
5	-1	75.4	-75.4	1
6	1	81.1	81.1	1
7	3	87.2	261.6	9
8	5	95.5	477.5	25
9	7	102.2	715.4	49
10	9	108.4	975.6	81
	$\sum u = 0$	$\sum v = 796.2$	$\sum uv = 1016.8$	$\sum u^2 = 330$

Substituting values of $\sum u, \sum v, \sum uv$ and $\sum u^2$ in ② and ③

$\Rightarrow 796.2 = 10B$ and $1016.8 = 330a$

$\therefore A = 3.081$ and $B = 79.62$

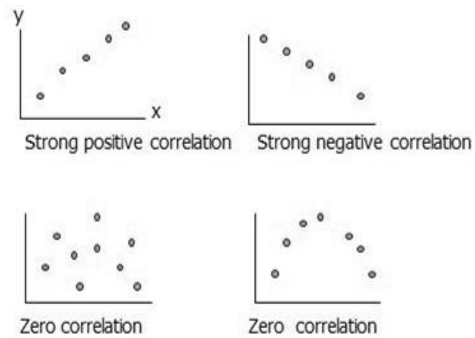
Substituting in ①, line of best fit is $v = 3.081u + 79.62$

$$\Rightarrow y = 3.081\left(\frac{x-5.5}{0.5}\right) + 79.62 \;\; \therefore \text{ Line of best fit is } y = 6.162x + 45.729$$

Average growth rate per month is given by: $\frac{dy}{dx} = 6.162$

4.3 Correlation

Correlation is a measure of association between two variables; which may be dependent or independent. Whenever two variables x and y are so related; that increase in one in accompanied by an increase or decrease in the other, then the variables are said to be correlated. Coefficient of correlation (r) lies between -1 and $+1$, i.e.$-1 \leq r \leq 1$.



If r is zero; no correlation between two variables, positive correlation ($0 < r \leq +1$); when both variables increase or decrease simultaneously, and negative correlation ($-1 \leq r < 0$); when increase in one is associated with decrease in other variable and vice-versa.

4.4 Karl Pearson Coefficient of Correlation

Coefficient of correlation (r) between two variables x and y is defined as

$$r = \frac{Covariance\ (x,y)}{\sqrt{Variance\ (x)}\ \sqrt{Variance\ (y)}} = \frac{\sum d_x d_y}{\sqrt{(\sum d_x^2)(\sum d_y^2)}} = \frac{\rho}{\sigma_x \sigma_y}$$

where $d_x = x - \bar{x}$, $d_y = y - \bar{y}$, $\bar{x}$, $\bar{y}$ are means of x and y data values.

$\rho = Cov\ (x, y) = \frac{\sum d_x d_y}{n}$ is the covariance between the variables x and y.

Also $\sigma_x = \sqrt{\frac{\sum d_x^2}{n}}$ and $\sigma_y = \sqrt{\frac{\sum d_y^2}{n}}$

Example 8 If $Cov\ (x, y) = 10$, $var(x) = 25$, $var(y) = 9$ find coefficient of correlation.

Solution: $r = \frac{Cov\ (x,y)}{\sqrt{Var(x)}\ \sqrt{Var(y)}} = \frac{10}{\sqrt{25}\ \sqrt{9}} = \frac{10}{5\times 3} = 0.67$

Example 9 Calculate coefficient of correlation from the following data:

x	9	8	7	6	5	4	3	2	1
y	15	16	14	13	11	12	10	8	9

Solution: Karl Pearson coefficient of correlation (r) is given by: $r = \frac{\sum d_x d_y}{\sqrt{(\sum d_x^2)(\sum d_y^2)}}$

where $d_x = x - \bar{x}$, $d_y = y - \bar{y}$, $\bar{x}$, $\bar{y}$ are means of x and y data values.

x	d_x $(x-\overline{x})$	d_x^2	y	d_y $(y-\overline{y})$	d_y^2	$d_x d_y$
9	4	16	15	3	9	12
8	3	9	16	4	16	12
7	2	4	14	2	4	4
6	1	1	13	1	1	1
5	0	0	11	-1	1	0
4	-1	1	12	0	0	0
3	-2	2	10	-2	4	4
2	-3	9	8	-4	16	12
1	-4	16	9	-3	9	12
$\sum x = 45$ $\overline{x} = 5$		$\sum d_x^2 = \mathbf{60}$	$\sum y = 108$ $\overline{y} = 12$		$\sum d_y^2 = 60$	$\sum d_x d_y = 57$

$$\therefore r = \frac{\sum d_x d_y}{\sqrt{(\sum d_x^2)(\sum d_y^2)}} = \frac{57}{\sqrt{60 \times 60}} = \frac{57}{60} = 0.95$$

4.4.1 Shortcut Method for Karl Pearson Coefficient of Correlation

We can also find Karl Pearson Coefficient of Correlation by taking assumed means as shown: If we take $d_x = x - a$, $d_y = y - b$,

where a and b are assumed means of x and y data values

Then $$r = \frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{n}}{\sqrt{(\sum d_x^2) - \frac{1}{n}(\sum d_x)^2}\sqrt{(\sum d_y^2) - \frac{1}{n}(\sum d_y)^2}}$$

- If x_i's are equispaced with height h , we can take $d_x = \frac{x-a}{h}$
  Similarly if y_i's are equispaced with height k , then $d_y = \frac{y-b}{k}$

Example 10 Calculate coefficient of correlation from the following data:

x	1	3	5	7	8	10
y	8	12	15	17	18	20

Solution: Let $d_x = x - 7$, $d_y = y - 15$

$$r = \frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{n}}{\sqrt{(\sum d_x^2) - \frac{1}{n}(\sum d_x)^2}\sqrt{(\sum d_y^2) - \frac{1}{n}(\sum d_y)^2}}$$

Calculating $\sum d_x, \sum d_y , \sum d_x^2 , \sum d_y^2$ and $\sum d_x d_y$

x	d_x $(x-7)$	d_x^2	y	d_y $(y-15)$	d_y^2	$d_x d_y$
1	-6	36	8	-7	49	42
3	-4	16	12	-3	9	12
5	-2	4	15	0	0	0
7	0	0	17	2	4	0
8	1	1	18	3	9	3
10	3	9	20	5	25	15
	$\sum d_x = -8$	$\sum d_x^2 = 66$		$\sum d_y = 0$	$\sum d_y^2 = 96$	$\sum d_x d_y = 72$

$$\therefore r = \frac{72 - \frac{(-8)(0)}{6}}{\sqrt{(66)-\frac{1}{6}(-8)^2}\sqrt{(96)-\frac{1}{6}(0)^2}} = 0.9879$$

4.4.2 Coefficient of Correlation of Bivariate Frequency Distribution

If given data is in the form of a bivariate frequency distribution,

$$\text{Then } r = \frac{\sum f d_x d_y - \frac{\sum f d_x \sum f d_y}{n}}{\sqrt{(\sum f d_x^2) - \frac{1}{n}(\sum f d_x)^2}\sqrt{(\sum f d_y^2) - \frac{1}{n}(\sum f d_y)^2}} \quad , n = \sum f$$

Example 11 Following table gives a bivariate distribution showing frequency of marks obtained according to age by a group of 52 students in an intelligent test:

Age / *Marks*	16-18	18-20	20-22	22-24
10-20	2	1	1	
20-30	3	2	3	2
30-40	3	4	5	6
40-50	2	2	3	4
50-60		1	2	2
60-70		1	2	1

Compute the correlation between marks and age of the students.

Solution: Let marks obtained by the students be denoted by *x* and age by *y*, then coefficient of correlation (*r*) for the bivariate frequency distribution is given by:

$$r = \frac{\sum f d_x d_y - \frac{\sum f d_x \sum f d_y}{n}}{\sqrt{(\sum f d_x^2) - \frac{1}{n}(\sum f d_x)^2}\sqrt{(\sum f d_y^2) - \frac{1}{n}(\sum f d_y)^2}} \quad , n = \sum f$$

Let $d_x = \frac{x-a}{10}$, $d_y = \frac{y-b}{2}$, where a& b denote assumed mean classes

Here a is taken as $30-40$, b is taken as $18-20$

Also quantities in brackets denote $d_x d_y$ for each cell.

$\therefore f d_x d_y$ for each cell is obtained by multiplying frequency of each cell with $d_x d_y$ and added across rows or columns to get $f d_x d_y$

x \ y	16-18	18-20	20-22	22-24	F	d_x	fd_x	fd_x^2	fd_xd_y
10-20	2 (2)	1 (0)	1 (-2)	-	4	-2	-8	16	2
20-30	3 (1)	2 (0)	3 (-1)	2 (-2)	10	-1	-10	10	-4
30-40	3 (0)	4 (0)	5 (0)	6 (0)	18	0	0	0	0
40-50	2 (-1)	2 (0)	3 (1)	4 (2)	11	1	11	11	9
50-60	-	1 (0)	2 (2)	2 (4)	5	2	10	20	12
60-70	-	1 (0)	2 (3)	1 (6)	4	3	12	36	12
F	10	11	16	15	**52**	Totals	15	93	31
d_y	-1	0	1	2	Totals				
fd_y	-10	0	16	30	36				
fd_y^2	10	0	16	60	86				
fd_xd_y	5	0	8	18	31				

$$\therefore r = \frac{31 - \frac{(15)(36)}{52}}{\sqrt{(93)-\frac{1}{52}(15)^2}\sqrt{(86)-\frac{1}{52}(36)^2}} = \frac{20.6154}{(9.4166)(7.8152)} = 0.2801$$

Thus there is a weak positive correlation between marks and age of the students.

4.4.3 Coefficient of Correlation by Rank differences

Rank correlation is used for attributes (like beauty, intelligence etc.) which cannot be measured quantitatively but can be provided with comparative ranks.

Spearman's Rank Correlation in given by: $r = 1 - \frac{6\sum D^2}{n(n^2-1)}$, where $D = R_1 - R_2$

<u>Tied Ranks</u>: If two or more observations in a data are equal, each observation is provided with an average rank and a correction factor is applied to correlation formula given as: Correction Factor (C.F.) = $\sum m(m^2-1)$, m is the number of times each observation is repeated.

Spearman's Rank Correlation for repeated ranks is given by:

$$r = 1 - \frac{6\left(\sum D^2 + \frac{1}{12}\text{C.F.}\right)}{n(n^2-1)}, \text{ where } D = R_1 - R_2$$

Example12 Calculate the coefficient of correlation from the following data; given ranks of 10 students in English and Mathematics.

Rank in English	3	1	5	4	2	6	8	10	9	7
Rank in Mathematics	2	4	3	1	5	10	7	9	8	6

Solution: Since comparative ranks are given; instead of marks, using Spearman's Rank Correlation is given by: $r = 1 - \frac{6\sum D^2}{n(n^2-1)}$, where $D = R_1 - R_2$

Rank in English $\boldsymbol{R_1}$	Rank in Mathematics $\boldsymbol{R_2}$	$\boldsymbol{D = R_1 - R_2}$	$\boldsymbol{D^2}$
3	2	1	1
1	4	-3	9
5	3	2	4
4	1	3	9
2	5	-3	9
6	10	-4	16
8	7	1	1
10	9	1	1
9	8	1	1
7	6	1	1
			$\sum D^2 = 52$

$$\therefore r = 1 - \frac{6(52)}{10(10^2-1)} = 0.6848$$

Example13 Eight competitors in a beauty contest got marks (out of 10) by three judges as given:

Judge A	9	6	5	10	3	1	4	2
Judge B	3	5	8	4	7	10	2	1
Judge C	6	4	9	8	1	2	3	10

Use rank correlation to discuss which pair of judges has the nearest approach to common tastes in beauty.

Solution: Since instead of ranks; marks are given by the three judges, converting the given data to comparative ranks for the eight competitors

Judge A		Judge B		Judge C		D_{AB}	D_{AB}^2	D_{BC}	D_{BC}^2	D_{AC}	D_{AC}^2
Marks	Rank	Marks	Rank	Marks	Rank						
9	2	3	6	6	4	-4	16	2	4	-2	4
6	3	5	4	4	5	-1	1	-1	1	-2	4
5	4	8	2	9	2	2	4	0	0	2	4
10	1	4	5	8	3	-4	16	2	4	-2	4
3	6	7	3	1	8	3	9	-5	25	-2	4
1	8	10	1	2	7	7	49	-6	36	1	1
4	5	2	7	3	6	-2	4	1	1	-1	1
2	7	1	8	10	1	-1	1	7	49	6	36

Here $D_{AB} =$ Rank by Judge A $-$ Rank by Judge B, also $\sum D_{AB}^2 = 100$

Similarly $D_{BC} =$ Rank by Judge B $-$ Rank by Judge C, also $\sum D_{BC}^2 = 120$

$D_{AC} =$ Rank by Judge A $-$ Rank by Judge C, also $\sum D_{AC}^2 = 58$

Rank Correlation between judges A and B is given by:

$$r_{AB} = 1 - \frac{6\sum D_{AB}^2}{n(n^2-1)} = 1 - \frac{6(100)}{8(8^2-1)} = -0.1905$$

Rank Correlation between judges B and C is given by:

$$r_{BC} = 1 - \frac{6\sum D_{BC}^2}{n(n^2-1)} = 1 - \frac{6(120)}{8(8^2-1)} = -0.4286$$

Rank Correlation between judges A and C is given by:

$$r_{AC} = 1 - \frac{6\sum D_{AC}^2}{n(n^2-1)} = 1 - \frac{6(58)}{8(8^2-1)} = 0.3095$$

Therefore Judges A and C have the nearest approach to common tastes in beauty, while Judges B and C have most different beauty tastes.

Example14: Obtain rank correlation coefficient for following marks in economics (x) and Mathematics (y) out of 25 for eight students.

x	20	24	12	20	10	12	24	20
y	18	19	16	22	14	16	19	12

Solution: Converting data into ranks: Ranks of x as R_x , Ranks of y as R_y

X	R_x	Y	R_y	$D = R_x - R_y$	D^2
20	4	18	4	0	0
24	1.5	19	2.5	-1	1
12	6.5	16	5.5	1	1
20	4	22	1	3	9
10	8	14	7	1	1
12	6.5	16	5.5	1	1
24	1.5	19	2.5	-1	1
20	4	12	8	-4	16
					$\sum D^2 = 30$

Correction Factor $= \sum m(m^2-1)$, m is the number of times each data value is repeated $\therefore$ $\text{C.F.} = 2(2^2-1) + 3(3^2-1) + 2(2^2-1) + 2(2^2-1) + 2(2^2-1)$

$$= 6 + 24 + 6 + 6 + 6 = 48$$

Spearman's Rank Correlation for repeated ranks is given by:

$$r = 1 - \frac{6\left(\sum D^2 + \frac{1}{12}\text{C.F.}\right)}{n(n^2-1)}, \text{ where } D = R_x - R_y$$

$$\therefore r = 1 - \frac{6\left(30 + \frac{48}{12}\right)}{8(8^2-1)} = \frac{25}{42} = 0.595$$

Example 15 Obtain rank correlation coefficient for following data

x	68	64	75	50	64	80	75	40	55	64
y	62	58	68	45	81	60	68	48	50	70

Solution: Converting data into ranks: Ranks of x as R_x , Ranks of y as R_y

x	R_x	Y	R_y	$D = R_x - R_y$	D^2
68	4	62	5	-1	1
64	6	58	7	-1	1
75	2.5	68	3.5	-1	1
50	9	45	10	-1	1
64	6	81	1	5	25
80	1	60	6	-5	25
75	2.5	68	3.5	-1	1
40	10	48	9	1	1
55	8	50	8	0	0
64	6	70	2	4	16
					$\sum D^2 = 72$

Correction Factor (C.F.) $=\sum m(m^2-1)$, m is the number of times each data value is repeated $\therefore \text{C.F.} = 2(2^2-1)+3(3^2-1)+2(2^2-1)=36$

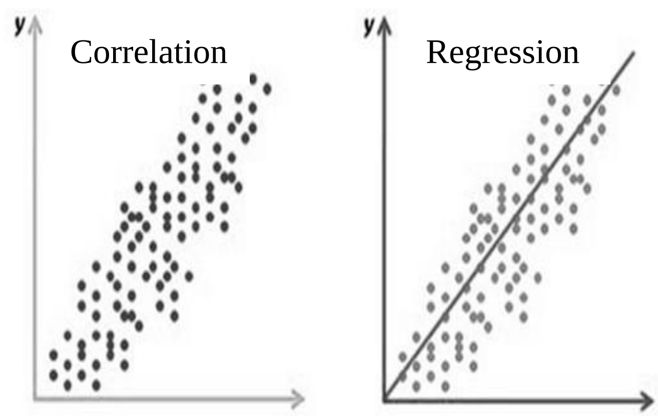
Spearman's Rank Correlation for repeated ranks is given by:

$$r = 1 - \frac{6\left(\sum D^2 + \frac{1}{12}\text{C.F.}\right)}{n(n^2-1)} \text{, where } D = R_x - R_y$$

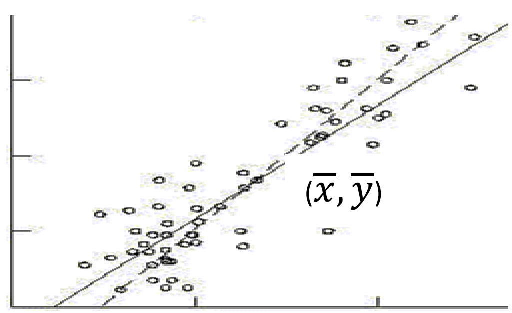
$$\therefore r = 1 - \frac{6\left(72+\frac{36}{12}\right)}{10(10^2-1)} = \frac{6}{11} = 0.545$$

4.5 Linear Regression

Regression describes the functional relationship between dependent and independent variables; which helps us to make estimates of one variable from the other. Correlation quantifies the association between the two variables; whereas linear regression finds the best line that predicts y from x and also x from y .The difference between correlation and regression is illustrated in the adjoining figure.



4.5.1 Lines of Regression: If we plot the observations of the linear regression between two variables, actually two straight lines can approximately be drawn through the scatter diagram. One line estimates values of y for specified values of x (known as line of regression of y on x); and other predicts values of x from given values of y (called line of regression of x on y).



Let line of regression of y on x be represented by $y = a + bx$ ... ①

Normal equations as derived by the method of least Square are:

$$\sum y = an + b\sum x \qquad \text{... ②}$$

and $\sum xy = a\sum x + b\sum x^2$ ... ③

Dividing ② by n , we get

$$\frac{\sum y}{n} = a + b\frac{\sum x}{n} \Rightarrow \overline{y} = a + b\overline{x}$$

Where $\overline{x}$ and $\overline{y}$ are the means of x series and y series. This shows that $(\overline{x}, \overline{y})$ lies on the line of regression given by ①.

Again as $(\overline{x}, \overline{y})$ satisfies ① , shifting the origin to $(\overline{x}, \overline{y})$ in equation ③ , we get

$$\sum(x-\overline{x})(y-\overline{y}) = a\sum(x-\overline{x}) + b\sum(x-\overline{x})^2$$

$$\Rightarrow \sum(x-\overline{x})(y-\overline{y}) = b\sum(x-\overline{x})^2 \qquad \because \sum(x-\overline{x}) = 0$$

$$\Rightarrow b = \frac{\sum(x-\overline{x})(y-\overline{y})}{\sum(x-\overline{x})^2} = \frac{\sum d_x d_y}{\sum d_x^2} \qquad \ldots ④$$

$$\text{Again } r = \frac{\sum d_x d_y}{\sqrt{(\sum d_x^2)(\sum d_y^2)}} = \frac{\sum d_x d_y}{n\sqrt{\frac{\sum d_x^2}{n}}\sqrt{\frac{\sum d_y^2}{n}}} = \frac{\sum d_x d_y}{n\sigma_x\sigma_y} \quad \because \sigma_x = \sqrt{\frac{\sum d_x^2}{n}},\ \sigma_y = \sqrt{\frac{\sum d_y^2}{n}}$$

Here σ_x , σ_y are standard deviations of x and y data points respectively

$$\Rightarrow \sum d_x d_y = nr\sigma_x\sigma_y \qquad \ldots ⑤$$

Using ⑤ in ④ , we get

$$b = \frac{n\, r\, \sigma_x\sigma_y}{\sum d_x^2} = \frac{r\, \sigma_x\sigma_y}{\sigma_x^2}$$

$\Rightarrow b = \frac{r\,\sigma_y}{\sigma_x}$ which is slope of line of regression line of y on x

$\therefore\ b_{yx} = \frac{r\,\sigma_y}{\sigma_x}$, b_{yx} denotes slope of line of regression line of y on x.

Thus line of regression of y on x given by ①, passes through $(\overline{x}, \overline{y})$ and is having slope $b_{yx} = \frac{r\,\sigma_y}{\sigma_x}$

$\therefore$ Equation of line of regression of y on x is given by $y - \overline{y} = b_{yx}(x - \overline{x})$

Similarly line of regression of x on y is given by: $x - \overline{x} = b_{xy}(y - \overline{y})$

where $b_{xy} = \frac{r\,\sigma_x}{\sigma_y}$ is slope of line of regression line of x on y

Here b_{xy} and b_{yx} are known coefficients of regression and are connected by the relation:

$$b_{xy}b_{yx} = \left(\frac{r\,\sigma_x}{\sigma_y}\right)\left(\frac{r\,\sigma_y}{\sigma_x}\right) = r^2$$

4.5.2 Properties of Regression Coefficients

- As $\sqrt{b_{xy}b_{yx}} = r$, the coefficient of correlation is the geometric mean between the two regression coefficients.
- Since $\frac{b_{xy}+b_{yx}}{2} \geq \sqrt{b_{xy}b_{yx}} = r$, $\therefore$ arithmetic mean of the two regression coefficients is greater than or equal to the correlation coefficient (r).

- If there is a perfect correlation between the two variables under consideration, then $b_{xy} = b_{yx} = r$; and the two lines of regression coincide. Converse is also true, i.e. if two lines of regression coincide, then there is a perfect correlation; $r = \pm 1$.
- Since $b_{xy} b_{yx} = r^2 > 0$, the signs of both regression coefficients b_{xy} and b_{yx} and coefficient of correlation (r) must be same; either all three negative or all positive.
- $\because b_{xy} b_{yx} = r^2 \leq 1$, if one of the regression coefficients is greater than unity, other must be less than unity.
- Point of intersection of two lines of regression is $(\overline{x}, \overline{y})$, Where $\overline{x}$ and $\overline{y}$ are the means of x series and y series.
- If both lines of regression cut each other at right angle, there is no correlation between the two variables; i.e. $r = 0$.

Example 16 Prove that arithmetic mean of coefficients of regression is greater than the coefficient of correlation.

Solution: We know that $b_{xy} = \frac{r\,\sigma_x}{\sigma_y}$ and $b_{yx} = \frac{r\,\sigma_y}{\sigma_x}$

To prove $\frac{b_{xy}+b_{yx}}{2} > r$

or $\frac{1}{2}\left[\frac{r\,\sigma_x}{\sigma_y} + \frac{r\,\sigma_y}{\sigma_x}\right] > r$

or $\frac{1}{2}\left[\frac{\sigma_x^2 + \sigma_y^2}{\sigma_x \sigma_y}\right] > 1$

or $\left[\frac{\sigma_x^2 + \sigma_y^2}{\sigma_x \sigma_y}\right] - 2 > 0$

or $\sigma_x^2 + \sigma_y^2 - 2\sigma_x \sigma_y > 0$

or $[\sigma_x - \sigma_y]^2 > 0$

which is true

Note : A.M. $= r$ if $b_{xy} = b_{yx} = r = \pm 1$

4.5.3 Angle between the Lines of Regression

If θ be the acute angle between the two regression lines for two variables x and y, then $\tan\theta = \frac{1-r^2}{r} \frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}$

Proof: The two lines of regression are given by:

$$y - \overline{y} = \frac{r\,\sigma_y}{\sigma_x}(x - \overline{x}) \qquad \text{...①}$$

and $$x - \overline{x} = \frac{r\,\sigma_x}{\sigma_y}(y - \overline{y}) \qquad \text{...②}$$

If m_1 and m_2 are slopes of lines ① and ②, then

$\tan\theta = \frac{m_2 - m_1}{1+m_1 m_2}$, where $m_1 = \frac{r\,\sigma_y}{\sigma_x}$, $m_2 = \frac{\sigma_y}{r\,\sigma_x}$

$$\Rightarrow \tan\theta = \frac{\frac{\sigma_y}{r\,\sigma_x} - \frac{r\,\sigma_y}{\sigma_x}}{1+\frac{r\sigma_y}{\sigma_x}\,\frac{\sigma_y}{r\sigma_x}} = \frac{\left(\frac{1}{r} - r\right)\frac{\sigma_y}{\sigma_x}}{1+\frac{\sigma_y^2}{\sigma_x^2}} = \frac{1-r^2}{r}\,\frac{\sigma_x\,\sigma_y}{\sigma_x^2 + \sigma_y^2} \qquad \text{...③}$$

- When $r = 0$, $\tan\theta = \infty \;\Rightarrow \theta = \frac{\pi}{2}$ from ③

  ∴ when $r = 0$, the two lines of regression are perpendicular to each other.
- When $r = \pm 1$, $\tan\theta = 0 \;\Rightarrow \theta = 0$ from ③

  ∴ when $r = \pm 1$, the two lines of regression are coincident

Example17 Find the correlation coefficient between *x* and *y*, when the two lines of regression are given by: $2x - 9y + 6 = 0$ and $x - 2y + 1 = 0$

Solution: Let the line of regression of *x* on *y* be $2x - 9y + 6 = 0$ ...①

Then the line of regression of *y* on *x* is $x - 2y + 1 = 0$ ...②

Now ① $\Rightarrow x = \frac{9}{2}y - 3 \quad \therefore b_{xy} = \frac{9}{2}$

Also ② $\Rightarrow y = \frac{1}{2}x + \frac{1}{2} \quad \therefore b_{yx} = \frac{1}{2}$

$\therefore r = \sqrt{b_{xy} b_{yx}} = \sqrt{\frac{9}{2} \times \frac{1}{2}} = \frac{3}{2}$, which is not possible as $-1 \le r \le 1$

So our choice of regression lines is incorrect.

∴Line of regression of *x* on *y* is $x - 2y + 1 = 0$

$\Rightarrow x = 2y - 1 \quad \therefore b_{xy} = 2$

Also line of regression of *y* on *x* is $2x - 9y + 6 = 0$

$\Rightarrow y = \frac{2}{9}x + \frac{2}{3} \quad \therefore b_{yx} = \frac{2}{9}$

$\therefore r = \sqrt{b_{xy} b_{yx}} = \sqrt{2 \times \frac{2}{9}} = \frac{2}{3}$

Hence coefficient of correlation between *x* and *y* is $\frac{2}{3}$

Example18 The regression equations calculated from a given set of observations for two random variables are: $x = -0.4y + 6.4$ and $y = -0.6x + 4.6$

Calculate $\overline{x}, \overline{y}$ and r.

Solution: The two equations of regression are:

$x = -0.4y + 6.4$ ...①

$$y = -0.6x + 4.6 \qquad \ldots ②$$

$\Rightarrow b_{xy} = -0.4$ and $b_{yx} = -0.6$

$\therefore r^2 = b_{xy}b_{yx} = 0.24$

$\Rightarrow \quad r = \pm 0.49$

We know that the signs of b_{xy} , b_{yx} and r must be same

$\therefore r = -0.49$

Again we know that the point of intersection of two regression lines is $(\overline{x}, \overline{y})$

Therefore solving ① and ②, we get $\overline{x} = 6$, $\overline{y} = 1$

Example19 From a partially destroyed lab data, following results were retrieved:

Lines of regression are:

$x = 0.45y + 5.35$ and $y = 0.8x + 6.6$, $\sigma_x^2 = 9$

Find $\overline{x}, \overline{y}, \sigma_y$ and r for the existing data.

Solution: The two equations of regression are:

$$x = 0.45y + 5.35 \qquad \ldots ①$$

$$y = 0.8x + 6.6 \qquad \ldots ②$$

We know that the point of intersection of two regression lines is $(\overline{x}, \overline{y})$

Therefore solving ① and ②, we get $\overline{x} = 13$, $\overline{y} = 17$

Again ① $\Rightarrow b_{xy} = 0.45$ and $b_{yx} = 0.8$

$\therefore r^2 = b_{xy}b_{yx} = 0.36$

$\Rightarrow \quad r = \pm 0.6$

We know that the signs of b_{xy} , b_{yx} and r must be same

$\therefore r = 0.6$

Also $b_{yx} = \frac{r\,\sigma_y}{\sigma_x} \Rightarrow 0.8 = \frac{(0.6)\sigma_y}{3} \Rightarrow \sigma_y = \frac{0.8 \times 3}{0.6} = 4$

Example 20 Find the regression line of *y* on *x* from the following data:

x	1	3	4	6	8	9	11	14
y	1	2	4	4	5	7	8	9

Also estimate the value of *y*, when $x = 10$

Solution: Let line of regression of *y* on *x* be:

$$y = a + bx \qquad \ldots ①$$

Then normal equations are given by:

$$\sum y = an + b\sum x \qquad \ldots ②$$

and $\sum xy = a\sum x + b\sum x^2 \qquad \ldots ③$

Calculating $\sum x, \sum y, \sum xy$ and $\sum x^2$

$\boldsymbol{x}$	$\boldsymbol{y}$	$\boldsymbol{x^2}$	$\boldsymbol{xy}$
1	1	1	1
3	2	9	6
4	4	16	16
6	4	36	24
8	5	64	40
9	7	81	63
11	8	121	88
14	9	196	126
$\sum \boldsymbol{x} = 56$	$\sum y = 40$	$\sum x^2 = 524$	$\sum xy = 364$

Substituting values of $\sum x, \sum y, \sum xy$ and $\sum x^2$ in ② and ③

$$\Rightarrow 40 = 8a + 56b \qquad \ldots ④$$

and $364 = 56a + 524b \qquad \ldots ⑤$

Solving ④and ⑤, we get $a = \frac{6}{11}$ and $b = \frac{7}{11}$

Substituting in ①, line of regression of y on x is $y = \frac{6}{11} + \frac{7}{11}x$

$\Rightarrow 7x - 11y + 6 = 0$

Also at $x = 10$, $y = \frac{76}{11}$

Example 21 Following data depicts the statistical values of rainfall and production of wheat in a region for a specified time period.

	Mean	Standard Deviation
Production of Wheat (kg. per unit area)	10	8
Rainfall (cm)	8	2

Estimate the production of wheat when rainfall is 9cm if correlation coefficient between production and rainfall is given to be 0.5.

Solution: Let the variables x and y denote production and rainfall respectively.

Given that $\overline{x} = 10$, $\overline{y} = 8$ also $\sigma_x = 8$, $\sigma_y = 2$

Now equation of regression of x on y is given by:

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

$$\Rightarrow x - 10 = \frac{(0.5)8}{2}(y - 8)$$

$$\Rightarrow x = 2y - 6$$

$\therefore$ When rainfall is 9cm, production of wheat is estimated to be

$2(9) - 6 = 12$ kg. per unit area

Example 22 Find the coefficient of correlation and the lines of regression for the data given below:

$$n = 18, \sum x = 12, \sum y = 18 , \sum x^2 = 60 , \sum y^2 = 96 \text{ and } \sum xy = 48$$

Solution: $\overline{x} = \frac{\sum x}{n} = \frac{12}{18} = 0.67$, $\overline{y} = \frac{\sum y}{n} = \frac{18}{18} = 1$

$$\sigma_x^2 = \frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2 = \frac{60}{18} - \left(\frac{12}{18}\right)^2 = 2.89 \therefore \sigma_x = 1.7$$

$$\sigma_y^2 = \frac{\sum y^2}{n} - \left(\frac{\sum y}{n}\right)^2 = \frac{96}{18} - \left(\frac{18}{18}\right)^2 = 4.33 \therefore \sigma_y = 2.08$$

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{(\sum x^2) - \frac{1}{n}(\sum x)^2}\sqrt{(\sum y^2) - \frac{1}{n}(\sum y)^2}}$$

$$= \frac{48 - \frac{(12)(18)}{18}}{\sqrt{(60) - \frac{1}{18}(12)^2}\sqrt{(96) - \frac{1}{18}(18)^2}} = \frac{36}{(7.2)(8.83)} = 0.57$$

$$b_{xy} = \frac{r\,\sigma_x}{\sigma_y} = \frac{(0.57)(1.7)}{2.08} = 0.47 \quad , b_{yx} = \frac{r\,\sigma_y}{\sigma_x} = \frac{(0.57)(2.08)}{1.7} = 0.7$$

Equations of lines of regression are:

$$y - \overline{y} = b_{yx}(x - \overline{x}) \quad , x - \overline{x} = b_{xy}(y - \overline{y})$$

$$\Rightarrow y - 1 = 0.7(x - 0.67) \text{ and } x - 0.67 = 0.47(y - 1)$$

$$\Rightarrow y = 0.7x + 0.53 \qquad \text{and } x = 0.47y + 0.2$$

Example 23 Marks obtained by 11 students in statistics papers are given below:

Paper I	60	65	68	70	75	85	80	45	55	56	58
Paper II	62	64	65	70	74	90	82	56	50	48	60

Calculate the coefficient of correlation for the above data. Also find the equations of lines of regression.

Solution: Let marks obtained in paper I be denoted by x and marks obtained in paper II be denoted by y.

Let $A_x = 65,\ A_y = 70\ \therefore\ d_x = x - 65\ ,\ d_y = y - 70$

Calculating $\sum d_x, \sum d_y\ , \sum d_x^2\ ,\ \sum d_y^2$ and $\sum d_x d_y$

$\boldsymbol{x}$	$\boldsymbol{d_x}$ $(\boldsymbol{x-65})$	$\boldsymbol{d_x^2}$	$\boldsymbol{y}$	$\boldsymbol{d_y}$ $(\boldsymbol{y-70})$	$\boldsymbol{d_y^2}$	$\boldsymbol{d_x d_y}$
60	-5	25	62	-8	64	40
65	0	0	64	-6	36	0
68	3	9	65	-5	25	-15
70	5	25	70	0	0	0
75	10	100	74	4	16	40
85	20	400	90	20	400	400
80	15	225	82	12	144	180
45	-20	400	56	-14	196	280
55	-10	100	50	-20	400	200
56	-9	81	48	-22	484	198
58	-7	49	60	-10	100	70
	$\sum d_x = 2$	$\sum d_x^2$ =1414		$\sum d_y = -49$	$\sum d_y^2$ =1865	$\sum d_x d_y = 1393$

Karl Pearson coefficient of correlation (r) is given by:

$$r = \frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{n}}{\sqrt{(\sum d_x^2) - \frac{1}{n}(\sum d_x)^2}\sqrt{(\sum d_y^2) - \frac{1}{n}(\sum d_y)^2}}$$

$$\therefore r \ = \frac{1393 - \frac{(2)(-49)}{11}}{\sqrt{(1414) - \frac{1}{11}(2)^2}\sqrt{(1865) - \frac{1}{11}(-49)^2}} = \frac{1401.9091}{(37.5984)(40.5799)} = 0.9188$$

Now $\bar{x} = A_x + \frac{\sum d_x}{n} = 65 + \frac{2}{11} = 65.1818$

$$\bar{y} = A_y + \frac{\sum d_y}{n} = 70 + \frac{-49}{11} = 65.5455$$

Also $\sigma_x = \sqrt{\frac{\sum d_x^2}{n} - \left(\frac{\sum d_x}{n}\right)^2} = \sqrt{\frac{1414}{11} - \left(\frac{2}{11}\right)^2} = 11.3363$

and $\sigma_y = \sqrt{\frac{\sum d_y^2}{n} - \left(\frac{\sum d_y}{n}\right)^2} = \sqrt{\frac{1865}{11} - \left(\frac{-49}{11}\right)^2} = 12.2353$

$$\therefore b_{xy} = \frac{r\,\sigma_x}{\sigma_y} = \frac{(0.9188)(11.3363)}{12.2353} = 0.8513$$

$$b_{yx} = \frac{r\,\sigma_y}{\sigma_x} = \frac{(0.9188)(12.2353)}{11.3363} = 0.9917$$

Equations of lines of regression are:

$y - \overline{y} = b_{yx}(x - \overline{x})$, $x - \overline{x} = b_{xy}(y - \overline{y})$

$\Rightarrow y - 65.55 = 0.99(x - 65.18)$ and $x - 65.18 = 0.85(y - 65.55)$

$\Rightarrow y = 0.99x + 1.02$ and $x = 0.85y + 9.46$

Example24 The regression equations calculated from a given set of observations for two variables x and y are: $x = 9y + 5$ and $y = kx + 9$

Show that $0 < k < \frac{1}{9}$. Also if $k = \frac{1}{10}$, find $\overline{x}, \overline{y}$ and r

Solution: The two equations of regression are:

$x = 9y + 5$ ...①

$y = kx + 9$ ...②

$\Rightarrow b_{xy} = 9$ and $b_{yx} = k$

$\therefore r^2 = b_{xy} b_{yx} = 9k$

$\Rightarrow r = 3\sqrt{k}$ $\because b_{xy} = 9$ is positive, therefore k and r are also positive

Now $0 < r < 1$ or $0 < 3\sqrt{k} < 1$

$\Rightarrow 0 < 9k < 1$ or $0 < k < \frac{1}{9}$

Now if $k = \frac{1}{10}$, equation ② becomes $10y = x + 90$ ...③

Solving ① and ③, the point of intersection of two regression lines is

$$\overline{x} = 860,\ \overline{y} = 95, \text{ also } r = 3\sqrt{k} = 3\sqrt{\frac{1}{10}} = 0.949$$

Exercise 4

1. Fit a straight line $y = ax + b$ to the following data

x	0	1	3	6	8
y	1	3	2	5	4

2. Fit a straight line $y = a + bx$ to the following data

x	25	19	50	36	40	45	30
y	77	76	85	80	82	83	79

3. Fit a second degree parabola $y = ax^2 + bx + c$ to the following data

x	0	1	2	3	4
y	1	1.8	1.3	2.5	6.3

4. Fit a second degree parabola $y = a + bx + cx^2$ to the following data

x	2	4	6	8	10
y	3.07	12.85	31.47	57.38	91.29

5. Find the coefficient of correlation between x and y from the given data. Also find the two lines of regression.

x	1	2	3	4	5	6	7	8	9	10
y	10	12	16	28	25	36	41	49	40	50

6. Find the rank correlation for the following data:

x	56	42	72	36	63	47	55	49	38	42	68	60
y	147	125	160	118	149	128	150	145	115	140	152	155

7. Following table shows ages of husband and wife of 53 married couples.

Wife→ *Husband*↓	15-25	25-35	35-45	45-55	55-65	65-75
15-25	1	1	-	-	-	-
25-35	2	12	1	-	-	-
35-45	-	4	10	1	-	-
45-55	-	-	3	6	1	-
55-65	-	-	-	2	4	2
65-75	-	-	-	-	1	2

Calculate the coefficient of correlation between the age of the husband and that of wife.

8. The regression equations of two variables x and y are $x = 0.7y + 5.2$, $y = 0.3x + 2.8$. Find the means of the two variables and the coefficient of correlation between them.
9. If the coefficient of correlation between two variables x and y is 0.5 and the acute angle between their lines of regression is $\tan^{-1}\frac{3}{8}$, show that $\sigma_x = \frac{\sigma_y}{2+\sqrt{3}}$
10. From a partially destroyed lab data, following results were retrieved: Lines of regression are:

$$8x = 10y - 66 \text{ and } 18y = 40x - 214\ ,\ \sigma_x = 3$$

Find $\overline{x}, \overline{y}, \sigma_y$ and r for the existing data.

Answers

1. $y = 0.38x + 1.6$
2. $y = 70.052 + 0.292x$
3. $y = 0.55x^2 - 1.07x + 1.42$
4. $y = 0.34 - 0.78x + 0.99x^2$
5. $r = 0.96$, $x = 0.2y - 0.64$, $y = 4.69x + 4.9$
6. 0.932
7. 0.91
8. $\bar{x} = 9.06$, $\bar{y} = 5.52$, $r = 0.46$

10. $\overline{x} = 13$, $\bar{y} = 17$, $\sigma_y = 4$, $r = 0.6$