UNIT-V

CORRELATION AND REGRESSION

Curve fitting- method of least squares- correlation-rank correlation-regression-multiple and partial correlation-plane of regression-coefficient of multiple correlation- coefficient of partial correlation

Curve fitting by the method of least square:

1. Straight line, $y = ax + b$ (or) $y = a + bx$

2. Parabola, $y = ax^2 + bx + c$ (or) $y = a + bx + cx^2$

3. Exponential, $y = ae^{bx}$

Straight line:

St. line equation	Normal equation
1. $y = ax + b$	$\sum xy = a\sum x^2 + b\sum x$ $\sum y = a\sum x + nb$
2. $y = a + bx$	$\sum xy = a\sum x + b\sum x^2$ $\sum y = na + b\sum x$

1. Fit a st. line for the following independent variable by the method of least squaring. [Nov 2014]

x :	**1**	**2**	**3**	**4**	**5**
y :	**4**	**3**	**6**	**7**	**11**

Solution:

Let the st. line eqn be, $y = ax + b$ ----------(1)

The normal equation is,

$\sum xy = a\sum x^2 + b\sum x$ ----------(2)

$\sum y = a\sum x + nb$ ----------(3)

Hence, n=5

x	y	x^2	xy
1	4	1	4
2	3	4	6
3	6	9	18
4	7	16	28
5	11	25	55
$\sum x$ =15	$\sum y$ =31	$\sum x^2$ = 55	$\sum xy$ =111

Sub these values in eqn (2) and (3)

111=55a +15b

31=15a+5b

Solving this eqn we get

a =1.8 and b =0.8

Substitute in $y = ax + b$ we get

$y = 1.8x + 0.8$

2. Fit a 1st degree curve to the following data ad eliminate the value of y when x=73.

x :	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**
y :	**1**	**3**	**5**	**10**	**6**	**4**	**2**	**1**

Solution:

The St. line equation is $y = ax + b$ ----------(1)

The normal equation is,

$\sum xy = a\sum x^2 + b\sum x$ ----------(2)

$\sum y = a\sum x + nb$ ----------(3)

Hence, n=8

x	y	x^2	xy
10	1	100	10
20	3	400	60
30	5	900	150
40	10	1600	400

50	6	2500	300
60	4	3600	240
70	2	4900	140
80	1	6400	80
$\sum x = 360$	$\sum y = 32$	$\sum x^2 = 20400$	$\sum xy = 1380$

Substitute these values in equation (1) and (2) we get

1380=20400a+360b

32=360a+8b

Solving these eqn we get

a =-0.014 and b=4.642

Substitute in $y = ax + b$

y =-0.014 x +4.642

Given When x =73, sub in above equation

y =-0.014(73)+4.642

y =3.61

3.Fit a st. line for the curve $y = ax + \frac{b}{x}$ for the following data: [NOV 2017]

x :	**1**	**2**	**3**	**4**
y :	**- 1.5**	**0.99**	**3.88**	**7.66**

Solution:

The given st. line equation is,

$$y = ax + \frac{b}{x} \text{----------(*)}$$

$$xy = ax^2 + b$$

let, $Y = xy$

$X = x^2$

Sub in eqn (*)

$Y = aX + b$ ----------(1)

The normal equations are,

$XY = a\sum X^2 + b\sum X$ ----------(2)

$\sum Y = a\sum X + nb$ ----------(3)

x	y	$X = x^2$	$Y = xy$	X^2	XY
1	-1.5	1	-1.5	1	-1.5
2	0.99	4	1.98	16	7.92
3	3.88	9	11.64	81	104.76
4	7.66	16	30.64	256	490.24
$\sum x$ =10	$\sum y$ =6.16	$\sum X$ = 30	$\sum Y$ = 42.76	$\sum X^2$ = 354	$\sum XY$ = 601.42

Sub in eqn (1) and (2) we get

601.42=354a+30b

42.76=30a+ 4b

Solving these two equations we get a =2.17 and b =-5.63

Y =2.17X -5.63

Replace X and Y values here

xy =2.17 x^2 - 5.63

y =2.17 x -5.63/ x

Fitting a Parabola:

Parabola Equation	Normal Equation
1. $y = ax^2 + bx + c$	$\sum x^2 y = a\sum x^4 + b\sum x^3 + c\sum x^2$ $\sum xy = a\sum x^3 + b\sum x^2 + c\sum x$ $\sum y = a\sum x^2 + b\sum x + nc$
2. $y = a + bx + cx^2$	$\sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4$ $\sum xy = a\sum x + b\sum x^2 + c\sum x^3$ $\sum y = na + b\sum x + c\sum x^2$

1. Fitting a parabola by the method of least square. [Nov 2016]

x :	**1**	**2**	**3**	**4**	**5**
y :	**2**	**3**	**5**	**8**	**10**

Solution:

Let, the parabola equation be,

$$y = ax^2 + bx + c \text{ ----------} (1)$$

The normal equations are,

$$\sum x^2 y = a\sum x^4 + b\sum x^3 + c\sum x^2 \text{ ----------} (2)$$
$$\sum xy = a\sum x^3 + b\sum x^2 + c\sum x \text{ ----------} (3)$$
$$\sum y = a\sum x^2 + b\sum x + nc \text{ ----------} (4)$$

x	y	x^2	x^3	x^4	xy	x^2y
1	2	1	**1**	**1**	2	2
2	3	4	8	16	6	12
3	5	9	27	81	15	45
4	8	16	64	256	32	128
5	10	25	125	625	50	250
$\sum x = 15$	$\sum y = 28$	$\sum x^2 = 55$	$\sum x^3 = 225$	$\sum x^4 = 979$	$\sum xy = 105$	$\sum x^2y = 437$

Substitute in eqn (1),(2) and (3)

$$437 = 979a + 225b + 55c$$

$$105 = 225a + 55b + 15c$$

$$28 = 55a + 15b + 5c$$

Solving this we get

a =0.2142, b =0.8142, c =0.8

Sub a,b,c values in $y = ax^2 + bx + c$

$$y = 0.2142x^2 + 0.8142x + 0.8$$

2. Fit a parabola of the curve $y = a + bx + cx^2$ by the method of least square. [Dec 2011]

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

Solution:

The equations of the parabola is,

$$y = a + bx + cx^2 \text{ ----------} (1)$$

The normal equations are,

$$\sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4 \text{ ----------} (2)$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3 \text{ ----------(3)}$$

$$\sum y = na + b\sum x + c\sum x^2 \text{ ----------(4)}$$

x	y	x^2	x^3	x^4	xy	x^2y
2	3.07	4	8	16	6.14	12.28
4	12.85	16	64	256	51.4	205.6
6	31.47	36	216	1296	188.82	1132.92
8	57.38	64	512	4096	459.04	3672.32
10	91.29	100	1000	100000	912.9	9129
$\sum x = 30$	$\sum y = 196.06$	$\sum x^2 = 220$	$\sum x^3 = 1800$	$\sum x^4 = 15664$	$\sum xy = 1618.3$	$\sum x^2y = 14152.12$

Sub in eqn (2), (3), (4) we get

$$14152.112 = 220a + 1800b + 15664c$$

$$1618.3 = 30a + 220b + 1800c$$

$$196.06 = 5a + 30b + 220c$$

$$a = 0.696$$

$$b = -0.8550$$

$$c = 0.9919$$

sub in eqn (1)

$$y = 0.696 - 0.855x + 0.9919x^2$$

3. Fit a parabola $y = ax^2 + bx + c$ in least squares sense to the data: [MAY 2017, NOV 2017]

X: 10 12 15 23 20
Y: 14 17 23 25 21

Let, the parabola equation be,

$y = ax^2 + bx + c$ ----------(1)

The normal equations are,

$\sum x^2y = a\sum x^4 + b\sum x^3 + c\sum x^2$ ----------(2)

$\sum xy = a\sum x^3 + b\sum x^2 + c\sum x$ ----------(3)

$\sum y = a\sum x^2 + b\sum x + nc$ ----------(4)

X	**y**	**x^2**	**x^3**	**x^4**	**xy**	**x^2y**
10	14	100	1000	10000	140	1400
12	17	144	1728	20736	204	2448
15	23	225	3375	50625	345	5175
23	25	529	12167	279841	575	13225
20	21	400	8000	160000	420	8400
$\sum x = 80$	**$\sum y = 100$**	**$\sum x^2 = 1398$**	**$\sum x^3 = 26270$**	**$\sum x^4 = 521202$**	**$\sum xy = 1684$**	**$\sum x^2y = 30648$**

Substitute in eqn (1),(2) and (3)

30648 = 521202a + 26270b + 1398c

$1684 = 26270a + 1398b + 80c$

$100 = 1398a + 80b + 5c$

Solving this we get

$a = -0.0695, \quad b = 3.00, \quad c = -8.728$

Sub a,b,c values in $y = ax^2 + bx + c$

$y = -0.0695x^2 + 3.00x - 8.728$

4. Fit a parabola to the following data; also estimate y at x=6.

x :	**1**	**2**	**3**	**4**	**5**
y:	**5**	**12**	**26**	**60**	**97**

Solution:

Let, the parabola equation be,

$y = ax^2 + bx + c$ ----------(1)

The normal equations are,

$\sum x^2y = a\sum x^4 + b\sum x^3 + c\sum x^2$ ----------(2)

$\sum xy = a\sum x^3 + b\sum x^2 + c\sum x$ ----------(3)

$\sum y = a\sum x^2 + b\sum x + nc$ ----------(4)

x	**y**	**x^2**	**x^3**	**x^4**	**xy**	**x^2y**
1	5	1	1	1	5	5
2	12	4	8	16	24	48
3	26	9	27	81	78	234
4	60	16	64	256	240	960
5	97	25	125	625	485	2425
$\sum x = 15$	**$\sum y = 200$**	**$\sum x^2 = 55$**	**$\sum x^3 = 225$**	**$\sum x^4 = 979$**	**$\sum xy = 832$**	**$\sum x^2y = 3672$**

Substitute in eqn (1),(2) and (3)

$3672 = 979a + 225b + 55c$

$832 = 225a + 55b + 15c$

$200 = 55a + 15b + 5c$

Solving this we get

$a = 5.714, \quad b = -11.085, \quad c = 10.4$

Sub a,b,c values in $y = ax^2 + bx + c$

$y = 5.714x^2 - 11.085x + 10.4$ and when $x = 6$, $y = 149.59$

Fitting of exponential curve:

Type-I:

Let, $y = ae^{bx}$

1.Find the curve for $y=ae^{bx}$ for the following data:

x :	**0**	**2**	**4**	**6**	**8**
y :	**150**	**63**	**28**	**12**	**5.6**

Solution:

Let $y = ae^{bx}$ ----------(1)

Taking log on both sides,

$$\log y = \log\left(ae^{bx}\right)$$

$$\log y = \log a + \log e^{bx}$$

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

$Y = A + Bx$ ----------(2) which is a straight line equation

Where $Y = \log y$, $A = \log a$, $B = b \log e$

The normal equations are,

$$\sum xY = A\sum x + B\sum x^2 \text{ ----------(3)}$$
$$\sum Y = nA + B\sum x \text{ ----------(4)}$$

x	y	x^2	$Y = \log y$	xY
0	150	0	2.1760	0
2	63	4	1.7993	3.5986
4	28	16	1.4471	5.7886
6	12	36	1.0791	6.4750
8	5.6	64	0.7481	5.9855
$\sum x = 20$		$\sum x^2 = 120$	$\sum Y = 7.2496$	$\sum xY = 21.8477$

Sub these values in eqn (3) and (4) we get

21.8477 = 20 A + 120B

7.2496 = 5A + 20B

A=2.1649

B=-0.1786

Sub in (2) we get

Y = 2.1649X - 0.1786

But

A = log a

a = antilog A

a = 146.1840

B = b log e

b = B/log e

=-0.1786/0.4343

b = -0.4112

sub in (1)

$$y = (146.1840)e^{-0.4112x}$$

2. Find the curve $y = ae^{bx}$ for the following data: [Nov 15]

x :	**0**	**5**	**8**	**12**	**20**
y :	**3.0**	**1.5**	**1.0**	**0.55**	**0.18**

Given:

$$y = ae^{bx}$$

Taking log on both sides,

$$\log y = \log\left(ae^{bx}\right)$$

$$\log y = \log a + \log e^{bx}$$

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

$Y = A + Bx$ ----------(2) which is a straight line equation

Where $Y = \log y$, $A = \log a$, $B = b\log e$

The normal equations are,

$$\sum xY = A\sum x + B\sum x^2 \text{ ----------(2)}$$
$$\sum Y = nA + B\sum x \text{ ----------(3)}$$

x	y	$Y = \log y$	x^2	xY
0	3	0.4771	0	0
5	1.5	0.1760	25	0.8804

8	1	0	64	0
12	0.55	-0.2596	144	-3.1156
20	0.18	-0.7447	400	-14.8945
$\sum x = 45$		$\sum Y = -0.3512$	$\sum x^2 = 633$	$\sum xY = -17.1297$

Sub in eqn (2) and (3)

-17.1297 = 45 A + 633 B

-0.3512 = 5 A + 45 B

A=0.4811

B=-0.0612

Y=0.4811 – 0.0612x

But a = antilog A

a = 3.0276

b= B/log e

= -0.0612/0.4343

b = -0.1409

sub in eqn (1)

$y=(3.0276)e^{-0.1409x}$

Type – II: $y = ab^x$

1.Fit a curve of the form $y = ab^x$ from the following table: [Nov 16, Dec 12]

x :	**0**	**1**	**2**	**3**	**4**	**5**	**6**
y :	**32**	**47**	**65**	**92**	**132**	**190**	**275**

Solution:

Let, $y = ab^x$ ----------(1)

Taking log on both sides,

$\log y = \log\left(ab^x\right)$

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

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

$Y = A + Bx$ Where $Y = \log y$, $A = \log a$, $B = \log b$

$Y = A + Bx$, which is a straight line equation

The normal equations are,

$$\sum xY = A\sum x + B\sum x^2 \text{ ----------} (2)$$
$$\sum Y = nA + B\sum x \text{ ----------} (3)$$

x	y	$Y = \log y$	x^2	xY
0	32	1.5051	0	0
1	47	1.6720	1	1.6720
2	65	1.8129	4	3.6258
3	92	1.9637	9	5.8913
4	132	2.1205	16	8.4822
5	190	2.2787	25	11.3937
6	275	2.4393	36	14.6359
$\sum x = 21$		$\sum Y = 13.7922$	$\sum x^2 = 91$	$\sum xY = 45.7009$

Sub in eqn (3) and (4) we get

45.7009 = 21 A + 91 B

13.7922 =7 A + 21 B

A=1.5069

B=0.1544

Sub in eqn (2)

$Y = 1.5069 + 0.1544x$

But a = antilog A

a = 32.1292

b = antilog B

b = 1.4269

Sub in eqn (1)

$y = 32.1292(1.4269)^x$ y = 32.1292 $(1.4269)^x$

2. Fit a curve of the form $y = ab^x$ for the following data by the method of least squares:

x: 1 2 3 4 5 6

y : **151** **100** **61** **50** **20** **8**

Solution:

Let, $y = ab^x$ ----------(1)

Taking log on both sides,

$\log y = \log\left(ab^x\right)$

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

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

$Y = A + Bx$ Where $Y = \log y$, $A = \log a$, $B = \log b$

$Y = A + Bx$, which is a straight line equation

The normal equations are,

$\sum xY = A\sum x + B\sum x^2$ ----------(2)

$\sum Y = nA + B\sum x$ ----------(3)

x	y	$Y = \log y$	x^2	xY
1	151	2.179	1	2.179
2	100	2	4	4
3	61	1.7853	9	5.3559
4	50	1.699	16	6.796
5	20	1.301	25	6.505
6	8	0.9031	36	5.4186
$\sum x = 21$		$\sum Y = 9.8674$	$\sum x^2 = 91$	$\sum xY = 30.2545$

Sub in eqn (3) and (4) we get

30.2545 = 21 A + 91 B

9.8674 = 6 A + 21 B

A = 2.498

B = - 0.244

Sub in eqn (2)

$y = 2.498 - 0.244x$

But a = antilog A

a = 314.775

b = antilog

B b = 0.570

Sub in eqn (1)

$$y = 314.775(0.570)^x$$

3. Fit a curve of the form $y=ab^x$ to the data.

x :	**2**	**3**	**4**	**5**	**6**
y :	**144**	**172.8**	**207.4**	**248.8**	**298.5**

Solution:

Let, $y = ab^x$ ----------(1)

Taking log on both sides, log y = log $[ab^x]$

log y = log a + x log b

Y = A + B x ----------(2)

Where Y = log y ; A = log a, B = log b which is a straight line equation

The normal equations are,

$\sum xY = A\sum x + B\sum x^2$----------(3)

$\sum Y = n A + B\sum x$ ----------(4)

x	y	Y = log y	x^2	xY
2	144	2.15836	4	**4.316**
3	172.8	2.2375	9	**6.712**
4	207.4	2.3168	16	**9.264**
5	248.8	2.3958	25	11.979
6	298.5	2.4749	36	14.8994
$\sum x = 20$		$\sum Y = 11.58336$	$\sum x^2 = 90$	$\sum xY = 45.125$

Sub in eqn (3) and (4) we get

47.125 = 20 A + 90 B

11. 58336 = 5A + 20 B

A = 2.00

B = 0.079

Sub in eqn (2)

Y = 1.5069 + 0.1544 x

But a = antilog A

a = 100

b = antilog B

b = 1.199

Sub in eqn (1)

$y = 100\ (1.199)^x$

Type – III :

Let , $y = ax^b$ be the power curve.

1.Fit the power curve of the form $y = ax^b$ for the following data:

x :	**1**	**2**	**3**	**4**	**5**
y :	**7.1**	**27.8**	**62.1**	**110**	**161**

Solution:

$y = ax^b$

Taking log on both sides,

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

$Y = A + bX$ (Straight line equation)

The normal equations are,

$\sum XY = A\sum X + b\sum X^2$ ----------(1)

$\sum Y = nA + b\sum X$ ----------(2)

x	y	$X = \log x$	$Y = \log y$	X^2	XY
1	7.1	0	0.8512	0	0
2	27.8	0.3010	1.440	0.0906	0.4346
3	62.1	0.4771	1.7930	0.2276	0.8554
4	110	0.6020	2.0413	0.3624	1.2289
5	161	0.6989	2.2068	0.4885	1.5423
$\sum x = 15$		$\sum X = 2.079$	$\sum Y = 8.3363$	$\sum X^2 = 1.1691$	$\sum XY = 4.0612$

Substitute in eqn (1) and (2)

4.0612 = 2.079 A + 1.1691 b

8.3363 = 5 A + 2.079 b

Solving this we get

A=0.8552 a =anti log (A) = 7.1647

b =1.9529 Sub a and b values in (1) $y = 7.1647x^{1.9529}$

$Y = 0.8552 + 1.9529X$

3.5. CORRELATION

In this section, we introduce the concept of correlation which is one of the methods of studying relationship between two variables. If the change in one variable affects a change in the other variables, the variables are said to be correlated.

Consider a set of bivariate data $(x_i, y_i), i = 1, 2, \ldots, n$. If there is a change in one variable corresponding to a change in the other variable, we say that the variables are correlated.

If the two variables deviate in the same direction the correlation is said to be direct or positive. If they always deviate in the opposite direction, then the correlation is said to be inverse or negative. If the change in one variable corresponds to a proportional change in the other variable, then the correlation is said to be perfect.

Height and weight of a batch of students; Income and expenditure of a family are examples of variables with positive correlation. Price and demand; volume *v* and pressure *p* of a perfect gas which obeys the law *pv=k* where *k* is a constant, are examples of variables with negative correlation.

The Karl Pearson's coefficients of correlation is defined by,

$$r = \frac{N\Sigma dxdy - \Sigma dx \Sigma dy}{\sqrt{N\Sigma dx^2 - (\Sigma dx)^2}\sqrt{N\Sigma dy^2 - (\Sigma dy)^2}}$$

$dx = x - A$ or $x - \bar{x}$, $dy = y - B$ or $y - \bar{y}$

Note

Correlation lies between –1 and 1
If r is positive, then it is called positive correlation
If r is negative, then it is called negative correlation

Example 1

Calculate the correlation coefficient for the following data

x	70	75	60	71	80	67	63	81
y	60	65	70	73	82	69	68	70

Solution

x	y	$dx = x - \bar{x}$	$dy = y - \bar{y}$	dx^2	dy^2	$dxdy$
70	60	–1	–10	1	100	10
75	65	4	–5	16	25	–20
60	70	–11	0	121	0	0
71	73	0	3	0	9	0
80	82	9	12	81	144	108
67	69	–4	–1	16	1	–16
63	68	–8	–2	64	4	–128
81	70	10	0	100	0	0
Σx =567	Σy 557	Σdx –1	Σdy =–3	Σdx^2 =399	Σdy^2 =283	$\Sigma dxdy = 82$

$$\bar{x} = \frac{\Sigma x}{N} = \frac{567}{8} = 70.8 = 71$$

$$\bar{y} = \frac{\Sigma y}{N} = \frac{557}{8} = 69.6 = 70$$

The correlation coefficient is

$$r = \frac{N\Sigma dxdy - \Sigma dx \Sigma dy}{\sqrt{N\Sigma dx^2 - (\Sigma dx)^2}\sqrt{N\Sigma dy^2 - (\Sigma dy)^2}} = \frac{8(82) - (-1)(-3)}{\sqrt{8(399) - (-1)^2}\sqrt{8(283) - (-3)^2}}$$

$$= \frac{656}{(56.5)(47.4)} = \frac{656}{2734.6}$$

$r = 0.023.$

Example 2

Find the correlation coefficient for the following data ***(November- 2001)***

x	60	65	68	70	71	75	85	82
y	65	62	65	73	70	72	80	85

Solution

x	y	$dx = x - \bar{x}$	$dy = y - \bar{y}$	dx^2	dy^2	$dxdy$
60	65	– 12	–7	144	49	84
65	62	–7	–10	49	100	70
68	65	–4	–7	16	49	28

70	73	–2	1	4	1	–2
71	70	–1	–2	1	4	2
75	72	3	0	9	0	0
85	80	13	8	169	64	104
82	85	10	13	100	169	130
$\sum x = 576$	$\sum y = 572$	$\Sigma dx = 0$	$\Sigma dy = -4$	$\Sigma dx^2 = 492$	$\Sigma dy^2 = 436$	$\Sigma dxdy = 416$

$$\bar{x} = \frac{\Sigma x}{N} = \frac{576}{8} = 72$$

$$\bar{y} = \frac{\Sigma y}{N} = \frac{572}{8} = 71.5 = 72$$

The correlation coefficient is

$$r = \frac{N\Sigma dxdy - \Sigma dx\Sigma dy}{\sqrt{N\Sigma dx^2 - (\Sigma dx)^2}\sqrt{N\Sigma dy^2 - (\Sigma dy)^2}}$$

$$= \frac{8(416) - 0}{\sqrt{8(492) - 0}\sqrt{8(436) - (-4)^2}} = \frac{3328}{(62.737)(59.19)}$$

$$r = 0.896.$$

Example 3

Find the correlation for the following data ***(November- 2003)***

x	68	64	69	65	66	70
y	81	85	80	88	87	82

Solution

x	y	$dx = x - \bar{x}$	$dy = y - \bar{y}$	dx^2	dy^2	$dxdy$
68	81	1	–3	1	9	–3
64	85	–3	1	9	1	–3
69	80	–2	–4	4	16	–8
65	88	–2	4	4	16	–8
66	87	–1	3	1	9	–3
70	82	3	-2	9	4	–6
$\sum x = 402$	$\sum y = 503$	$\sum dx = 0$	$\sum dy = -1$	$\sum dx^2 = 28$	$\sum dy^2 = 55$	$\sum dxdy = -31$

$$\bar{x} = \frac{\Sigma x}{N} = \frac{402}{6} = 67$$

$$y = \frac{\Sigma y}{N} = \frac{503}{6} = 83.8 = 84$$

The correlation coefficient is,

$$r = \frac{N\Sigma dxdy - \Sigma dx\Sigma dy}{\sqrt{N\Sigma dx^2 - (\Sigma dx)^2}\sqrt{N\Sigma dy^2 - (\Sigma dy)^2}}$$

$$= \frac{6(-31) - 0}{\sqrt{6(28) - 0}\sqrt{6(55) - (-1)^2}} = \frac{-186}{(12.96)(18.138)}$$

$$= \frac{-186}{235.073} = 0.791$$

$\therefore r = 0.791.$

Example 4

Find the correlation coefficient for the following data ***(May- 2002)***

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

Solution

x	y	$dx = x - \bar{x}$	$dy = y - \bar{y}$	dx^2	dy^2	$dxdy$
1	9	–4	–3	16	9	12
2	8	–3	–4	9	16	12
3	10	–2	–2	4	4	4
4	12	–1	0	1	0	0
5	11	0	–1	0	1	0
6	13	1	1	1	1	1
7	14	2	2	4	4	4
8	16	3	4	9	16	12
9	15	4	3	16	9	12
$\sum x = 45$	$\sum y = 108$	$\sum dx = 0$	$\sum dy = 0$	$\sum dx^2 = 60$	$\sum dy^2 = 60$	$\sum dxdy = 57$

$$\bar{x} = \frac{\Sigma x}{N} = \frac{45}{9} = 5$$

$$\bar{y} = \frac{\Sigma y}{N} = \frac{108}{9} = 12$$

The correlation coefficient is

$$r = \frac{N\Sigma dxdy - \Sigma dx\Sigma dy}{\sqrt{N\Sigma dx^2 - (\Sigma dx)^2}\sqrt{N\Sigma dy^2 - (\Sigma dy)^2}}$$

$$= \frac{9(57) - 0}{\sqrt{9(60) - 0}\sqrt{9(60) - 0}} = \frac{513}{(23.2)(23.2)}$$

$$= \frac{513}{540}$$

$$r = 0.95$$

Exercise

1. Calculate the correlation coefficient for the following data

Height of fathers	65	66	67	67	68	69	70	72
Height of sons	67	68	65	68	72	72	69	71

(Ans. r =0.603) ***(April-2010)***

2. Ten individuals got the following percentage of marks in two subjects I and II

Roll No	1	2	3	4	5	6	7	9	8	10
Marks in sub I	78	36	98	25	75	82	90	62	65	39
Marks in Sub II	84	51	91	60	68	62	86	58	53	47

Calculate the coefficient of correlation. *(Ans. r = 0.741)*

3. Calculate the coefficient of correlation between the values *x* and *y* from the following data.

x	78	89	97	69	59	79	68	61
y	125	137	156	112	107	136	123	108

(November- 2002) *(Ans. r =0.948)*

4. Calculate the correlation coefficient for the following data.

x	65	66	67	67	68	69	70	72
y	67	68	65	68	72	72	69	71

(May- 2009) *(Ans. r =0.347)*

5. Find the coefficient of correlation

x	62	64	65	69	70	71	72	74
y	126	125	139	145	165	152	180	208

(Ans. r =0.9)

6. Calculate the correlation coefficient for the following data.

x	60	62	64	66	68	70	72
y	61	63	63	63	64	65	67

(Ans. r =0.939)

3.6. RANK CORRELATION

Let $(x_i, y_i), i = 1, 2, \ldots, n$ be the ranks of 'i' individuals in two characteristics A and B respectively. Pearson's coefficients of correlation between the ranks x_i's and y_i's is called the rank correlation coefficients between the two characteristic A and B for that group of individuals. It is given by, $\rho = 1 - \dfrac{6\Sigma d_i^2}{n(n^2-1)}$, where $d_i = (x_i, y_i)$.

If the rank is repeated, then the rank correlation coefficient is,

$$\rho = 1 - \frac{6\left[\Sigma d^2 + \dfrac{m(m^2-1)}{12}\right]}{n(n^2-1)}$$

Here $\dfrac{m(m^2-1)}{12}$ is the correction factor that is to be calculated for each repeated rank.

Example 1

The ranking of 10 students in 2 subjects x and y are

x	3	5	8	4	7	10	2	1	6	9
y	6	4	9	8	1	2	3	10	5	7

Find the coefficient of rank correlation ***(May- 2001)***

Solution

x	*y*	*d = x -y*	d^2
3	6	– 3	9
5	4	1	1
8	9	– 1	1

4	8	– 4	16
7	1	6	36
10	2	8	64
2	3	– 1	1
1	10	– 9	81
6	5	1	1
9	7	2	4
			$\Sigma d^2 = 214$

$$\rho = 1 - \frac{6\Sigma di^2}{n(n^2-1)} = 1 - \frac{6(24)}{10(100-1)}$$

$$= 1 - \frac{1284}{990} = 1.2961$$

Example 2

Find the rank correlation coefficient for the following data

x	40	80	64	35	90
y	35	40	28	80	84

Solution

x	y	Rank in x	Rank in y	$d = x - y$	d^2
40	35	4	4	0	0
80	40	2	3	-1	1
64	28	3	5	-2	4
35	80	5	2	3	9
90	84	1	1	0	0
					$\Sigma d^2 = 14$

Here n = 5

The rank correlation coefficient is defined by,

$$\rho = 1 - \frac{6\Sigma di^2}{n(n^2-1)}$$

$$= 1 - \frac{6(14)}{5(25-1)} = 1 - \frac{84}{120} \qquad = 0.3$$

Example 3

Ten students got the following marks in chemistry and physics

Marks in chemistry	78	36	98	25	75	82	90	62	65	39
Marks in Physics	84	51	91	60	68	62	86	58	63	47

Solution

x	y	Rank in x x_i	Rank in y y_i	$d = x_i - y_i$	d^2
78	84	4	3	1	1
36	51	9	9	0	0
98	91	1	1	0	0
25	60	10	7	3	9
75	68	5	4	1	1
82	62	3	6	– 3	9
90	86	2	2	0	0
62	58	7	8	– 1	1
65	63	6	5	1	1
39	47	8	10	– 2	4
					$\Sigma d^2 = 26$

The rank correlation coefficient is

$$\rho \qquad = 1 - \frac{6\Sigma d_i^{\,2}}{n(n^2-1)} = 1 - \frac{6(26)}{10(100-1)} = 1 - \frac{156}{990} \qquad = 0.842$$

Example 4

Find the rank correlation coefficient for the following data ***(November-2011)***

x	11	10	14	12	8	9
y	7	6	8	6	1	2

Solution

x	y	Rank in x	Rank in y	$d = x - y$	d^2
11	7	3	5	– 2	4
10	6	4	3.5	0.5	0.25
14	8	1	6	– 5	25
12	6	2	3.5	– 1.5	2.25
8	1	6	1	5	25
9	2	5	2	3	9
					$\Sigma d^2 = 65.5$

We observe that in the rank of y, 6 occur twice. Hence in the calculation of the rank correlation coefficient $\sum (x-y)^2$ is to be corrected by adding the Correction factor .

Correction factor = $\frac{m(m^2-1)}{12} = \frac{2(4-1)}{12} = 0.5$

The spearman's rank correlation coefficient is defined by,

$$\rho = 1 - \frac{6\left[\Sigma d^2 + \frac{m(m^2-1)}{12}\right]}{n(n^2-1)} = 1 - \frac{6(65.5+0.5)}{6(35)}$$

$$= 1 - \frac{6(66)}{210} = -0.8857$$

Example 5

Find the rank correlation coefficient from the following data.

x	65	71	65	68	70	72	67	74
y	70	74	68	69	70	65	69	65

Solution

x	y	Rank in x x_i	Rank in y y_i	$d_i = x_i - y_1$	di^2
65	70	7.5	2.5	5	25
71	74	3	1	2	4
65	68	7.5	6	1.5	2.25
68	69	5	4.5	0.5	0.25
70	70	4	2.5	1.5	2.25

72	65	2	7.5	– 5.5	30.25
67	69	6	4.5	1.5	2.25
74	65	1	7.5	– 6.5	42.25
					$\Sigma d^2 = 108.5$

We observe that 65 is repeated for rank 7, 8 and average = $\frac{7+8}{2}$ = 7.5. Hence in the calculation of the rank correlation coefficient, $\sum (x-y)^2$ is to be corrected by adding the Correction factor $CF_1 = \frac{2(2^2-1)}{12} = \frac{1}{2}$.

70 is repeated for rank 2,4 and average = $\frac{2+3}{2} = 2.5, CF_2 = \frac{2(2^2-1)}{12} = \frac{1}{2}$

69 is repeated for rank 4, 5 and average = $\frac{4+5}{2} = 4.5$, $CF_3 = \frac{2(2^2-1)}{12} = \frac{1}{2}$

65 is repeated for rank 7, 8 and average = $\frac{7+8}{2} = 7.5$, $CF_4 = \frac{2(2^2-1)}{12} = \frac{1}{2}$

Correction factor = $CF_1 + CF_2 + CF_3 + CF_4 = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} = 2$

The rank correlation coefficient is

$$\rho = \frac{1-6\left[\Sigma d_i^2 + \frac{m(m^2-1)}{12}\right]}{n(n^2-1)} = 1 - \frac{6\left[108.5+2\right]}{8(64-1)} = -0.315$$

Example 6

Find the rank correlation factor for the following data. ***(January-2011)***

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

Solution

x	y	Rank in x x_i	Rank in y y_i	$d_i = x_i - y_i$	d_i^2
68	62	4	5	– 1	1
64	58	6	7	– 1	1
75	68	2.5	3.5	– 1	1

50	45	9	10	– 1	1
64	81	6	1	+5	25
80	60	1	6	– 5	25
75	68	2.5	3.5	– 1	1
40	48	10	9	1	1
55	50	8	8	0	0
64	70	6	2	4	16
					$\Sigma d^2 = 72$

We observe that 75 is repeated for the rank 2, 3 and average = $\frac{3+2}{2}$ = 2.5. Hence in the calculation of the rank correlation coefficient, $\sum (x-y)^2$ is to be corrected by adding the Correction factor $CF_1 = \frac{2(2^2-1)}{12} = \frac{1}{2}$.

64 is repeated for the rank 5, 6, 7, average = $\frac{5+6+7}{3} = 6$, $CF_2 = \frac{3(3^2-1)}{12} = 2$

68 is repeated for the rank 3,4, average = $\frac{3+4}{2} = 3.5$, $CF_3 = \frac{2(2^2-1)}{12} = \frac{1}{2}$

Correction factor = $CF_1 + CF_2 + CF_3 = \frac{1}{2} + 2 + \frac{1}{2} = 3$

The rank correlation coefficient is

$$\rho = 1 - \frac{6\left[\Sigma d_i^2 + \frac{m(m^2-1)}{12}\right]}{n(n^2-1)} = 1 - \frac{6[72+3]}{10} = 0.54$$

Example7

Ten competitions in a musical test were ranked by three judges A, B, and C in the following order.

Ranked by *A*	1	6	5	10	3	2	4	9	7	8
Ranked by *B*	3	5	8	4	7	10	2	1	6	9
Ranked by *C*	6	4	9	8	1	2	3	10	5	7

Using rank correlation method, discuss which pairs of judges have the nearest approach to common links in music. ***(May-2011)***

Solution

n =10

Ranked by *A*	Ranked by *B*	Ranked by *C*	d_1 = *x* – *y*	d_2 = *x*-*z*	d_3 = *y*–*z*	d_1^2	d_2^2	d_3^2
1	3	6	– 2	– 5	– 3	4	25	9
6	5	4	1	2	1	1	4	1
5	8	9	– 3	– 4	– 1	9	16	1
10	4	8	6	2	– 4	36	4	16
3	7	1	– 4	2	6	16	4	36
2	40	2	– 8	0	8	64	0	64
4	2	3	2	1	– 1	4	1	1
9	1	10	8	– 1	– 9	64	1	81
7	6	5	1	2	1	1	4	1
8	9	7	– 1	1	2	1	1	4
			Σd_1 = 0	Σd_2=0	Σd_3 =0	Σd_1^2=200	Σd_2^2 =60	Σd_3^2=214

$$\rho(x,y) = 1 - \frac{6\Sigma d_1^2}{n(n^2-1)}$$

$$= 1 - \frac{6\times 200}{10\times 99} = 1 - \frac{40}{33}$$

$$= -0.212$$

$$\rho(x,z) = 1 - \frac{6\Sigma d_2^2}{n(n^2-1)}$$

$$= 1 - \frac{6\times 60}{10\times 99}$$

$$= 1 - \frac{4}{11}$$

$$= 0.636$$

$$\rho(y,z) = 1 - \frac{6\Sigma d_3^2}{n(n^2-1)}$$

$$= 1 - \frac{6\times 214}{10\times 99} = 1 - 1.296$$

$$= -0.296$$

Since $\rho(x, z)$ is maximum, we conclude that the pair of judges x and z have the nearest approach to common links in music.

Example 8

The data given below are marks scored by the students in statistics and mathematics. Find the rank correlation coefficient

x	18	78	46	37	47	56	82	47	46	28	46	75	81	47	50
y	46	81	38	44	38	47	75	56	63	71	63	78	95	45	67

Solution

x	y	Rank in x x_i	Rank in y y_i	$d_i = x_i - y_i$	d_i^2
18	46	15	11	4	16
78	81	3	2	1	1
46	38	11	14.5	−3.5	12.25
37	44	13	13	0	0
47	38	8	14.5	−6.5	42.25
56	47	5	10	−5	25
82	75	1	4	−3	9
47	56	8	9	−1	1
46	63	11	7.5	7.5	12.25
28	71	14	5	9	81
46	63	11	7.5	7.5	12.25
75	78	4	3	1	1
81	95	2	1	1	1
47	45	8	12	−4	16
50	67	6	6	0	0
					$\Sigma d^2 = 230$

X: 47 is repeated for the ranks 7, 8, 9, average $= \frac{7+8+9}{3} = 8, CF_1 = \frac{3(3^2-1)}{12} = 2$

46 is repeated for the ranks 10, 11, 12, average $= \frac{10+11+12}{3} = 11, CF_2 = \frac{3(3^2-1)}{12} = 2$

Y: 63 repeated for the ranks 7, 8, average $= \frac{7+8}{2} = 7.5,\ CF_3 = \frac{2(2^2-1)}{12} = \frac{1}{2}$

38 repeated for the ranks 14, 15, average = $\frac{14+15}{2} = 14.5$, $CF_4 = \frac{2(2^2-1)}{12} = \frac{1}{2}$

Correction factor = $CF_1 + CF_2 + CF_3 + CF_4 = \frac{1}{2} + 2 + \frac{1}{2} + \frac{1}{2} = 5$

The rank correlation coefficient is

$$\rho = 1 - \frac{6\left[\Sigma d_i^2 + \frac{m(m^2-1)}{12}\right]}{n(n^2-1)} = 1 - \frac{6[230+5]}{15(15^2-1)} = 0.580$$

Exercise

1. Find the rank correlation coefficient for the following data.

(i)

x	48	60	72	62	56	40	39	52	30
y	62	78	65	70	38	54	60	32	31

(Ans. 0.668)

(ii)

x	45	56	39	54	45	40	56	60	30	36
y	40	36	30	44	36	32	45	42	20	36

(Ans. 0.764)

(iii)

x	1	4	2	3	5
y	3	1	2	5	4

(Ans.0.1)

3.7. REGRESSION

If we fit a straight line by the principle of least squares to the points of the scatter diagram in such a way that the sum of the squares of the distance parallel to the y-axis from the points to the line is minimized, we obtain a line of best fit for the data and it is called the regression line of best fit for the data and it is also called the regression line of *y* on *x*.

Regression is the estimation or prediction of unknown values of one variable from known values of another variable. It is the mathematical measure of the average relationship between two or more variables in terms of the original limits of the data.

The regression coefficient of *x* on *y* is given by

$$b_{xy} = r\frac{\sigma x}{\sigma y} \quad \ldots(1)$$

$$= \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(y-\bar{y})^2}$$

The regression coefficient of y on x is

$$b_{yx} = r\frac{\sigma y}{\sigma x} \quad \ldots(2)$$

$$b_{yx} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}$$

The line of regression x on y is

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

The line of regression y on x is

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

Note 1

From the equations (1) and (2)

$$b_{xy}b_{yx} = r\frac{\sigma x}{\sigma y} \times r\frac{\sigma y}{\sigma x} = r^2$$

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

$$r = \pm\sqrt{b_{xy}\, b_{yx}}$$ *Where r is the correlation coefficient of x and y.*

Note 2

1. *If $r = 0$, then the two regression lines becomes $y = \bar{y}$ and $x = \bar{x}$ which are two straight lines parallel to x and y axes respectively and passing through their means y and x*

   They are mutually perpendicular.

2. *If $r = \pm 1$ the two lines of regression will coincide.*

Properties of regression coefficients.

1. Correlation coefficients is the geometric mean between the regression coefficients.
2. If one of the regression coefficients is greater than unity numerically the other must be less than unity numerically.
3. Arithmetic mean of regression coefficients is greater than the correlation coefficient.
4. The correlation coefficient and the two regression coefficients have the same sign.

Angle between two regression lines

If θ is the acute angle between the two regression lines then

$$\tan\theta = \frac{1-r^2}{r}\frac{\sigma x\,\sigma y}{\sigma x^2+\sigma y^2}$$

Proof

The line of regression of *x* on *y* is,

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

$$y-\bar{y} = \frac{1}{b_{xy}}\left(x-\bar{x}\right), \text{ where } b_{xy} = r\frac{\sigma_x}{\sigma_y} \qquad \text{...(1)}$$

Let m_1 be the slope of equation (1), then $m_1 = \frac{1}{b_{xy}} = \frac{\sigma_y}{r\sigma_x}$

The line of regression of y on x is,

$$y-\bar{y} = b_{yx}\left(x-\bar{x}\right), \text{ where } b_{yx} = r\frac{\sigma_y}{\sigma_x} \qquad \text{...(2)}$$

m_2 be the slope of equation (2), then $m_2 = b_{yx}$

We know that

$$\tan\theta = \pm\frac{m_1-m_2}{1+m_1m_2}$$

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

$$= \pm \frac{\sigma_y - r_2\sigma_y}{\frac{r\sigma_x}{1+\frac{\sigma_y^2}{\sigma_x^2}}} = \pm \frac{\sigma_y(1-r_2)}{\frac{r\sigma_x}{\frac{\sigma_x^2+\sigma_y^2}{\sigma_x^2}}}$$

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

Since θ is acute,

$$\tan\theta = \left(\frac{1-r^2}{r}\right)\frac{\sigma x\,\sigma y}{\sigma x^2+\sigma y2}$$

Example 1

Obtain the equation of the lines of regression for the following data

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

Also obtain the estimation of x for y = 70 ***(November- 2002)***

Solution

X	Y	$x-\bar{x}$ $x-68$	$y-\bar{y}$ $y-69$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
65	67	–3	–2	9	4	6
66	68	–2	–1	4	1	2
67	65	–1	–4	1	16	4
67	68	0	–1	1	1	1
68	72	1	3	0	9	0
69	72	2	3	1	9	3
70	69	3	0	4	0	0
72	71	4	2	16	4	8
$\Sigma x =$ 544	$\Sigma y =$ 552	$\sum(x-\bar{x}) =0$	$\sum(y-\bar{y}) =0$	$\sum(x-\bar{x})^2 =$ 36	$\sum(y-\bar{y})^2 =$ 44	$\sum(x-\bar{x})(y-\bar{y}) =24$

$$b_{xy} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(y-\bar{y})^2} = \frac{24}{44} = 0.5454$$

$$b_{yx} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2} = \frac{24}{36} = 0.6667$$

Regression lines of x on y is

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

$$x - 68 = 0.5454\ (y - 69)$$

$$x = 0.545y + 30.395$$

Regression lines of y on x is

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

$$y - 69 = 0.667\ (x - 68)$$

$$y = 0.667x + 23.644$$

To estimate x for a given y, we use the line of regression of x on y.

If $y = 70$, estimated value of x is given by

$$x = 0.545 \times 70 + 30.395$$

$$x = 68.545$$

$\therefore$ The regression coefficient x on y is $b_{xy} = 0.5454$

The regression coefficient y on x is $b_{yx} = 0.667$

The regression line x on y is, $x = 0.545y + 30.395$

The regression line y on x is, $y = 0.667x + 23.644$

If $y = 70$, estimated value of x is, $x = 68.545$

Example 2

From the following data, find (i) two regression line equations

(i) Coefficients of correlation between the marks in statistics and economics

(ii) The most likely marks in statistics where marks in economics are 30.

Marks in Economics	25	28	35	32	31	36	29	38	34	32
Marks in Statistics	43	46	49	41	36	32	31	30	33	39

Solution

Let x any y be the marks in economics and statistics respectively (n = 10)

x	Y	$x-\bar{x}$ x-32	$y-\bar{y}$ y-38	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
25	43	–7	5	49	25	–35
28	46	–4	8	16	64	–32
35	49	3	11	9	121	33
32	41	0	3	0	9	0
31	36	–1	–2	1	4	2
36	32	4	–6	16	36	–24
29	31	–3	–7	9	49	21
38	30	6	-8	36	64	–48
34	33	2	–5	4	25	–10
32	39	0	1	0	1	0
Σx = 320	Σy = 380	$\sum(x-\bar{x})$ =0	$\sum(y-\bar{y})$ =0	$\sum(x-\bar{x})^2$ =140	$\sum(y-\bar{y})^2$ =398	$\sum(x-\bar{x})(y-\bar{y})$ =–93

$$b_{xy} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(y-\bar{y})^2} = \frac{-93}{398} = -0.23 \qquad \bar{x} = \frac{\sum x}{n} = \frac{320}{10} = 32$$

$$b_{yx} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2} = \frac{-93}{140} = -0.664 \qquad \bar{y} = \frac{\sum y}{n} = \frac{380}{10} = 38$$

The regression equation of x on y is

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

$$x - 32 = -0.23\,(y-38)$$

$$x = -0.23y + 8.74 + 32$$

$$x = -0.23y + 40.74$$

The regression equation of y on x is,

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

$$y - 38 = -0.66\,(x-32)$$

$$Y\text{-}38 = -0.66x + 21.12$$

$$y = -0.66x + 59.12 \qquad \ldots(1)$$

The coefficient of correlation is,

$r = \pm\sqrt{b_{yx}\, b_{xy}}$

$= \pm\sqrt{(-0.66)(-0.23)}$

$r = \pm 0.38$

The most likely marks in statistics where marks in economics is 30,

Put $x = 30$ in (1)

$y = -0.66(30) + 59.12 = 39.34 = 39$

Example 3

Obtain the lines of regression and the coefficient of correlation from the following data.

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

Solution

x	y	$x-\bar{x}$ x-4	$y-y$ y-11	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
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
Σx = 28	Σy = 770	$\sum(x-\bar{x})$ =0	$\sum(y-\bar{y})$ =0	$\sum(x-\bar{x})^2$ =28	$\sum(y-\bar{y})^2$ =28	$\sum(x-\bar{x})(y-\bar{y})$ =26

The regression coefficients x on y and y on x are,

$$b_{xy} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma\left(y-\bar{y}\right)^2} = \frac{26}{28} = 0.9285$$

$$b_{yx} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma\left(x-\bar{x}\right)^2} = \frac{26}{28} = 0.9285$$

The regression equation of x on y is

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

$x - 4 = 0.9285\,(y - 11)$

$x = 0.9285\,y - 10.21 + 4$

$x = 0.9285\,y - 6.21$

The regression equation of y on x is,

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

$y - 11 = 0.9285\,(x - 4)$

$y = 0.9285\,x - 3.714 + 11$

$y = 0.9285\,x - 7.286$ ...(1)

The coefficient of correlation is,

$r = \pm\sqrt{b_{yx}\,b_{xy}}$

$= \pm\sqrt{(0.9285)(0.9285)} = \pm 0.9285$

Example 4

The two regression equation of the variables x and y are x = 19.13 – 0.87y, y = 11.64 – 0.50x (i) Find the mean of x (ii) Find the mean of y (iii) Correlation coefficient between x and y. ***(May- 2007)***

Solution

Given that the two regression equations are

$x = 19.13 - 0.87y$ ...(1)

$y = 11.64 - 0.50x$ ...(2)

From (1) $b_{xy} = -0.87$

From (2) $b_{yx} = -0.50$

The correlation coefficient is,

$r = \pm\sqrt{b_{xy}\,b_{yx}}$

$= \pm\sqrt{(-0.87)(-0.50)} = \pm 0.659$

Let $\bar{x}$ and $\bar{y}$ satisfy (1) and (2)

$\bar{x} = 19.13 - 0.87\,\bar{y}$

$\bar{y} = 11.64 - 0.32\,\bar{x}$

$\bar{x} + 0.87\,\bar{y} = 19.13$ ...(3)

$0.5\,\bar{x} + \bar{y} = 11.64$ ...(4)

Eqn. (3) $\times$ 0.5 $\Rightarrow$ $0.5\bar{x} + 0.435\bar{y} = 9.565$ ...(5)

Equation (4) + (5) gives

$$\bar{y} = \frac{2.075}{0.565} = 3.67$$

Put $\bar{y}$ in (3)

$$\bar{x} + 0.87\,(3.67) = 19.13$$

$$\bar{x} = 15.93$$

Example 5

In a partially destroyed laboratory record of an analysis of a correlation data the following results only are eligible. Variance of x = 9, regression equations 8x – 10y + 66 = 0, 40x – 18y = 214,
Find (i) the means of x and y
(ii) The standard deviation of y and
(iii) The coefficient between x and y.

Solution

(i) Since both regression lines 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)

Eqn. (1) $\times \Rightarrow$ $40\bar{x} - 50\bar{y} = -330$ ...(3)

Equation (2) + (3) gives

$-32\bar{y} = -544$

$\bar{y} = 17$

Sub $\bar{y}$ in (1)

$8\bar{x} - 10(17) = -66$

$\bar{x} = 13$

$\therefore\ \bar{x} = 13,\ \bar{y} = 17$

(ii) Given Variance of $x = 9$

$\sigma_x = 3$

Consider $8x - 10y + 66 = 0$

$10y = 8x + 66$

$$y = \frac{8}{10}x + \frac{66}{10}$$

$$\Rightarrow byx = \frac{8}{10}$$

Consider $40x - 18y = 214$

$$40x = 214 + 18y$$

$$x = \frac{18}{40}y + \frac{214}{40}$$

$$\Rightarrow bxy = \frac{18}{40}$$

We know that

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

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

$$= \sqrt{\frac{18}{40} \times \frac{8}{10}} = 0.6$$

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

$$\frac{8}{10} = 0.6\frac{\sigma y}{\sigma x}$$

$$\sigma y = 4$$

$\therefore$ Mean of x $\sigma_x = 3$

Mean of y $\sigma_y = 4$

Coefficient of correlation = 0.6

Exercise

1. The following data pertains to length of service x is years and the annual income y (in ten thousands of rupees) for a sample of 10 employees of industry.

x	6	8	9	10	11	12	14	16	18	20
y	14	17	15	18	16	22	26	25	30	34

Compute the correlation coefficient between the length of service and the annual income. Also obtain a line of regression of y on x.

(Ans. x-0.631y =-1.882; 1.436x-y = -4.768; r = ± 0.952)

2. Obtain the two regression equations and the correlation coefficient.

x	68	64	69	65	66	70

y	81	85	80	88	87	82

(Ans. x+0.564y=114.376; 1.107x+y=158.169; r = ± 0.7902)

3. Obtain the two regression equations and the correlation coefficient.

x	78	89	97	69	59	79	68	57
y	125	137	156	112	107	138	123	108

(Ans. x-0.774y=-22.524; 1.162x-y=-38.85; r =0.948)

4. Obtain the two regression lines from the following data. Also compute the correlation coefficient.

x	22	26	29	30	31	31	34	35
y	20	20	21	29	27	24	27	31

(Ans. x-0.548y=16.848; 0.411x-y=-11.67; r = ± 0.475)

5. The regression equation of two variables x and y are $x = 0.7y+5.2$; $y = 0.3x+2.8$. Find the means of the variables and coefficient of correlation between them.

(*Ans*. $r = \pm 4.583; \overline{X} = -1.24; \overline{Y} = -0.92$)

Partial and Multiple Correlation and Regression

INTRODUCTION

The simple correlation and regression analysis discussed earlier measure the degree and nature of the effect of one variable on another. While it is useful to know how one phenomenon is influenced by another, it is also important to know how one phenomenon is affected by several other variables. One variable is related to a number of other variables, many of which may be interrelated among themselves. For example, yield of rice is affected by the type of soil, temperature, amount of rainfall, etc. It is part of the statistician's task to determine the effect of one case when the effect of others is estimated. This is done with the help of multiple and partial correlation analysis. Thus, it shall be possible for us to compare the relative importance of television advertisement and newspaper advertisement on increasing sales.

The basic distinction between multiple and partial correlation analysis is that whereas in the former, we measure the degree of association between the variable Y and all the variables, $X_1, X_2, X_3, \ldots X_n$, *taken together*; in the latter we measure the degree of association between Y and one of the variables $X_1, X_2, X_3, \ldots X_n$, *with the effect of all the other variables removed*. It should be noted that when only two variables are included in a study, the dependent variable is usually designated by Y, and the independent variable by X. However, when more than one independent variable is used it becomes advantageous to distinguish between the variables by means of subscripts and use only the letter X. The dependent variable is generally denoted by X_1 and the independent variables by X_2, X_3, etc. This scheme of notation can be expanded to take care of any number of independent variables.

PARTIAL CORRELATION

Partial Correlation Coefficients

Partial correlation coefficients provide a measure of the relationship between the dependent variable and other variables, with the effect of the rest of the variables eliminated.

If we denote by $r_{12.3}$, the coefficient of partial correlation between X_1 and X_2 keeping X_3 constant, we find that

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

Similarly,

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{23}^2}}$$

where, $r_{13.2}$ is the coefficient of partial correlation between X_1 and X_3 keeping X_2 constant.

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{13}^2}}$$

where $r_{23.1}$ is the coefficient of partial correlation between X_2 and X_3 keeping X_1 constant. Thus, for three variables X_1, X_2 and X_3, there will be three coefficients of partial correlation each studying the relationship between two variables when the third is held constant. It should be noted that the squares of partial correlation coefficients are called *coefficients of partial determination.*

Illustration 1. In a trivariate distribution it is found that

$r_{12} = 0.7,\ r_{13} = 0.61, r_{23} = 0.4$

Find the values of $r_{23.1}, r_{13.2}$ and $r_{12.3}$

Solution.

$$r_{23.1} = \frac{r_{23} - r_{12}\, r_{13}}{\sqrt{1-r_{12}^2}\ \sqrt{1-r_{13}^2}}$$

Substituting the given values

$$r_{23.1} = \frac{0.4 - 0.7 \times 0.61}{\sqrt{1-(0.7)^2}\ \sqrt{1-(0.61)^2}} = \frac{0.4 - 0.427}{\sqrt{0.51}\ \sqrt{1-0.3721}} = 0.048$$

$$r_{13.2} = \frac{r_{13} - r_{12}\, r_{23}}{\sqrt{1-r_{12}^2}\ \sqrt{1-r_{23}^2}}$$

$$= \frac{0.61 - 0.7 \times 0.4}{\sqrt{1-(0.7)^2}\ \sqrt{1-(0.4)^2}} = \frac{0.61 - 0.28}{\sqrt{1-0.49}\ \sqrt{1-0.16}} = 0.504$$

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{1-r_{13}^2}\ \sqrt{1-r_{23}^2}} = \frac{0.7 - 0.6 \times 0.4}{\sqrt{1-(0.61)^2}\ \sqrt{1-(0.4)^2}} = 0.633$$

Illustration 2. On the basis of observation made on agricultural production (X_1) the use of fertilizers (X_2) and the use of irrigation (X_3), the following zero order correlation coefficients were obtained :

$$r_{12} = 0.8, r_{13} = 0.65, r_{23} = 0.7$$

Compute the partial correlation between agricultural production and the use of fertilizers eliminating the effect of irrigation.

Solution. We have to calculate the value of $r_{12.3}$

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{1-r_{13}^2}\ \sqrt{1-r_{23}^2}}$$

Substituting the values $r_{12.3} = 0.8, r_{13} = 0.65$ and $r_{23} = 0.7$

$$r_{12.3} = \frac{0.8 - (0.65 \times 0.7)}{\sqrt{1-(0.65)^2}\ \sqrt{1-(0.7)^2}} = \frac{0.8 - 0.455}{\sqrt{1-0.4225}\ \sqrt{1-0.49}} = 0.636.$$

Illustration 3. Is it possible to get the following from a set of experimental data.

(*a*) $r_{23} = 0.8, r_{31} = -0.5, r_{12} = 0.6$ (*b*) $r_{23} = 0.7, r_{31} = -0.4, r_{12} = 0.6$.

Solution. In order to see whether there is any inconsistency, we should calculate $r_{12.3}$. If its value exceeds one, there is inconsistency otherwise not.

(*a*)
$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}} = \frac{0.6-(-0.5)(0.8)}{\sqrt{1-(-0.5)^2}\sqrt{1-(0.8)^2}}$$
$$= \frac{0.6+0.4}{\sqrt{0.75}\sqrt{0.36}} = \frac{1}{0.52} = 1.92$$

Since the value of $r_{12.3}$ is greater than one, therefore, there is some inconsistency in the given data.

(*b*)
$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}$$
$$= \frac{0.6-(-0.4)(0.7)}{\sqrt{1-(0.4)^2}\sqrt{1-(0.7)^2}} = \frac{0.6+0.28}{\sqrt{0.84}\sqrt{0.51}} = \frac{0.88}{0.65} = 1.35$$

This again is greater than one, therefore, there is some inconsistency in the given data.

Partial Correlation Coefficients in more than three variables

When four variables are involved in a correlation problem, there are twelve possible first-order coefficients. Some of these are :

$$r_{14.2} = \frac{r_{14} - r_{12}\, r_{24}}{\sqrt{1 - r_{12}^2}\ \sqrt{1 - r_{24}^2}}$$

$$r_{14.3} = \frac{r_{14} - r_{13}\, r_{34}}{\sqrt{1 - r_{13}^2}\ \sqrt{1 - r_{34}^2}}$$

$$r_{13.4} = \frac{r_{13} - r_{14}\, r_{34}}{\sqrt{1 - r_{14}^2}\ \sqrt{1 - r_{34}^2}}$$

$$r_{12.4} = \frac{r_{12} - r_{14}\, r_{24}}{\sqrt{1 - r_{14}^2}\ \sqrt{1 - r_{24}^2}}$$

$$r_{24.3} = \frac{r_{24} - r_{23}\, r_{34}}{\sqrt{1 - r_{23}^2}\ \sqrt{1 - r_{34}^2}}$$

$$r_{34.2} = \frac{r_{34} - r_{23}\, r_{24}}{\sqrt{1 - r_{23}^2}\ \sqrt{1 - r_{24}^2}}$$

$$r_{23.4} = \frac{r_{23} - r_{24}\, r_{34}}{\sqrt{1 - r_{24}^2}\ \sqrt{1 - r_{34}^2}}.$$

Similarly, the formulae for other partial correlation coefficients, *i.e.*, $r_{12.3}$, $r_{13.2}$, $r_{23.1}$, $r_{24.1}$, $r_{34.1}$, can also be written.

Second-order Partial Correlation Coefficients

Second-order coefficients may be obtained from order coefficients. In case of four variables, if $r_{12.34}$ is the coefficient of partial correlation between X_1 and X_2 keeping X_3 and X_4 constant, then

$$r_{12.34} = \frac{r_{12.4} - r_{13.4}\, r_{23.4}}{\sqrt{1 - r^2_{13.4}}\ \sqrt{1 - r^2_{23.4}}}.$$

Similarly,
$$r_{13.24} = \frac{r_{13.4} - r_{12.4}\, r_{23.4}}{\sqrt{1 - r^2_{12.4}}\ \sqrt{1 - r^2_{23.4}}}$$

and
$$r_{14.23} = \frac{r_{14.3} - r_{12.3}\, r_{24.3}}{\sqrt{1 - r^2_{12.3}}\ \sqrt{1 - r^2_{24.3}}}.$$

Alternative formulae giving the same results are available for all three of the second-order coefficients. They are :

$$r_{12.34} = \frac{r_{12.3} - r_{14.3}\, r_{24.3}}{\sqrt{1 - r^2_{14.3}}\ \sqrt{1 - r^2_{24.3}}}$$

$$r_{13.24} = \frac{r_{13.2} - r_{14.2}\, r_{34.2}}{\sqrt{1 - r^2_{14.2}}\ \sqrt{1 - r^2_{34.2}}}$$

$$r_{14.23} = \frac{r_{14.2} - r_{13.2}\, r_{34.2}}{\sqrt{1 - r^2_{13.2}}\ \sqrt{1 - r^2_{34.2}}}.$$

The value of a partial correlation coefficient is usually interpreted via the corresponding coefficient of partial determination, which is merely the square of the former. Thus, if $r_{12.3} = 0.4$, $r^2_{12.3} = 0.16$.

freedom is reduced by the number of

Characteristics and Uses of Partial Correlation Analysis. The purpose of partial correlation analysis is the measurement of relationship between two factors, with the effects of one or more other factors eliminated. If the assumptions of the method are true for a series of data, the power of partial analysis is great. The problem of holding certain variables constants while the relationship between the other is measured often presents itself in statistical analysis. Partial correlation is especially useful in the analysis of interrelated series. It is particularly pertinent to uncontrolled experiments of various kinds in which, such interrelationship usually exists. Most economic data fall in this category.

Partial correlation is of greatest value when used in conjunction with gross and multiple correlation in the analysis of factors affecting variations in many kinds of phenomena.

It has the advantage that the relationships are expressed concisely in a few well-defined coefficients. Also it is adaptable to small amounts of data and the reliability of the results can be rather easily tested.

Limitations of Partial Correlation Analysis. 1. The usefulness of the partial analysis is somewhat limited by the following basic assumptions of the method :

(*i*) The gross or zero-order correlation must have linear regressions.

(*ii*) The effects of the independent variable must be additively and not jointly related.

(*iii*) Because the reliability of partial coefficients decreases as its order increases. The number of observations in gross correlations should be fairly large. Often the student carries the analysis beyond the limits of the data. Thus, weakness to some extent can be guarded against by test of reliability.

2. When the above assumptions have been satisfied, partial correlation analysis still possess the disadvantages of laborious calculations and difficult interpretation even for statisticians.

The interpretation of the partial and multiple correlation results tends to assume that the independent variable have casual effects on dependent variable. The assumption is sometimes true, but more often untrue in varying degrees.

MULTIPLE CORRELATION

In problems of multiple correlation, we are dealing with situations that involve three or more variables. For example, we may consider the association between the yield of wheat per acre and both the amount of rainfall and the average daily temperature. We are trying to make estimates of the value of one of these variables based on the values of all the others. The variable whose value we are trying to estimate is called the dependent variable and the other variables on which our estimates are based are known as independent variables. The statistician himself chooses which variable is to be dependent and which variables are to be independent. It is merely a question of problem being studied. If we are trying to determine the most probable weight of men, we make weight, the dependent variable and height, age, etc., independent variables. If on the other hand, we are interested in estimating height, we will make height the dependent variable and weight, age, etc., the independent variables. Thus in problems of multiple correlation, we always have three or more variables (one dependent and the rest independent). In order that we may distinguish them easily, we follow the custom of representing them by the letter X with subscript. The dependent variable is always denoted by X_1 and the others by X_2, X_3, etc. Thus in the height, age and weight problem, if we are trying to estimate men's weight (that is, if weight be dependent variable), we might denote

$X_1 \rightarrow$ weight in lbs.

$X_2 \rightarrow$ height in inches.

$X_3 \rightarrow$ age in years.

The multiple correlation is of great practical significance—for rarely is it ever true that a variable is influenced solely or predominantly by one other factor. For example, the sales of a manufacturer are influenced, among other things, by his prices, his competitive position in the industry, his sales promotion campaign, industry sales, competitors' prices and national prosperity. In simple correlation, only one of the independent variables at a time could be correlated with the manufact ırer's sales and there is no direct way of determining the extent to which the observed correlation might have been caused by the interacting influence of other factors on the two variables under study. For instance, in times of prosperity a high level of national income may lead to increased industry sales, a share of which is captured by this manufacturer. But to what extent are the manufacturer's sales influenced by the universally buoyant affect of national prosperity and to what extent are his sales affected by the particular trend of the industry sales within the economy, *i.e.*, assuming that the nation's economy remain fairly stable ? To answer questions of this type which are of vital importance in framing suitable managerial policies, one has to depend on multiple correlation analysis.

Coefficient of Multiple Correlation

The coefficient of multiple linear correlation* is represented by R_1 and it is common to add subscript designating the variables involved. Thus $R_{1.234}$ would represent the coefficient of multiple linear correlation between X_1 on the one hand, and X_2, X_3 and X_4 on the other. The subscript of the dependent variable is always to the left of the point.

The coefficient of multiple correlation can be expressed in terms of r_{12}, r_{13} and r_{23} as follows :

$$R_{1.23} = \sqrt{\frac{r^2_{12} + r^2_{13} - 2r_{12}\, r_{13}\, r_{23}}{1 - r^2_{23}}}$$

$$R_{2.13} = \sqrt{\frac{r^2_{21} + r^2_{23} - 2r_{12}\, r_{13}\, r_{23}}{1 - r^2_{13}}}$$

and

$$R_{3.12} = \sqrt{\frac{r^2_{13} + r^2_{23} - 2r_{12}\, r_{13}\, r_{23}}{1 - r^2_{12}}}$$

A coefficient of multiple correlation such as $R_{1.23}$ lies between 0 and 1. The closer it is to 1, the better is the linear relationship between the variables. The closer it is to 0, the worse is the linear relationship. If the coefficient of multiple correlation is 1, the correlation is called *perfect*. Although a correlation coefficient of 0 indicates no linear relationship between the variables, it is possible that a nonlinear relationship may exist. It should be noted that whereas the simple correlation coefficients range from + 1.0 to 0 to − 1.0, the coefficients of multiple correlation are always positive in sign and range from + 1.0 to 0.

An alternative formula for calculating $R_{1.23}$ is :

$$R_{1.23} = \sqrt{r^2_{12} + r_{13.2}\,(1 - r^2_{12})}.$$

similarly, $$R_{1.24} = \sqrt{\frac{r^2_{12} + r^2_{14} - 2r_{12}\, r_{14}\, r_{24}}{1 - r^2_{24}}}$$

or $$R_{1.24} = \sqrt{r^2_{12} + r^2_{14.2}\,(1 - r^2_{12})}$$

and $$R_{1.34} = \sqrt{\frac{r^2_{13} + r^2_{14} - 2r_{13}\, r_{14}\, r_{34}}{1 - r^2_{34}}}$$

or $$R_{1.34} = \sqrt{r^2_{13} + r^2_{14.3}\,(1 - r^2_{13})}.$$

Coefficient of Multiple Determination

In chapter on correlation, we talked of coefficient of determination r^2 which measures the fit of a straight line to the two-variable scatter. In exactly the same way, the coefficient of multiple determination denoted by $R^2_{1.23}$ is also defined. Thus, $R^2_{1.23}$ may be thought of as a measure of closeness of fit of the regression plane to the actual points relative to the point of the means of the variable. Or, just as does r^2, $R^2_{1.23}$ measures the percentage of total error that is accounted for by the regression. Obviously, the greater the value of $R^2_{1.23}$, the smaller is the scatter and the better is the fit. Thus, if coefficient of multiple determination between yield of rice (X_1) and fertilizers (X_2) and rain (X_3) is 0.953, it means that 95.3 per cent of the variations in yield have been explained by the variation in fertilizers and rain. There remains only 4.7 per cent of the variations in yield of rice that can be explained only by factors which have not been taken into consideration in our analysis.

Illustration. 4. The following zero-order, correlation coefficients are given

$r_{12} = 0.98$, $r_{13} = 0.44$ and $r_{23} = 0.54$.

Calculate multiple correlation coefficient treating first variable as dependent and second and third variables as independent.

Solution. We have to calculate the multiple correlation coefficient treating first variable as dependent and second and third variables as independent, *i.e.*, we have to find $R_{1.23}$

$$R_{1.23} = \sqrt{\frac{r^2_{12} + r^2_{13} - 2r_{12}\, r_{13}\, r_{23}}{1 - r^2_{23}}}$$

Substituting the given values

$$R_{1.23} = \sqrt{\frac{(0.98)^2 + (0.44)^2 - 2(0.98)\,(0.44)\,(0.54)}{1 - (0.54)^2}}$$

$$= \sqrt{\frac{0.9604 + 0.1936 - 0.4657}{0.7084}} = +0.986.$$

Advantages of Multiple Correlation Analysis. The coefficient of multiple correlation serves the following purposes :

1. It serves as a measure of the degree of association between one variable taken as the dependent variable and a group of other variables taken as the independent variables.

2. Hence it also serves as a measure of goodness of fit of the calculated plane of regression and consequently, as a measure of the general degree of accuracy of estimates made by reference to equation for the plane or regression.

Limitations of Multiple Correlation Analysis. 1. Multiple correlation analysis is based on the assumption that the relationship between the variables is linear. In other words, the rate of change in one variable in terms of another is assumed to be constant for all values. In practice, most relationship are not linear but follow some other pattern. This limits somewhat the use of multiple correlation analysis. The linear regression coefficients are not accurately descriptive of curvilinear data.

2. A second important limitation is the assumption that effects of independent variables on the dependent variables are separate, distinct and additive. When the effects of variables are additive, a given change in one has the same effect on the dependent variable regardless of the sizes of the other two independent variables.

3. Linear multiple correlation involves a great deal of work relative to the results frequently obtained. When the results are obtained, only a few students well trained in the method are able to interest them. The misuse of correlation results has probably cast more doubt on the method than is justified. However, this lack of understanding and resulting misuses are due to the complexity of the method.

Multiple Regression

In the simple linear regression model discussed earlier, we talked of one dependent variable and one independent variable.

In multiple regression analysis which is a logical extension of two-variable regression analysis, instead of a single independent variable, two or more independent variables are used to estimate the values of a dependent variable. However, the fundamental concepts in the analysis remain the same. The multiple regression and correlation analysis serves highly useful purpose in practice. Its main objectives are :

(*a*) To derive an equation which provides estimates of the dependent variable from values of the two or more independent variables.

(*b*) To obtain a measure of the error involved in using this regression equation as a basis for estimation.

(*c*) To obtain a measure of the proportion of variance in the dependent variable accounted for or "explained by" the independent variables.

The first purpose is accomplished by deriving an appropriate regression equation by the method of least squares. The second purpose is achieved through the calculation of standard error of estimate and the third purpose is accomplished by computing the multiple coefficient of determination.

The multiple regression equation involving two independent variables shall take the form

$$Y_c = a + b_1X_1 + b_2X_2$$

The general form of the linear multiple regression function for k independent variables $X_1, X_2, \ldots\ldots, X_k$, is

$$Y_c = a + b_1X_1 + b_2X_2 + \ldots\ldots + b_kX_k$$

The linear function which is fitted to data for two variables is referred to as a *straight line*, for three variables a *plane*, for four or more variables a *hyperplane*.

If we have three variables X_1, X_2 and X_3, the multiple linear regression equation of X_1 on X_2 and X_3 shall have the following form :

$$X_1 = a_{1.23} + b_{12.3}X_2 + b_{13.2}X_3$$

Obviously X_1 is the dependent variable here and X_2 and X_3 are independent variables. The constant $a_{1.23}$ is the intercept made by the regression plane; it is zero when the regression line passes through the origin. The regression coefficients denoted by $b_{12.3}$ and $b_{13.2}$ represent the rate of change of the dependent variable per unit change in each of the independent variables when the other independent variables are held constant. The first subscript always represents the dependent variable and the second subscript denotes the particular independent variable being related to X_1. The subscripts after the period indicate the other independent variables, all of which are held constant while the effect of the particular independent variable on X_1 is measured. Thus, $b_{12.3}$ measures the amount by which a unit change in X_2 is expected to affect X_1 when X_3 is held constant and $b_{12.3}$ measures the amount of change in X_1 when X_3 is held constant and $b_{13.2}$ measures the amount of change in X_1 per unit change in X_3 when X_2 is held constant. Similarly, $b_{13.24}$ would represent the change in X_1 per unit change in X_3 when the values of X_2 and X_4 are held constant.

The regression coefficients, *i.e.*, b's in multiple linear regression are termed *coefficients of net regression* : the regression is *net* in the sense that the regression of the dependent variable on the particular independent variable is measured while holding the values of the other independent variables constant. It contrast, the coefficients in simple regression are called *coefficients of gross regression* because no allowance is made for indirect influences on the regression.

The follwing are the usual assumptions made in a linear multiple regression analysis illustrated for the case of two independent variables :

1. The conditional distributions of Y for given X_1 and X_2 are assumed to be normal.
2. These conditional distributions are assumed to have equal standard deviations.
3. The $Y - Y_c$ deviation are assumed to be independent of one another.

Normal Equations for the Least Square Regression Plane

Just as there exist, least square regression line approximating a set of N data points (X, Y) in a two-dimensional scatter diagram, so also there exist *least square regression* planes fitting a set of N data points (X_1, X_2, X_3) in a three-dimensional by scatter diagram.

The least square regression plane of X_1 on X_2 and X_3 has the equation (i) where $b_{12.3}$ and $b_{13.2}$ are determined by solving simultaneously, the *normal equations*.

$$\Sigma X_1 = Na_{1.23} + b_{12.3}\,\Sigma X_2 + b_{13.2}\Sigma X_3$$
$$\Sigma X_1X_2 = a_{1.23}\,\Sigma X_2 + b_{12.3}\,\Sigma X_2^{\,2} + b_{13.2}X_2X_3$$
$$\Sigma X_1X_3 = a_{1.23}\,\Sigma X_3 + b_{12.3}\,\Sigma X_2X_3 + b_{13.2}X_3^{\,2}.$$

These can be obtained formally by multiplying both sides of equation (i) by 1, X_2 and X_3 successively and summing on both sides.

When the number of variables is 4 or more, solving the above system of normal equation becomes a very tedious procedure. Efficient methods solving simultaneous equations require a knowledge of matrix algebra, which is not assumed for the reader of the text. Thus in our discussion that follows, we shall confine ourselves to the two independent variables cases, which of course, can be extended to cover cases with three or more independent variables.

The work involved in finding these regression equations can be reduced by proceeding in terms of deviations from the mean of the variables under consideration. The regression equation for these variables in this procedure is :

$$x_1 = b_{12.3}\,x_2 + b_{13.2}\,x_3$$

where $x_1 = (X_1 - \overline{X}_1),\ x_2 = (X_2 - \overline{X}_2),\ x_3 = (X_3 - \overline{X}_3)$.

The value $b_{12.3}$ and $b_{13.2}$ can be obtained by solving simultaneously the following two normal equations :

$$\Sigma x_1x_2 = b_{12.3}\,\Sigma x_2^{\,2} + b_{13.2}\,\Sigma x_2x_3$$
$$\Sigma x_1x_3 = b_{12.3}\,\Sigma x_2x_3 + b_{13.2}\Sigma x_3^{\,2}.$$

The value of $b_{12.3}$ and $b_{13.2}$ can also be obtained as follows :

$$b_{12.3} = r_{12.3}\,\frac{\sigma_{1.23}}{\sigma_{2.13}}$$

$$b_{13.2} = r_{13.2}\frac{\sigma_{13.2}}{\sigma_{3.12}}$$

The regression equation of X_1 on X_2 and X_3 can be expressed as follows :

$$X_1 - \bar{X}_1 = \left(\frac{r_{12} - r_{13}r_{23}}{1 - r^2_{23}}\right)\left(\frac{S_1}{S_2}\right)(X_2 - \bar{X}_2) + \left(\frac{r_{13} - r_{12}r_{23}}{1 - r^2_{23}}\right)\left(\frac{S_1}{S_3}\right)(X_3 - \bar{X}_3) \quad ...(i)$$

The regression equation of X_3 on X_2 and X_1 can be written as follows :

$$X_1 - \bar{X}_2 = \left(\frac{r_{23} - r_{13}r_{12}}{1 - r^2_{12}}\right)\left(\frac{S_1}{S_2}\right)(X_2 - \bar{X}_2) + \left(\frac{r_{13} - r_{12}r_{23}}{1 - r^2_{12}}\right)\left(\frac{S_3}{S_1}\right)(X_1 - \bar{X}_1) \quad ...(ii)$$

From (*i*) and (*ii*), the coefficients of X_3 and X_1 are respectively

$$b_{13.2} = \left(\frac{r_{13} - r_{12}r_{23}}{1 - r^2_{23}}\right)\left(\frac{S_1}{S_2}\right) \text{ and } \left(\frac{r_{13} - r_{23}r_{12}}{1 - r^2_{12}}\right)\left(\frac{S_2}{S_1}\right)$$

$$b_{13.2}\, b_{31.2} = \frac{(r_{13} - r_{12}r_{23})^2}{(1 - r^2_{23})(1 - r^2_{12})} = r^2_{13.2}$$

This method of obtaining regression equations is much simpler compared to one where simultaneously several normal equations are to be solved. For calculating regression equation for three variables, when the above procedure is used we need the following :

X_1	X_2	X_3
S_1	S_2	S_3
r_{12}	r_{13}	r_{23}

Other Equations of Multiple linear Regression

In the case of two variables, there were two equations of regression—one of them indicating regression of Y on X, and the other, that of X on Y. When there are three variables, there will be three equations of regression indicating the regression of X_1 on X_2 and X_3, the other indicating the regression of X_2 on X_1 and X_3 and the third indicating the regression of X_3 on X_1 and X_2. The first of these has been given earlier. If X_2 and X_3 were to be treated as dependent variables the regression equation will respectively be :

$$X_2 = a_{2.13} + b_{21.3}X_1 + b_{23.1}X_3 \quad ...(ii)$$

$$X_3 = a_{3.12} + b_{31.2}X_1 + b_{32.1}X_2 \quad ...(iii)$$

The normal equations for fitting (*ii*) will be :

$$\Sigma X_2 = Na_{2.13} + b_{21.3}\Sigma X_1 + b_{23.1}\Sigma X_3$$

$$\Sigma X_1X_2 = a_{2.13}\Sigma X_1 + b_{21.3}\Sigma X_1^2 + b_{23.1}\Sigma X_1X_2$$

$$\Sigma X_2X_3 = a_{2.13}\Sigma X_3 + b_{21.3}\Sigma X_1X_3 + b_{23.1}\Sigma X_3^2$$

In case we want to fit equation (*iii*), the normal equations will be :

$$\Sigma X_3 = Na_{3.12} + b_{31.2}\Sigma X_1 + b_{32.1}\Sigma X_2$$

$$\Sigma X_1X_3 = a_{3.12}\Sigma X_1 + b_{31.2}\Sigma X_1^2 + b_{32.1}\Sigma X_1X_2$$

$$\Sigma X_2X_3 = a_{3.12}\Sigma X_2 + b_{31.2}\Sigma X_1X_2 + b_{32.1}\Sigma X_2^2$$

Relationship between Partial and Multiple Correlation Coefficients

Interesting results connecting the multiple correlation coefficients and the various partial correlation coefficients can be found. For example, we find :

$$1-R^2_{1.23} = (1-r^2_{12})\,(1-r^2_{13.2})$$

$$1-R^2_{1.234} = (1-r^2_{13})\,(1-r^2_{13.2})\,(1-r^2_{14.23})$$

Illustration 5. Find multiple linear regression equation of X_1 on X_2 and X_3 from the data relating to three variables given below :

X_1 :	4	6	7	9	13	15
X_2 :	15	12	8	6	4	8
X_3 :	30	24	20	14	10	4

Solution. The regression equation of X_1 on X_2 and X_3 is

$$X_1 = a_{1.23} + b_{12.3}\,X_2 + b_{13.2}\,X_3$$

The value of the constants $a_{1.23}$, $b_{12.3}$ and $b_{13.2}$ are obtained by solving the following three normal equations :

$$\Sigma X_1 = Na_{1.23} + b_{12.3}\,\Sigma X_2 + b_{13.2}\,\Sigma X_3$$

$$\Sigma X_1 X_2 = a_{1.23}\,\Sigma X_2 + b_{12.3}\,\Sigma X_2^2 + b_{13.2}\,\Sigma X_2 X_3$$

$$\Sigma X_1 X_3 = a_{1.23}\,\Sigma X_3 + b_{12.3}\,\Sigma X_2 X_3 + b_{13.2}\,\Sigma X_3^2$$

Calculating the required values :

X_1	X_2	X_3	X_1X_2	X_1X_3	X_2X_3	X_2^2	X_3^2	X_1^2
4	15	30	60	120	450	225	900	16
6	12	24	72	144	288	144	576	36
7	8	20	56	140	160	64	400	49
9	6	14	54	126	84	36	196	81
13	4	10	52	130	40	16	100	169
15	3	4	45	60	12	9	16	225
$\Sigma X_1 = 54$	$\Sigma X_2 = 48$	$\Sigma X_3 = 102$	$\Sigma X_1X_2 = 339$	$\Sigma X_1X_3 = 720$	$\Sigma X_2X_3 = 1,034$	$\Sigma X_2^2 = 494$	$\Sigma X_3^2 = 2,188$	$\Sigma X_1^2 = 576$

Substituting the values in the normal equations

Substituting the values in the normal equations :

$$6a_{1.23} + 48b_{12.3} + 102b_{13.2} = 54 \quad \ldots (i)$$

$$48a_{1.23} + 49b_{12.3} + 1034b_{13.2} = 339 \quad \ldots (ii)$$

$$102a_{1.23} + 1034b_{12.3} + 2188b_{13.2} = 720 \quad \ldots (iii)$$

Multiplying Eqn. (*i*) by 8, we get

$$48a_{1.23} + 384b_{12.3} + 816b_{13.2} = 432 \quad \ldots (iv)$$

Subtracting Eqn. (*ii*) from (*iv*), we get

$$110b_{12.3} + 218b_{13.2} = -93 \quad \ldots (v)$$

Multiplying Eqn. (*i*) by 17, we get

$$102a_{1.23} + 816b_{12.3} + 1734b_{13.2} = 918 \quad \ldots (vi)$$

Subtracting Eqn. (*iii*) from Eqn. (*vi*) we get

$$218b_{12.3} + 454b_{13.2} = -198 \quad \ldots (vii)$$

Multiplying Eqn. (*v*) by 109, we obtain

$$11990b_{12.3} + 23762b_{13.2} = -10137 \quad \ldots (viii)$$

Multiplying Eqn. (*vii*) by 55, we get

$$11990b_{12.3} + 24970b_{13.2} = -10890 \quad \ldots (ix)$$

Subtracting Eqn. (*viii*) from Eqn. (*ix*), we get

$$1208b_{13.2} = -753$$

$$b_{13.2} = \frac{-753}{1208} = -0.623.$$

Substituting this value of $b_{13.2}$ in Eqn. (*v*), we get

$$110b_{12.3} + 218(-0.623) = -93$$

$$110b_{12.3} = 135.814 - 93$$

$$b_{12.3} = \frac{42.814}{110} = +0.389.$$

Substituting the value of $b_{12.3}$ and $b_{13.2}$ in Eqn. (*i*), we get

$$6a_{1.23} + 48(0.389) + 102(-0.623) = 54$$

$$6a_{1.23} = 54 + 63.546 - 18.672 = 98.874$$

$$a_{1.23} = 16.479.$$

Thus, the required regression equation is :

$$X_1 = 16.479 + 0.389\,X_2 - 0.623\,X_3.$$

Illustration. 6. Given the following, determine the regression equation of :

(*i*) x_1 on x_2 and x_3, and (*ii*) x_2 on x_1 and x_3

$r_{12} = 0.8$ $r_{13} = 0.6$ $r_{23} = 0.5$

$\sigma_1 = 10$ $\sigma_2 = 8$ $\sigma_3 = 5$.

Solution. Regression equation of X_1 on X_2 and X_3 is given by

$$X_1 = a + b_{12.3} X_2 + b_{13.2} X_3.$$

If the variates X_1, X_2 and X_3 are measured as deviations from their respective means, 'a' will be zero. The values of $b_{12.3}$ and $b_{13.2}$ can be calculated from the data given above but not for 'a'. So, let us assume x_1 and x_2 represent deviations from means. So the regression equation of x_1 on x_2 and x_3 is :

$$x_1 = b_{12.3} x_2 + b_{13.2} x_3$$

$$b_{12.3} = \frac{\sigma_1}{\sigma_2} \times \frac{r_{12} - r_{13} r_{23}}{1 - r^2_{23}}$$

$$= \frac{10}{8} \times \frac{0.8 - (0.6)(0.5)}{1 - (0.5)^2} = 0.833.$$

$$b_{13.2} = \frac{\sigma_1}{\sigma_2} \times \frac{r_{13} - r_{13} r_{23}}{1 - r^2_{23}}$$

$$= \frac{10}{5} \times \frac{0.6 - (0.8)(0.5)}{1 - (0.5)^2} = 0.533.$$

∴ Required regression equation is

$$x_1 = 0.833\, x_2 + 0.533\, x_3.$$

(*ii*) Regression equation of x_2 on x_1 and x_3

$$x_2 = b_{21.3} x_1 + b_{23.1} x_3$$

$$b_{21.3} = \frac{\sigma_1}{\sigma_2} \times \frac{r_{12} - r_{23} r_{13}}{1 - r^2_{13}}$$

$$= \frac{8}{10} \times \frac{(0.8) - (0.5)(0.6)}{1 - (0.6)^2}$$

$$= \frac{8}{10} \times \frac{0.8 - 0.3}{1 - 0.36} = \frac{8}{10} \times \frac{05}{0.64} = 0.625.$$

$$b_{23.1} = \frac{\sigma_1}{\sigma_2} \times \frac{r_{23} - r_{12}\, r_{13}}{1 - r^2_{13}}$$

$$= \frac{8}{5} \times \frac{05 - (0.8)(0.6)}{1 - (0.6)^2} = \frac{8}{5} \times \frac{0.02}{0.64} = 0.05.$$

Thus $x_2 = 0.625\, x_1 + 0.05\, x_3$ is the required regression equation.

14. If the correlation coefficient taking the values 1,-1, 0, then what is the significance of r?

Solution

If r = 1, there is perfect direct correlation between the two variables.
If r = -1, there is perfect inverse correlation between the two variables.
If r = 0, we infer that there is no linear correlation.

15. Write some of the properties of the correlation coefficient. (November-2011)

Solution

(i) The correlation coefficient always lies between -1 to 1. (i.e.) $-1 \le r \le 1$

(ii) If two variables are independent, their correlation coefficient is zero.

(iii) The degree of relationship between two variables is symmetric. (i.e.) $r_{xy} = r_{yx}$

(iv) It bears the same sign as cov(X,Y).

13. What is correlation and how it is measured?

Solution

Correlation refers to statistical tool for measuring the degree of relationship between two or more variables. If a change in variable X produces a corresponding change in variable Y, then X and Y are said to be correlated. $dx = x - A$.

Two variables may be highly, slightly or moderately correlated. The measure of such correlation is called correlation coefficient usually denoted by "r"

The Karl Pearson's coefficients of correlation is defined by,

$$r = \frac{N\Sigma dxdy - \Sigma dx\Sigma dy}{\sqrt{N\Sigma dx^2 - (\Sigma dx)^2}\sqrt{N\Sigma dy^2 - (\Sigma dy)^2}}$$

$dx = x - A$ or $x - \bar{x}$

$dy = y - B$ or $y - \bar{y}$

16. Find the correlation coefficient between x and y of the data (2, 7), (11, 9).

Solution

x	y	x^2	y^2	xy
2	7	4	49	14
11	9	121	81	99
$\sum x=13$	$\sum y=16$	$\sum x^2=125$	$\sum y^2=130$	$\sum xy=113$

Coefficient of correlation= $r = \dfrac{N\Sigma dxdy - \Sigma dx\Sigma dy}{\sqrt{N\Sigma dx^2-(\Sigma dx)^2}\sqrt{N\Sigma dy^2-(\Sigma dy)^2}}$

$$= \frac{2(113)-(13)(16)}{\sqrt{2(125)-(13)^2}\sqrt{2(130)-(16)^2}} = \frac{18}{18} = 1$$

17. Find the correlation coefficient between x and y of the data (5, 13), (-7, 21).

Solution

X	y	x^2	y^2	xy
5	13	25	169	65
-7	21	49	441	-147
$\sum x$=-2	$\sum y=34$	$\sum x^2=74$	$\sum y^2=610$	$\sum xy$=-82

Coefficient of correlation $r = \dfrac{N\Sigma dxdy - \Sigma dx\Sigma dy}{\sqrt{N\Sigma dx^2-(\Sigma dx)^2}\sqrt{N\Sigma dy^2-(\Sigma dy)^2}}$

$$= \frac{-2(82)-(-2)(34)}{\sqrt{2(74)-(-2)^2}\sqrt{2(610)-(34)^2}} = -1$$

18. Define "rank correlation" ***(May-2011)***

Solution

The rank is a variate which takes only the values 1, 2,…, n. Let $(x_i, y_i), i=1,2,\ldots,n$, be the ranks of '$i$' individuals in two characteristics A and B respectively.

Pearson's coefficients of correlation between the ranks x_i's and y_i's is called the rank correlation coefficients between the two characteristic A and B for that group of individuals is given by, $\rho = 1 - \dfrac{6\Sigma d_i^2}{n(n^2-1)}$, where $d_i = x_i - y_i$

If the rank is repeated, then the rank correlation coefficient is,

$\rho = 1 - \dfrac{6\left[\Sigma d^2 + \dfrac{m(m^2-1)}{12}\right]}{n(n^2-1)}$ where m is the number of times a rank is repeated.

Here $\dfrac{m(m^2-1)}{12}$ is the correction factor to be calculated for each repeated ranks.

19. Find the rank correlation coefficient for the following data.

x	1	2	3	4
y	3	4	2	1

Solution

x	y	R_1	R_2	$d = R_1 - R_2$	d^2
1	3	4	2	2	4
2	4	3	1	2	4
3	2	2	3	-1	1
4	1	1	4	3	9
					$\sum d^2 = 18$

$$\rho = 1 - \frac{6\Sigma di^2}{n(n^2-1)} = 1 - \frac{6(18)}{4(17)} = -0.588$$

20. Find the rank correlation coefficient for the following data.

X	55	56	58	59
Y	35	39	38	42

Solution

x	y	R_1	R_2	$d = R_1 - R_2$	d^2
55	35	4	4	0	0
56	39	3	2	1	1
58	38	2	3	−1	1
59	42	1	1	0	0
					$\sum d^2 = 2$

$$\rho = 1 - \frac{6\Sigma di^2}{n(n^2-1)} = 1 - \frac{6(2)}{4(15)} = 0.8$$

21. Explain Regression.

Solution

Regression is the estimation or prediction of unknown values of one variable from known values of another variable. It is the mathematical measure of the average relationship between two or more variables in terms of the original limits of the data.

22. Write the two regression lines.

Solution

The line of regression x on y is $x - \bar{x} = b_{xy}(y - \bar{y})$

where $b_{xy} = r\dfrac{\sigma x}{\sigma y} = \dfrac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(y-\bar{y})^2}$

The line of regression y on x is $y - \bar{y} = b_{yx}(x - \bar{x})$

where $b_{yx} = r\dfrac{\sigma y}{\sigma x} = \dfrac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}$

23. What is the angle between two regression lines?

Solution

The angle θ between the two regression lines is given by $\tan\theta = \dfrac{1-r^2}{r}\dfrac{\sigma_x\,\sigma_y}{\sigma_x^{\,2}+\sigma_y^{\,2}}$

24. If the equation of regression line are $x + 2y = 5$ ***and*** $2x + 3y = 8$ ***then find the correlation coefficient between x and y.***

Solution

$$x + 2y = 5 \Rightarrow 2y = -x + 5$$

$$y = -\frac{1}{2}x + \frac{5}{2} \Rightarrow b_{yx} = -\frac{1}{2}$$

$$2x + 3y = 8 \Rightarrow 2x = -3y + 8 \Rightarrow b_{xy} = -\frac{3}{2}\ x = -\frac{3}{2}y + 4$$

$$r = \sqrt{b_{yx}b_{xy}} = \sqrt{\left(-\frac{1}{2}\right)\left(-\frac{3}{2}\right)} = \sqrt{\frac{3}{4}}$$

25. Can $2x + 3y = 4, x - y = 5$ ***be equations of valid regression lines?***

Solution

$2x+3y=4$ and $x-y=5$

$3y=-2x+4$ and $x=y+5$

$y=-\frac{2}{3}x+\frac{4}{3}$

$r=\sqrt{\left(-\frac{2}{3}\right)(1)}$ is not possible.

Or else, $2x=-3y+4$ and $x-y=5$

$x=-\frac{3}{2}y+2$ and $y=x-5$

$r=\sqrt{\left(-\frac{3}{2}\right)(1)}$ is not possible.

$\therefore$ The equations are not valid regression lines.

26. If the regression line of y on x is $y=2x-4$ ***and if the mean of x is 3. Find the mean of y.***

Solution

Given $\bar{x}=3$

The regression line passes through $(\bar{x},\bar{y})$

$\bar{y}=2\bar{x}-4 \Rightarrow \bar{y}=2(3)-4=2$

Mean of $y=2$

27. If -0.4 and -1.6 are the regression coefficient of X and Y, then the correlation coefficient between them is?

Solution

$r=\pm\sqrt{b_{xy}b_{yx}}=\sqrt{(-0.4)(-1.6)}=\sqrt{.64}=0.8$