Module III

Correlation And Regression

Chapter 1:

Curve fitting

Let x be an independent variable & y be a variable depending on x. Here, we say that y is a function of x & write it as $\boxed{y = f(x)}$.

If $f(x)$ is a known fun, then for any allowable values $x_1, x_2 \dots x_n$ of x, we can find the corresponding values $y_1, y_2 \dots y_n$ of y & there by determine the pairs $(x_1, y_1) (x_2, y_2) \dots (x_n, y_n)$ which constitute a bivariate data. These pairs of values of x & y give us n points on the curve $y = f(x)$.

Suppose consider the converse prbm. Suppose we are given n values $x_1, x_2 \dots x_n$ of an independent variable x & the corresponding values $y_1, y_2 \dots y_n$ of a variable y depending on x. Then the pairs $(x_1, y_1) (x_2, y_2) \dots (x_n, y_n)$ gives us n points in xy plane.

Generally it is not possible to find the

actual curve $y=f(x)$ that passes through these points.
Hence we try to find a curve that serve as best approximation to the curve $y=f(x)$. Such a curve is referred as <u>curve of bestfit</u>. The process of determining a curve of bestfit → <u>curve fitting</u>. The method for curve fitting is called method of <u>least squares</u>.

→ <u>Method of least squares :-</u>

This method is for finding the unknown coefficients in a curve that serves as best approximation to the curve $y=f(x)$.
By A.M Legendre ie C.F Gauss, the <u>principle of least squares</u> says that the sum of squares of the error b/w the observed values ie the corresponding estimated values should be the least.
Suppose to fit a k^{th} deg curve, given by $y=a_0+a_1x+a_2x^2 \dots a_kx^k$ — ① to the given pairs of observations (x_1,y_1) (x_2,y_2) ... (x_n,y_n). The curve has $k+1$ unknown const. ie hence if $n = k+1$.

we get $k+1$ eq on subset the values of (x_i,y_i) in ①
This gives unique soln to the values as $a_1 \dots a_n$
However if $n > k+1$, no unique soln is possible ie we use method of least □.
now let,
$y_e = a_0+a_1x+a_2x^2+ \dots a_kx^k$ be the estimated value of y when x takes the value x_i. But the corresponding observed value of y is y_i.
Hence if e_i is the residual as error for this point.

$e_i = y_i - y_e = y_i - a_0 - a_1x_i - a_2x_i^2 \dots a_kx_i^k$
To make the sum of □ min, we have to minimise,

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1x_i - a_2x_i^2 - a_kx_i^k)^2 \quad ②$$

By differencial calculus,
S will have its min-value when,

$$\frac{\partial S}{\partial a_0} = 0 \quad \frac{\partial S}{\partial a_1} = 0 \quad \dots \frac{\partial S}{\partial a_k} = 0.$$

which gives $k+1$ eq → <u>Normal eq.</u>
Solving these eq we get the best values of $a_0, a_1 \dots a_k$
Subst. on ① we get the curve of bestfit.

1) ## Fitting of a straight line :-

$$y = a + bx$$

$$\left.\boxed{\begin{aligned}\Sigma y_i &= na + b\Sigma x_i \\ \Sigma x_i y_i &= a\Sigma x_i + b\Sigma x_i^2\end{aligned}}\right\}\text{Normal eq.}$$

Q) fit a straight line

$x = 1 \quad 2 \quad 3 \quad 4 \quad 5$

$y = 14 \quad 13 \quad 9 \quad 5 \quad 2$

Estimate the value of y when $x = 3.5$

A) $N = 5$

x_i	y_i	x_i^2	$x_i y_i$
1	14	1	14
2	13	4	26
3	9	9	27
4	5	16	20
5	2	25	10
$\Sigma x_i = 15$	$\Sigma y_i = 43$	$\Sigma x_i^2 = 55$	$\Sigma x_i y_i = 97$

$\Sigma y_i = na + b\Sigma x_i \longrightarrow 43 = 5a + 15b$ — ①

$\Sigma x_i y_i = a\Sigma x_i + b\Sigma x_i^2 \longrightarrow 97 = 15a + 55b$ — ②

① × 3

$$\begin{aligned} \cancel{15a} + 45b &= 129 \\ \cancel{15a} + 55b &= 97 \\ \hline -10b &= 32 \\ b &= \underline{\underline{-3.2}} \end{aligned}$$

$15a + 55 \times -3.2 = 97$

$15a - 176 = 97$

$15a = 19 + 176$

$15a = 273 \qquad a = \underline{\underline{18.2}}$

$y = 18.2 - 3.2 \times 3.5 \qquad y = a + bx$

$= 18.2 - 11.2 \qquad x = 3.5$

$\underline{\underline{y = 7}}$

* Suppose to have a straight line that serves as best approximation to the actual curve $y = f(x)$ passing through given points $(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)$. This line will be referred as <u>line of best fit</u>

$$y = a + bx \quad \text{— ①}$$

$a, b \rightarrow$ parameters to be determined.

Let 'y_e' be the value of y corresponding to the value x_i of x as determined by eq — ①. The value $y_e \rightarrow$ <u>Estimated value of y</u>.

when $x = x_i$, the observed value of y is y_i. Then the diff.. $y_i - y_e \rightarrow$ <u>Residual / error</u>.

By the principle of least squares, we have

$$S = \Sigma (y_i - y_e)^2 \quad \text{— ②}.$$

Determine a & b, so that S is min.

Two necessary conditions for this are,

$$\frac{\partial S}{\partial a} = 0 \,, \quad \frac{\partial S}{\partial b} = 0.$$

using — ② these conditions yield the follow-eq.

$\Sigma (y_i - a - bx_i) = 0 \quad$ or $\quad \Sigma y_i = na + b\Sigma x_i$ — ③

$\Sigma (y_i - a - bx_i)\, x_i = 0 \quad$ or

$\Sigma x_i y_i = a\, \Sigma x_i + b\, \Sigma x_i^2$ — ④

— ③ & — ④ $\longrightarrow$ <u>Normal eq.</u> for determining a & b.

Putting the values of a & b so determined — ①, we get the eq of line of least fit for the given data.

② Fitting of parabola :-

$$\boxed{y = a + bx + cx^2}$$

Suppose we wish to hue a parabola as the curve of best fit for a data consisting of 'n' given pairs (x_i, y_i), $i = 1, 2 \ldots n$

(P) of best fit in the form $y = a + bx + cx^2$ — ①

where a, b, c = constants.

Let 'y_e' be the value of 'y' corresponding to the value 'x_i' of x determined by eq — ①, then sum of ~~erro~~ squares of error b/w observed value of 'y' & estimated value of 'y' is given by,

$$S = \left(\sum_{i=1}^{n} y_i - y_e\right)^2 \qquad y_e \rightarrow \text{estimated value.}$$

using eq — ①

$$S = \sum_{i=1}^{n}\left(y_i - (a + bx + cx^2)\right)^2 \text{ — ②}$$

we determined a, b, c, so that 'S' is least

3 necessary conditions for this are, (2+Tue)

$$\frac{\partial s}{\partial a} = 0 \qquad \frac{\partial s}{\partial b} = 0 \qquad \frac{\partial s}{\partial c} = 0$$

using eq ② This conditions yield the following normal eq -

$$\boxed{\begin{aligned} \Sigma y_i &= na + b\,\Sigma x_i + c\,\Sigma x_i^2 \text{ — ③} \\ \Sigma x_i y_i &= a\,\Sigma x_i + b\,\Sigma x_i^2 + c\,\Sigma x_i^3 \text{ — ④} \\ \Sigma x_i^2 y_i &= a\,\Sigma x_i^2 + b\,\Sigma x_i^3 + c\,\Sigma x_i^4 \text{ — ⑤} \end{aligned}}$$

Solve eq — ③, — ④, — ⑤ for determining a, b, c. Putting the values a, b, c, so determining in eq — ① we get the eq of parabola of best fit for the given data.

1) Fit a parabola. $y = a + bx + cx^2$ to the following data.

x : 1 2 3 4 5 6 7

y : 2.3 5.2 9.7 16.5 29.4 35.5 54.4.

A) $n = 7$

x_i	y_i	x_i^2	x_i^3	x_i^4	$x_i y_i$	$x_i^2 y_i$
1	2.3	1	1	1	2.3	2.3
2	5.2	4	8	16	10.4	20.8
3	9.7	9	27	81	29.1	87.3
4	16.5	16	64	256	66.0	264.0
5	29.4	25	125	625	147	735
6	35.5	36	216	1296	213	1278
7	54.4	49	343	2401	380.8	2665.6
$\Sigma x_i = 28$	$\Sigma y_i = 153$	$\Sigma x_i^2 = 140$	$\Sigma x_i^3 = 784$	$\Sigma x_i^4 = 4676$	$\Sigma x_i y_i = 848.6$	$\Sigma x_i^2 y_i = 5053$

$153 = 7a + 28b + 140c$ — ①

$848.6 = 28a + 140b + 784c$ — ②

$5053 = 140a + 784b + 4676c$ — ③

153 × 4 = 612

28 × 4 = 112

① × 4

~~28a + 112b + 560c = 612~~

① & ②

$28a + 140b + 784c = 848.6$

$28a \pm 112b \pm 560c = \pm 612$

$28b + 224c = 236.6$ — ④

eq—② × 5 → $140a + 700b + 3920c = 4243$

$$\begin{aligned} \cancel{140a} + 784b + 4676c &= 5053 \\ \cancel{140a} \pm 700b \pm 3920c &= \underline{-}4243 \\ \hline 84b + 756c &= 810 \quad \text{—⑤} \end{aligned}$$

eq—④ × 3 → $84b + 672c = 709.8$

$$\begin{aligned} \cancel{84b} + 756c &= 810 \\ \cancel{84b} \pm 672c &= \underline{-}709.8 \\ \hline 84c &= 100.2 \\ c &= 1.1928 \rightarrow \underline{1.193} \end{aligned}$$

$84b + 672 \times 1.193 = 709.8$

$84b + 801.696 = 709.8$

$84b = 709.8 - 801.696$

$84b = -91.896$

$b = \underline{-1.094}$

$7a + 28b + 140c = 153$

$7a + 28 \times -1.094 + 140 \times 1.193 = 153$

$7a - 30.632 + 167.02 = 153$

$7a + 136.388 = 153$

$7a = 153 - 136.388$

$7a = 16.612$

$a = \underline{2.37}$

$a = 2.37$

$b = -1.094$

$c = 1.193$

③ **Fitting a parabola :-** $y = ab^x$.

$$\left[\begin{aligned} \log(ab) &= \log a + \log b \\ \log(a^b) &= b \log a. \end{aligned}\right]$$

Fitting of a curve of the form $y = ab^x$

Suppose we wish to fit a curve whose eq → $y = ab^x$ —①

As the curve of best fit for a data consisting of pairs (x_i, y_i), $i = 1, 2, \ldots, n$. taking log on both sides of eq —① we get,

$y = ab^x$

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

$\log y = u$

$\log a = A$ $\qquad u = A + Bx$ —②

$\log b = B$

Is a linear eq, ∴ the normal eq that yield A & B —

$$\boxed{\begin{aligned} \Sigma u_i &= nA + B\Sigma x_i \quad \text{—③} \\ \Sigma x_i u_i &= A\Sigma x_i + B\Sigma x_i^2 \quad \text{—④} \end{aligned}}$$

where $u_i = \log y_i$

Solving this eq we obtain A & B —

$a = \text{antilog } A$

$b = \text{antilog } B$

Substituting this values of a & b in eq—① we obtain the curve best fit for the given data.

1) Fit a curve. $y = ab^x$

x : 1 2 3 4 5 6 7 8

y : 1.0 1.2 1.8 2.5 3.6 4.7 6.6 9.1

A) $n = 8$

(ln + no.)

x_i	y_i	$u_i = \log_e y_i$	x_i^2	$x_i u_i$
1	1.0	0.000	1	0
2	1.2	0.1823	4	0.3646
3	1.8	0.5878	9	1.7634
4	2.5	0.9163	16	3.6652
5	3.6	1.2809	25	6.4045
6	4.7	1.5475	36	9.285
7	6.6	1.8870	49	13.209
8	9.1	2.2082	64	17.6656
$\Sigma x_i = 36$	$\Sigma y_i = 30.5$	$\Sigma u_i = 8.61$	$\Sigma x_i^2 = 204$	$\Sigma x_i u_i = 52.3573$

$\Sigma u_i = nA + B\Sigma x_i \rightarrow 8.61 = 8A + 36B$ $\times 36$

$\Sigma x_i u_i = A\Sigma x_i + B\Sigma x_i^2 \rightarrow 52.3573 = 36A + 204B$ $\times 8$

$309.96 = 288A + 1296B$

$418.8584 = 288A + 1632B$

$\cancel{288A} + 1632B = 418.8584$

$\cancel{288A} \pm 1296B = 309.96$

$336B = 108.8984$

$B = 0.3241$

$8A + 36 \times 0.3241 = 8.61$

$8A + 11.6676 = 8.61$

$8A = 8.61 - 11.6676$

$8A = -3.0576$

$A = -0.3822$

$a = \text{anti}\log A \rightarrow$ shift + ln → no

$= \text{antilog}(-0.3822) = 0.684$

$b = \text{antilog } B$

$= \text{antilog}(0.3241) = 1.3828$

$y = ab^x$

$= (0.684)(1.3828)^x$

H/W

1) Fix a parabola –

x : 1 2 3 4 5 6 7 8 9

y : 2 6 7 8 10 11 11 10 9

$x = 4.5$ $y = ?$

A) $n = 9$.

x_i	y_i	x_i^2	x_i^3	x_i^4	$x_i y_i$	$x_i^2 y_i$
1	2	1	1	1	2	2
2	6	4	8	16	12	24
3	7	9	27	81	21	63
4	8	16	64	256	32	128
5	10	25	125	625	50	250
6	11	36	216	1296	66	396
7	11	49	343	2401	77	539
8	10	64	512	4096	80	640
9	9	81	729	6561	81	729

$\Sigma x_i = 45$	$\Sigma y_i = 74$	$\Sigma x_i^2 = 285$	$\Sigma x_i^3 = 2025$	$\Sigma x_i^4 = 15333$	$\Sigma x_i y_i = 421$	$\Sigma x_i^2 y_i = 2771$

$74 = 9a + 45b + 285c$ —— ①

$421 = 45a + 285b + 2025c$ —— ②

$2771 = 285a + 2025b + 15333c$ —— ③

① × 5 $\rightarrow$ $370 = 45a + 225b + 1425c$

$45a + 285b + 2025c = 421$
$45a \pm 225b \pm 1425c = 370$

$60b + 600c = 51$ —— ④

② × 285
③ × 45

$12825a + 91125b + 689985c = 124695$
$12825a \pm 81225b \pm 577125c = 119985$

$9900b + 112860c = 4710$ —— ⑤

$9900b + 112860c = 4710$ × 60
$60b + 600c = 51$ × 9900

$594000b + 6771600c = 282600$
$594000b \pm 5940000c = 504900$

$831600c = -222300$

$c = -0.2673$

$60b + 600 \times -0.2673 = 51$

$60b - 160.38 = 51$

$60b = 51 + 160.38$

$60b = 211.38$

$b = 3.523$

$45a + 285 \times 3.523 + 2025 \times 0.2673 = 421$

$45a + 1004.055 - 541.2825 = 421$

$45a + 462.7725 = 421$

$45a = 421 - 462.7725$

$45a = -41.7725$

$a = 0.9282$

$a = 0.9282$
$b = 3.523$
$c = -0.2673$

④ Fitting of a curve $y = ax^b$:-

Suppose we wish to have a curve whose eq is in the form $y = ax^b$ —①
As the curve of best fit for a data consisting of the pairs (x_i, y_i), $i = 1, 2, \ldots n$
Taking log on both sides of eq ①
we get,

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

$\log y = u$
$\log a = A$
$\log x = X$

$u = A + bX$ —② is a linear eq

$$\boxed{\begin{aligned} \Sigma u_i &= nA + b\Sigma x_i \quad ③ \\ \Sigma x_i u_i &= A\Sigma x_i + b\Sigma x_i^2 \quad ④ \end{aligned}}$$

⑤ Fitting of a curve ~~y = ae~~ $y = ae^{bx}$:-

If $y = ae^{bx}$ —①
As the curve of best fit for a data.

$\log y = \log a + bx \log e$ —②

$u = \log y$
$A = \log a$
$B = b \log e$

$u = A + BX$ —③

$$\boxed{\begin{aligned} \Sigma u_i &= nA + B\Sigma x_i \quad \text{---} (4) \\ \Sigma x_i u_i &= A\Sigma x_i + B\Sigma x_i^2 \quad \text{---} (5) \end{aligned}}$$

Solving eq (4) & (5) we get A & B

$$\boxed{\begin{aligned} a &= \text{antilog } A \\ b &= \cancel{\text{antilog } B}\ \frac{B}{\log e} \end{aligned}}$$

Substituting a, b in eq (1) we get the curve of best fit for the given pairs of observations.

(6) Fitting of a curve $y = ax^2 + \frac{b}{x}$:-

To fit this curve for a set of observation $(x_1, y_1)\ (x_2, y_2) \dots (x_n, y_n)$. The normal eq –

$$\boxed{\begin{aligned} \Sigma y &= a\Sigma x^2 + b\Sigma \frac{1}{x} \\ \Sigma xy &= a\Sigma x^3 + nb \end{aligned}}$$

Solve 2 eq for a & b, substituting a, b in eq, we get the required curve of best fit.

1) Fit a straight line (by least □'s) to following data –

x : 1 2 3 4 5

y : 35 68 100 138 170

A) $n = 5$

$y = a + bx$.

$\Sigma y_i = na + b\Sigma x_i$ — (1) $511 = 5a + 15b$ ×3

$\Sigma x_i y_i = a\Sigma x_i + b\Sigma x_i^2$ — (2) $1873 = 15a + 55b$

x_i	y_i	x_i^2	$x_i y_i$
1	35	1	35
2	68	4	136
3	100	9	300
4	138	16	552
5	170	25	850
$\Sigma x_i = 15$	$\Sigma y_i = 511$	$\Sigma x_i^2 = 55$	$\Sigma x_i y_i = 1873$

$15a + 45b = 1533$

$15a + 55b = 1873$

$10b = 340$

$b = 34$

$15a + 45 \times 34 = 1533$

$15a + 1530 = 1533$

$15a = 1533 - 1530$

$15a = 3$

$a = 0.2$

Line of best fit –

$y = a + bx$

$= 0.2 + 34x$

H/W

2) x : 0 1 2 3 4

y : 1 1.8 3.3 4.5 6.3

3) x : 1 2 3 4 5 6

y : 2.4 3 3.6 4 6 8

4) x : 1 2 3 4 5 6 7

y : 80 90 92 83 94 99 92.

2.A)

x_i	y_i	x_i^2	x_iy_i
0	1	0	0
1	1.8	1	1.8
2	3.3	4	6.6
3	4.5	9	13.5
4	6.3	16	25.2
$\Sigma x_i=10$	$\Sigma y_i=16.9$	$\Sigma x_i^2=30$	$\Sigma x_iy_i=47.1$

$n=5$

$16.9 = 5a + 10b \quad \times 3$

$47.1 = 10a + 30b$

$15a + \cancel{30b} = 50.7$

$10a + \cancel{30b} = 47.1$

$5a = 3.6 \quad a = 0.72$

$\cancel{a = 1.8}$

$5 \times 0.72 + 10b = 16.9$

$3.6 + 10b = 16.9$

$10b = 16.9 - 3.6$

$10b = 13.3$

$b = 1.33$

$y = 0.72 + 1.33x.$

3.A)

x_i	y_i	x_i^2	$x_i y_i$
1	2.4	1	2.4
2	3	4	6
3	3.6	9	10.8
4	4	16	16
5	6	25	30
6	8	36	48
$\Sigma x_i=21$	$\Sigma y_i=27$	$\Sigma x_i^2=91$	$\Sigma x_iy_i = 113.2$

$27 = 6a + 21b \quad \times 21$

$113.2 = 21a + 91b. \quad \times 6 \qquad n=6.$

$126a + 546b = 679.2$

$\cancel{126a} \pm 441b = 567$

$105b = 112.2$

$b = 1.068$

$6a + 21 \times 1.068 = 27$

$6a + 22.428 = 27$

$6a = 27 - 22.428$

$6a = 4.572$

$a = 0.762$

$y = a + bx$

$y = 0.762 + 1.068x.$

4.A)

x_i	y_i	x_i^2	x_iy_i
1	80	81	80
2	90	4	180
3	92	9	276
4	83	16	332
5	94	25	470
6	99	36	594
7	92	49	644
$\Sigma x_i=28$	$\Sigma y_i=630$	$\Sigma x_i^2=140$	$\Sigma x_iy_i=2576$

$630 = 7a + 28b \quad \times 28 \qquad n = 7$

$2576 = 28a + 140b \quad \times 7$

$196a + 980b = 18032$

$196a \pm 784b = \mp 17640$

$196b = 392$

$b = 2.$

$7a + 28\times2 = 630$

$7a + 56 = 630$

$7a = 630 - 56$

$7a = 574$

$a = 82$

$y = a + bx$

$y = 82 + 2x$

1) Derive the least □ eqs for fitting a curve of the type $y = ax + \frac{b}{x}$ to a set of n points (x_i, y_i) $i = 1,2,3 \ldots n$.

A) The error of estimate (e_i) for i^{th} point (x_i, y_i) is given by, $\quad y = ax + \frac{b}{x}$

$$e_i = \left(y_i - ax_i - \frac{b}{x_i}\right)$$

According to principle of least □, we have to determine the values a & b, so that the sum of the □'s of errors

$$S = \Sigma e_i^2 = \sum_{i=1}^{n}\left(y_i - ax_i - \frac{b}{x_i}\right)^2 \text{ is min.}$$

Consequently the normal eq -

$$\frac{\partial S}{\partial a} = 0 \Rightarrow -2\sum_{i=1}^{n} x_i \sum_{i=1}^{n}\left(y_i - ax_i - \frac{b}{x_i}\right) \text{ — ①}$$

$$\frac{\partial S}{\partial b} = 0 \Rightarrow -2\sum_{i=1}^{n} \frac{1}{x_i}\left(y_i - ax_i - \frac{b}{x_i}\right) \text{ — ②}$$

① × ②

② × ②

$$\sum_{i=1}^{n} x_iy_i = a\Sigma x_i^2 + nb$$

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

H/w

2) $y = ae^{bx}$.

x :	1	2	3	4	5	6
y :	1.6	4.5	13.8	40.2	125.0	300

3) $y = ax^2 + \frac{b}{x}$

x :	1	2	3	4
y :	1.51	0.99	3.88	7.66

4) $y = \frac{a}{x} + bx$

x :	1	2	3	4	5	6	7	8
y :	5.40	6.30	8.20	10.30	12.60	14.40	17.30	19.58

2.A)

Σx_i	Σy_i	Σx_i^2	$\Sigma u_i = \log_e y_i$	$x_i u_i$
1	1.6	1	0.4700	0.47
2	4.5	4	1.5040	3.008
3	13.8	9	2.6246	7.8738
4	40.2	16	3.6938	14.7752
5	125.0	25	4.8283	24.1415
6	300	36	5.7037	34.2222
21	448.62	91	18.8244	84.4907

$18.8244 = 6A + 21B \quad \times 21$
$84.4907 = 21A + 91B \quad \times 6$

$126A + 546B = 506.9424$
$126A + 441B = 395.3124$

$105B = 111.6318$

$B = 1.06316$

$6A + 21 \times 1.06316 = 18.8244$
$6A + 22.32636 = 18.8244$
$6A = 3.50196$
$A = -0.58366$

$a = \text{antilog}(-0.58366)$
$= 0.557$

$b = B/\log e$ (value→ $\ln = 2.71$) → $\frac{1.06316}{0.9996} = 1.05$

$= 1.05$

$y = 0.557 e^{1.05x}$

3.A)

x	y	x^2	x^3	xy	$1/x$
1	1.51	1	1	1.51	1
2	0.99	4	8	1.98	0.5
3	3.88	9	27	11.64	0.33
4	7.66	16	64	30.64	0.25
10	14.04	30	100	45.77	2.08

$14.04 = 30a + 2.08b \quad \times 100$
$45.77 = 100a + 4b \quad \times 30$

$3000a + 2.08b = 1404$
$3000a + 120b = 1373.1$

$88b = 30.9$

$b = 0.35$

$100a + 4 \times 0.35 = 45.77$
$100a + 1.4 = 45.77$
$100a = 44.37$
$a = 0.4437$

$y = 0.4437x^2 + \frac{0.35}{x}$

4.A) $y = \frac{a}{x} + bx$

$y_i + \frac{a}{x_i} + bx_i = 0$

$S = e_i^2 = \left(y_i - \frac{a}{x_i} - bx_i\right)^2$

$(0 - 1 - 0 = -1$
2×-1
$= -2)$

$\frac{\partial S}{\partial a} = 0 \quad -2 \frac{1}{x_i}\left(y_i - \frac{a}{x_i} - bx_i\right) = 0/-2$

$\Rightarrow \frac{1}{x_i}\left(y_i - \frac{a}{x_i} - bx_i\right) = 0 \text{ — } ①$

$\frac{\partial S}{\partial b} = 0 \quad -2 x_i \left(y_i - \frac{a}{x_i} - bx_i\right) = 0 \quad (0 - 0 - 1 = -1)$

$x_i\left(y_i - \frac{a}{x_i} - bx_i\right) = 0 \text{ — } ②$

$① \Rightarrow \boxed{\frac{y_i}{x_i} = a\frac{1}{x_i^2} + nb}$

$② \Rightarrow \boxed{x_i y_i = na + bx_i^2}$

x_i	y_i	x_i^2	$\frac{1}{x_i^2}$	$x_i y_i$	y_i/x_i
1	5.40	1	1	5.4	5.4
2	6.30	4	0.25	12.6	3.15
3	8.20	9	0.1111	24.6	2.73
4	10.30	16	0.0625	41.2	2.575
5	12.60	25	0.04	63	2.52
6	14.90	36	0.0277	89.4	2.48

7	17.30	49	0.0204	121.1	2.47
8	19.50	64	0.015625	156	2.43
		204	1.527325	513.3	23.755

$\Sigma x_i y_i = na + b\Sigma x_i^2$

$\Sigma \frac{y_i}{x_i} = a\Sigma \frac{1}{x_i^2} + nb$

$513.3 = 8a + 204b \quad \times 1.527$

$23.755 = 1.527a + 8b \quad \times 8$

$12.216a + 311.508b = 783.8091$

$12.216a + 64b = 190.04$

$247.508b = 593.7691$

$b = 2.398$

$12.216a + 64 \times 2.398 = 190.04$

$12.216a + 153.672 = 190.04$

$12.216a = 36.568$

$a = 2.99$

$y = \frac{2.99}{x} + 2.398x$

2: Correlation & Regression.

* Correlation :- It is a statistical ~~measure~~ ~~metho~~ measure for finding out degree of association b/w 2/more variable by association mean the tentency of the variables to move together.

2 variables x & y are related x tend to accombined by the corresponding movements in the other ~~x~~ 'y'. Then x & y are said to be correlated. The movements may be in the same (dir)/ opposite (dir). (x↑ y↑ → same (dir), x↓ y↑ → oppo) (c) said to be +ve/-ve according as this movemnts are in the same (in the opposite (dir). If 'y' is unaffected by any change in 'x' then x & y are said to be uncorrelated.

* LR conner's Definition :-

If 2/more quantities vary in the sympathy, so that movemnts in the 1 tend to be accombined by corresponding movemnts in the other, then they are said to be correlated

→ Linear (c) :-

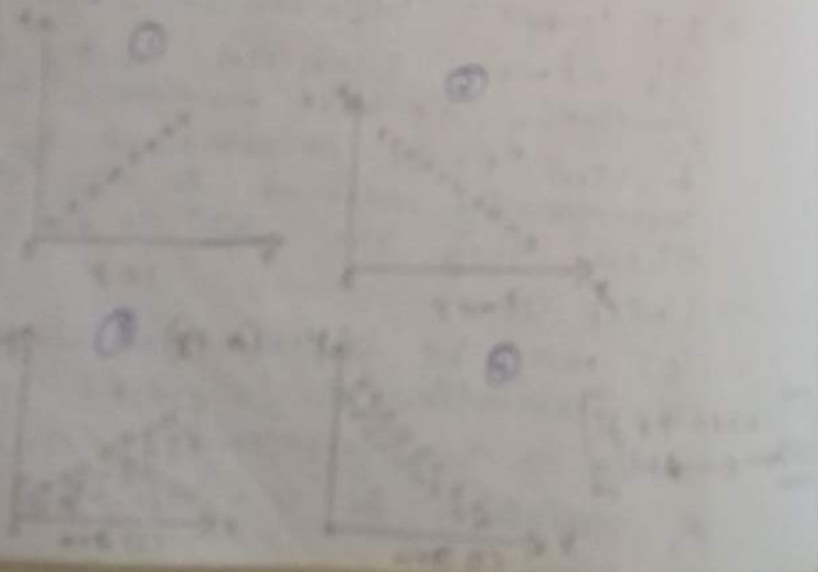
(c) may be linear / non-linear. The variation in 'x' bears a constant ratio to the corresponding amount of variation in 'y'. Then (c) b/s x & y is said to be linear otherwise it is non-linear.

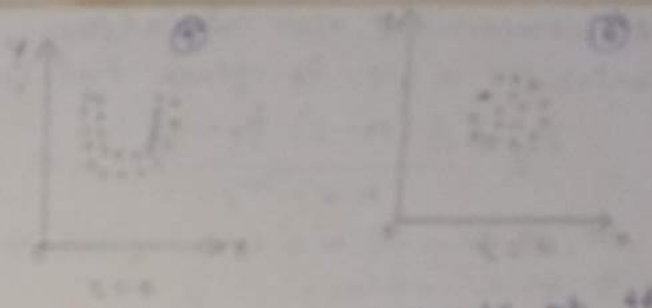
→ Interpretation of CC :-

1) Scatter Diagram
2) Covariance method / Karl Pearson's method
3) Rank method.

1) Scatter Diagram :-

* Statistical data relating to simultaneous movements of 2 variables can be graphically represented by points.
* x is along the horizontal axis & y is along the vertical axis.
* All the pairs of values of x & y are shown by points on graph paper. This diagrammatic representation of bivariate data → scatter diagram.





* In fig ① & ② the movements of the 2 variables are in the same (dir) & scatter diagram shows a linear path. Hence (C) is +ve / direct.
* In fig ③ & ④ movements of 2 variables are in opposite (dir) & scatter diagram shows a linear path. In this case (C) is -ve / indirect.
* In fig ⑤ & ⑥ it is ... showed a curve as given a ...

→ Karl Pearson's correlation coefficient :

(r) is a statistical tool which studies the relationship b/w 2 variables. We cannot simply look at a scatter diagram & conclude that more than ½ of the points appear to lie nearly in a straight line, there is a +ve/-ve (r) b/w the variables. On the other hand neither can conclude that there is no (C) at all. We need a quantity which is a measure of the extent to which x & y are related. The quantity which is used for this purpose → coefficient of (C) (r_{xy}) or (r)

It measures the ~~r~~relation b/w 2 variables x & y is given the formula,

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n \sigma_x \sigma_y} \quad \text{—} (1).$$

x_i, y_i $(i = 1, 2, \dots n)$ = 2 sets of values of x & y.

$\bar{x}, \bar{y}$ = means

σ_x, σ_y = standard deviation of x & y.

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

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

$$\sigma_y^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i^2 - \bar{y}^2).$$

If $(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)$ be 'n' pairs of observations on 2 variables x & y, then covariance of x & y is,

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

Cov indicates the joint variations b/w 2 variables. ∴ (c) coefficient b/w x & y is –

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

'r' can be written in diff forms –

$$X_i = x - \bar{x} \qquad Y_i = y - \bar{y}$$

$$r = \frac{\sum x_i y_i}{n \sigma_x \sigma_y}$$

~~form~~ – (1)

$$r = \frac{\cancel{\frac{1}{n}} \sum_{i=1}^{n} x_i y_i}{\sqrt{\frac{\sum x_i^2}{\cancel{n}}} \times \sqrt{\frac{\sum y_i^2}{\cancel{n}}}} \qquad \sigma_x^2 = \frac{\sum x_i^2}{n} \qquad \sigma_y^2 = \frac{\sum y_i^2}{n}$$

cancel $1/n$ from numrtr & denomentr.

$$r = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum x_i^2} \times \sqrt{\sum y_i^2}}$$

$$\text{cov}(x, y) = \frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})$$

$[(a-b)(c-d) = ac - ad - bc + bd]$

$$= \frac{1}{n} \sum (x_i y_i - x_i \bar{y} - y_i \bar{x} + \bar{x}\bar{y})$$

$$= \frac{\sum x_i y_i}{n} - \frac{\bar{y} \sum x_i}{n} - \frac{\bar{x} \sum y_i}{n} + \bar{x}\bar{y}$$

$$= \frac{\sum x_i y_i}{n} - \cancel{\bar{x}\bar{y}} - \bar{x}\bar{y} + \cancel{\bar{x}\bar{y}}$$

$$= \frac{\sum x_i y_i}{n} - \bar{x}\bar{y} \qquad \bar{x} = \frac{\sum x_i}{n}$$

$$\text{cov}(x, y) = \frac{\sum x_i y_i}{n} - \frac{\sum x_i}{n} \times \frac{\sum y_i}{n}$$

$$r = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} = \frac{\frac{\sum x_i y_i}{n} - \frac{\sum x_i}{n} \times \frac{\sum y_i}{n}}{\sqrt{\frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n}\right)^2} \times \sqrt{\frac{\sum y_i^2}{n} - \left(\frac{\sum y_i}{n}\right)^2}}$$

multiply by n^2,

$$\boxed{r = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}}$$

* Theorem :- The (c) coefficient is independent of change of origin & scale of measurement.

Proof

Let $(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)$ be a set of 'n' pairs of observations.

$$r_{uv} = \frac{\frac{1}{n}\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum(x_i-\bar{x})^2}\sqrt{\frac{1}{n}\sum(y_i-\bar{y})^2}} \quad \text{—(1)}$$

Let us transform $x_i = u_i$, $y_i = v_i$

$$u_i = \frac{x_i - x_0}{c_1} \qquad v_i = \frac{y_i - y_0}{c_2} \quad \text{—(2)}$$

x_0, y_0, c_1, c_2 are arbitary constants.

from —(2), we have

$$x_i = c_1 u_i + x_0 \qquad y_i = c_2 v_i + y_0$$

$$\bar{x} = x_0 + c_1\bar{u} \qquad \bar{y} = y_0 + c_2\bar{v}.$$

$$x_i - \bar{x} = c_1(u_i - \bar{u})$$
$$y_i - \bar{y} = c_2(v_i - \bar{v})$$

Substitute these into —(1)

$$r_{uv} = \frac{\frac{1}{n}\sum_{i=1}^{n} c_1(u_i-\bar{u})\, c_2(v_i-\bar{v})}{\sqrt{\frac{1}{n}\sum c_1^2(u_i-\bar{u})^2} \times \sqrt{\frac{1}{n}\sum c_2^2(v_i-\bar{v})^2}}$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n}(u_i-\bar{u})(v_i-\bar{v})}{\sqrt{\frac{1}{n}\sum(u_i-\bar{u})^2}\sqrt{\frac{1}{n}\sum(v_i-\bar{v})^2}}$$

$$r_{uv} = \frac{\sum_{i=1}^{n}(u_i-\bar{u})(v_i-\bar{v})}{n\,\sigma_u\sigma_v}$$

(c) coefficient remains unchanged

Here r_{uv} simplified as,

$$r_{uv} = \frac{Cov(u,v)}{\sigma_u\sigma_v}$$

[find Cov (u,v) like Cov (x,y)]

$$= \frac{\frac{1}{n}\sum_{i=1}^{n} u_i v_i - \bar{u}\bar{v}}{\sqrt{\frac{1}{n}\sum u_i^2 - \bar{u}^2}\sqrt{\frac{1}{n}\sum v_i^2 - \bar{v}^2}}$$

$\times n^2$

$$\boxed{r_{uv} = \frac{n\sum u_i v_i - \sum u_i \sum v_i}{\sqrt{n\sum u_i^2 - (\sum u_i)^2}\sqrt{n\sum v_i^2 - (\sum v_i)^2}}}$$

Remark – In general $\sigma_x = |c|\sigma_u$

$\sigma_y = |d|\sigma_v$

$\therefore r_{xy} = \frac{cd}{|c|\cdot|d|} \cdot r_{uv}$

Now $\frac{cd}{|c|\cdot|d|} = +1$ or -1 according as c.

$\therefore r_{xy} = \pm r_{uv}$ according as c, d have same / opposite sign.

Note – In actual computations, we can take $c_1 = c_2 = 1$ so we assume $u_i = x_i - x_0$ & $v_i = y_i - y_0$,

$x_0, y_0 \rightarrow$ should be chosen that most of u_i & v_i respectively, numerically $>$ x_i & y_i.

* 235

~~→ Limits of (c) co-efficient :-~~

⟹ * Derivation of Spearman's formula for Rank (c) coefficient :-

$$R = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

Let $(x_1, y_1) \ldots (x_n, y_n)$ be the ranks of n individuals in 2 characters

Edward Spearman's Rank (ρ) Coefficient

R is the product moment (C) coefficient b/w these ranks,

$$R = \frac{cov(x,y)}{\sigma x \sigma y}$$

$$cov(x,y) = \frac{\Sigma\{(x_i-\bar{x})(y_i-\bar{y})\}}{n}$$

n = natural no.

$x_1, x_2 \ldots x_n$ are $1, 2, \ldots n$ in same order.

$$\Sigma x = 1+2+\ldots n = \frac{n(n+1)}{2}$$

(Σx^2)

$$\Sigma x^2 = 1^2+2^2 \ldots n^2 = \frac{n(n+1)(2n+1)}{6}$$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{n+1}{2} \qquad \frac{\frac{\cancel{n}(n+1)}{2}}{\cancel{n}} = \frac{n+1}{2}$$

$$\frac{\Sigma x^2}{n} = \frac{(n+1)(2n+1)}{6}$$

$$\therefore \sigma x^2 = \frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2 = \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4}$$

$$= \frac{4\left((n+1)(2n+1) - 6(n+1)^2\right)}{24} \quad (a+b)^2$$

$$= \frac{4(2n^2+n+2n+1) - 6(n^2+2n+1)}{24}$$

$$= \frac{8n^2+4n+8n+4-6n^2-12n-6}{24}$$

$$= \frac{8n^2+12n+4-6n^2-12n-6}{24}$$

$8n^2-6n^2 = 2n^2$ $\quad 4-6=-2$

$$= \frac{2n^2-2}{24} = \frac{\cancel{2}(n^2-1)}{\cancel{24}_{12}}$$

$$= \frac{n^2-1}{12}$$

Similarly,

$$\bar{y} = \frac{n+1}{2} \qquad \sigma y^2 = \frac{n^2-1}{12}$$

Let $d_i = x_i - y_i$, $d_i = (x_i-\bar{x})-(y_i-\bar{y})$

$$\therefore \frac{\Sigma d_i^2}{n} = \frac{\Sigma[(x_i-\bar{x})(y_i-\bar{y})]^2}{n} \quad (a-b)^2$$

$$= \frac{\Sigma(x_i-\bar{x})^2}{n} + \frac{\Sigma(y_i-\bar{y})^2}{n} - 2\frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{n}$$

$$= \sigma x^2 + \sigma y^2 - 2\,cov(x,y)$$

$$2\,cov(x,y) = \frac{n^2-1}{12} + \frac{n^2-1}{12} - \frac{\Sigma d_i^2}{n}$$

$$\cancel{2}\,cov(x,y) = \frac{\cancel{2}(n^2-1)}{12} - \frac{\Sigma d_i^2}{n}$$

$$cov(x,y) = \frac{n^2-1}{12} - \frac{\Sigma d_i^2}{2n}$$

from eq-①,

$$R = \frac{\frac{n^2-1}{12} - \frac{\Sigma d_i^2}{2n}}{\sqrt{\frac{n^2-1}{12}} \times \sqrt{\frac{n^2-1}{12}}} \qquad \sqrt{\ } \times \sqrt{\ } = x$$

$$= \frac{\frac{n^2-1}{12} - \frac{\Sigma d^2}{2n}}{\left(\frac{n^2-1}{12}\right)} = \frac{\cancel{\frac{n^2-1}{12}}}{\cancel{\frac{n^2-1}{12}}} - \frac{\frac{\Sigma d^2}{2n}}{\frac{n^2-1}{12}}$$

$$\frac{a/b}{c/d} = \frac{ad}{bc} = \frac{\Sigma d^2 \times 12^6}{2n(n^2-1)}$$

$$\boxed{R = 1 - \frac{6\Sigma d^2}{n(n^2-1)}} \qquad = 1 - \frac{6 \Sigma d^2}{n(n^2-1)}$$

→ Limits of (c) coefficient :-

Now find the limits of (c) coefficient b/w 2 variables & show that it lies b/w −1 & +1.

(i.e) $-1 \le r_{xy} < +1$

Proof
Let $(x_1, y_1) \dots (x_n, y_n)$ be given pairs of observations.

$$r_{xy} = \frac{\frac{1}{n} \Sigma (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\Sigma(x_i-\bar{x})^2}\sqrt{\frac{1}{n}\Sigma(y_i-\bar{y})^2}} \quad \text{— (1)}$$

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

$\sigma_y^2 = \frac{1}{n}\Sigma(y_i-\bar{y})^2 = \frac{\Sigma y_i^2}{n}$

Then eq,

$$r_{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{n \sigma_x \sigma_y} \quad \text{— (2)}$$

Now split eq.

$$\sum_{i=1}^{n}\left(\frac{x_i}{\sigma_x} \pm \frac{y_i}{\sigma_y}\right)^2 = \frac{\Sigma x_i^2}{\sigma_x} + \frac{\Sigma y_i^2}{\sigma_y} + \frac{2\Sigma x_i y_i}{\sigma_x \sigma_y} \quad \text{— (3)}$$

$(a+b)^2$

(3) ÷ n

$$= \frac{n\sigma_x^2}{\sigma_x^2} + \frac{n\sigma_y^2}{\sigma_y^2} + 2nr_{xy}.$$

$$= n + n + 2nr_{xy}$$

$$= 2n \pm 2nr_{xy} = 2n(1 \pm r_{xy})$$

Left hand side of above identity is the sum of □ of n non-ve hence it is +ve / 0.

Hence, $1 \pm r_{xy} > 0$ (or) $r_{xy} \le 1$ & $r_{xy} \ge -1$

(or) $-1 \le r_{xy} \le +1$

∴ (c) coefficient lies b/w −1 & +1.

1) find coefficient of (c),

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

$$\boxed{r = \frac{\frac{1}{n}\Sigma(x_i-\bar{x}) \cdot \Sigma(y_i - \bar{y})}{\sqrt{\frac{1}{n}(\Sigma x_i - \bar{x})^2} \cdot \sqrt{\frac{1}{n}(\Sigma y_i - \bar{y})^2}}}$$

$$\Rightarrow \boxed{\frac{(x-\bar{x})(y-\bar{y})}{\sqrt{(x-\bar{x})^2}\sqrt{(y-\bar{y})^2}}} \Rightarrow \boxed{\frac{xy}{\sqrt{x^2}\sqrt{y^2}}}$$

$\Sigma x = \Sigma x \qquad y = \Sigma y.$

$\bar{x} = \frac{\Sigma x}{n}$

$\bar{y} = \frac{\Sigma y}{n}$

A)

x	y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	xy	x^2	y^2
1	6	-3	-4	12	9	16
2	8	-2	-2	4	4	4
3	11	-1	1	-1	1	1
4	9	0	-1	0	0	1
5	12	1	2	2	1	4
6	10	2	0	0	4	0
7	14	3	4	12	9	16
28	70	0	0	29	28	42

$\bar{Y} = \frac{\Sigma Y}{n} = \frac{70}{7} = 10$ $\qquad \bar{X} = \frac{28}{7} = 4$

$$r = \frac{\frac{1}{n}\Sigma x_i - \bar{x} \cdot \Sigma y_i - \bar{y}}{\sqrt{\frac{1}{n}\Sigma(x_i-\bar{x})^2} \cdot \sqrt{\frac{1}{n}\Sigma(y_i-\bar{y})^2}}$$ $\qquad \sqrt{1/n} \times \sqrt{1/n} = \frac{1}{n}$

$$r = \frac{\frac{1}{n}\Sigma x_i - \bar{x} \cdot \Sigma y_i - \bar{y}}{\frac{1}{n}\Sigma(x_i-\bar{x})^2 \Sigma(y_i-\bar{y})^2}$$

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

$x_i - \bar{x} = x$

$y_i - \bar{y} = y$

$$\boxed{r = \frac{xy}{\sqrt{x^2} \cdot \sqrt{y^2}}}$$

$r = \frac{29}{\sqrt{28}\sqrt{42}}$ $\qquad x^2 = 28 \qquad y^2 = 42$

$= \frac{29}{5.291 \times \cancel{6.928}\ 6.4807}$ $= \frac{}{\cancel{36.650}}$

$= \frac{29}{34.289} = 0.845$

2) ~~Kent~~ Karl pearson's coefficient of (c) b/w 2 variables x & y is 0.28. Their covarience is +7.6. If the variance of x is 9, find standard deviation of y series.

A) $r_{xy} = \frac{cov(x,y)}{\sigma x \, \sigma y}$ $\qquad$ std. dev - of variance $\qquad \sqrt{\text{variance}}$ = standard dev.

$0.28 = \frac{7.6}{3 \cdot \sigma y}$ $\qquad \Rightarrow \sigma x^2 = 9$

$\sigma x = 3 \qquad \sigma y = ?$

$\sigma y = \frac{7.6}{3 \times 0.28}$ $\qquad r_{xy} = 0.28$

$cov(x,y) = 7.6$

$\sigma y = \frac{7.6}{0.84} = 9.047$

3) Calculate pearson's coefficient of (c) from the following taking 100 & 50 as the assumed avg of x & y.

x:	104	111	104	114	118	117	105	108	106	100	104	105
y:	57	55	47	45	45	50	64	63	66	62	69	61

X	y	$u = x - \bar{x}'$	$v = y - \bar{y}$ 50	u^2	v^2	uv
104	57	4	7	16	49	28
111	55	11	5	121	25	55
104	47	4	-3	16	9	-12
114	45	14	-5	196	25	-70
118	45	18	-5	324	25	-90
117	50	17	0	289	0	0
105	64	5	14	25	196	70
108	63	8	13	64	169	104
106	66	6	16	36	256	96
100	62	0	12	0	144	0
104	69	4	19	16	361	76
105	61	5	11	25	121	55
		96	84	1128	1380	312

$n = 12$

$$r = \frac{n \sum u_i v_i - \sum u_i \sum v_i}{\sqrt{n \sum u_i^2 - \sum (u_i)^2}\sqrt{n \sum v_i^2 - \sum (v_i)^2}}$$

$$= \frac{12 \times 312 - 96 \times 84}{\sqrt{12 \times 1128 - (96)^2}\sqrt{12 \times 1380 - (84)^2}}$$

$$= \frac{3744 - 8064}{\sqrt{13536 - (96)^2}\sqrt{16560 - (84)^2}}$$

$$\cancel{= \frac{-4320}{\sqrt{13440}\sqrt{16476}}} \cancel{= \frac{-4320}{115}}$$

$$= \frac{-4320}{\sqrt{13536 - 9216}\sqrt{16560 - 7056}} = \frac{-4320}{\sqrt{4320}\sqrt{9504}}$$

$$= \frac{-4320}{6407.595493}$$

65.7267069
97.4884608

$$= -0.674$$

4) Cal the (c) of correlation for the following ages of husbands & wifes.

Age of (H) (x) = 23 27 28 29 30 31 33 35 36 39
Age of (W) (y) = 18 22 23 24 25 26 28 29 30 32

x	y	$x_i = x - \bar{x}$	$y_i = y - \bar{y}$	x_i^2	y_i^2	$x_i y_i$
23	18	-8.1	-7.7	65.61	59.29	62.37
27	22	-4.1	-3.7	16.81	13.69	15.187
28	23	-3.1	-2.7	9.61	7.29	8.37
29	24	-2.1	-1.7	4.41	2.89	3.57
30	25	-1.1	-0.7	1.21	0.49	0.77
31	26	-0.1	0.3	0.01	0.09	-0.03
33	28	1.9	2.3	3.61	5.29	4.37
35	29	3.9	3.3	15.21	10.89	12.87
36	30	4.9	4.3	24.01	18.49	21.07
39	32	7.9	6.3	62.41	39.69	49.77
				202.9	158.1	178.36

$$\bar{x} = \frac{311}{10} = 31.1$$

$$\bar{y} = \frac{257}{10} = 25.7$$

$$r = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}} = \frac{178.36}{\sqrt{202.9}\sqrt{158.1}} = \frac{178.36}{14.2442971 \times 12.5737822}$$

$$= \frac{178.36}{179.10469} = \underline{0.995}$$

5) Cal co of correlation.

x : 6 2 10 4 8

y : 9 11 5 8 7

A) $$\gamma = \frac{n\,\Sigma xy - \Sigma x\,\Sigma y}{\sqrt{n\,\Sigma x^2 - (\Sigma x)^2}\sqrt{n\,\Sigma y^2 - (\Sigma y)^2}}$$

x	y	x^2	y^2	x y
6	9	36	81	54
2	11	4	121	22
10	5	100	25	50
4	8	16	64	32
8	7	64	49	56
30	40	220	340	214

n=5

$$\gamma = \frac{n\,\Sigma xy - \Sigma x\,\Sigma y}{\sqrt{n\,\Sigma x^2 - (\Sigma x)^2}\sqrt{n\,\Sigma y^2 - (\Sigma y)^2}}$$

$$= \frac{5 \times 214 - 30 \times 40}{\sqrt{5 \times 220 - (30)^2}\sqrt{5 \times 340 - (40)^2}} = \frac{1070 - 1200}{\sqrt{200}\sqrt{100}}$$

$$= \frac{-130}{14.14213562 \times 10} = \frac{-130}{141.4213562}$$

$$= \underline{-0.919}$$

6)

Student	1	2	3	4	5	6	7	8	9	10
Marks in Maths	78	36	98	25	75	82	90	62	65	39
Marks in Stat	84	51	91	60	68	62	86	58	53	47

Cal. Rank correlation coefficient.

A)

Roll no	Mathematics		Statistics		di	di^2
	Mrk	Rank	Mrk	Rank		
1	78	4	84	3	1	1
2	36	9	51	9	0	0
3	98	1	91	1	0	0
4	25	10	60	6	4	16
5	75	5	68	4	1	1
6	82	3	62	5	-2	4
7	90	2	86	2	0	0
8	62	7	58	7	0	0
9	65	6	53	8	-2	4
10	39	8	47	10	-2	4
						30

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

$$= 1 - \frac{6 \times 30}{10(10^2-1)}$$

$$= 1 - \frac{180}{10 \times 99}$$

$$= 1 - \frac{180}{990} = 1 - 0.1818$$

$$= \underline{0.8181}$$

⇒ Spearman's formula for repeated ranks :-

If in a series 2|more individuals have the same score when we find the avg of the ranks of these individuals ie allote these avg rank to each of them.

eg = 98 → 7, 8, 9. ⇒ The score 98 occurs 3 times, there is a time of 7, 8, 9 place.

$$= \frac{7+8+9}{3} = \frac{24}{3} = \underline{\underline{8}}$$

$$\boxed{S\cdot\text{formula} = 1 - \frac{6 \Sigma d^2 + \frac{t^3 - t}{12}}{n(n^2-1)}}$$

1) find rank (c) coefficient —

Series A :	115	109	112	87	98	120	98	100	98	118
Series B :	75	73	85	70	76	82	65	73	68	80

as

Series A		Series B		$d_i = x - y$	d_i^2
Score	Rank(x)	Score	Rank(y)		
115	3	75	5	−2	4
109	5	73	6.5	−1.5	2.25
112	4	85	1	3	9
87	10	70	8	2	4
9th – 98	8	76	4	4	16
120	1	82	2	1	1
8th – 98	8	65	10	2	4
100	6	73	6.5	0.5	0.25
7th – 98	8	68	9	1	1
118	2	80	3	−1	1
					42.5

based on B. 73 → 2 times = $\frac{6+7}{2}$ = 6.5

98 occurs 3 times $\frac{7+8+9}{3}$ = 8

$$R = 1 - \frac{6\left(\sum d^2 + \frac{t^3-t}{12}\right)}{n(n^2-1)}$$

(98) $t=3$, $\frac{3^3-3}{12} = \frac{24}{12} = 2$

(73) $t=2$, $\frac{2^3-2}{12} = 6/12 = 0.5$

$$\sum \frac{t^3-t}{12} = 2+0.5 = 2.5$$

$$R = 1 - \frac{6\times[42.5+2.5]}{10(10^2-1)}$$

$$= \frac{1-\cancel{6}[45]}{10\times 99} = 1 - \frac{3}{11}$$

$$= \frac{11-3}{11} = \frac{8}{11} = 0.72$$

$$= 1 - \frac{270}{990} = 0.72$$

2) Cal Rank (C) coefficient from following data specifying the ranks of 7 students in 2 subjects.

(xi) Rank in 1st sub :	1.	2	3	4	5	6	7
(yi) Rank in 2nd sub :	4	3	1	2	6	5	7

3) The coefficient of rank (C) of marks obtained by 10 students in 2 subjects was computed as 0.5. It was later discovered that the differencing marks in 2 subjects obtained by 1 of the students was wrongly taken as 3 instead of 7. find the correct coefficient of rank (C)?

A) $R = 0.5$

$n = 10$

$$R = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

$$0.5 = 1 - \frac{6\sum(x-y)^2}{10\times 99}$$

$$0.5 = 1 - \frac{6\sum(x-y)^2}{990}$$

$$(0.5-1)\,990 = 6\sum(x-y)^2$$

$$989.5 = 6\sum(x-y)^2$$

$$0.5 = \frac{990 - 6\sum(x-y)^2}{990}$$

$$0.5\times 990 - 990 = -6\sum(x-y)^2$$

$$-495 = -6\sum(x-y)^2$$

$$\sum(x-y)^2 = \frac{495}{6} = 82.5$$

$$\Rightarrow 82.5 - \overset{\text{wrong}}{3^2} + \overset{\text{correct}}{7^2}$$

$$= 82.5 - 9 + 49 = 122.5$$

$$R = 1 - \frac{6\times 122.5}{990}$$

$$= \frac{990 - 6\times 122.5}{990} = \frac{990-735}{990}$$

$$R = 0.2575$$

$d = x_i - y_i$

7 ✓

3 ✓

$(x_i-y_i)^2 = 82.5$

$82.5 - 3^2 + 7^2$

122.5

$\sum (x_i-y_i)^2$

~~⇒ Sum rule~~

2.A)

x_i	y_i	$d_i = x_i - y_i$	d_i^2
1	4	-3	9
2	3	-1	1
3	1	2	4
4	2	2	4
5	6	-1	1
6	5	1	1
7	7	0	0
			20

$$R = 1 - \frac{6 \Sigma d_i^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 20}{7(48)}$$

$$= 1 - \frac{120}{336} = 1 - 0.357 = \underline{\underline{0.643}}$$

2 (R) (coe) cannot be > 1 ~~%~~

It is not possible that 1 regression (coe) is -ve while the other is +ve. The value of (coe) of (corr) cannot exceed unity

∴ If 1 of the (R) (coe) is > than unity, the other must be < than unity

$$\cancel{\frac{1+3x^2 - 12x^2}{2\sqrt{x}(1+3x)^2}} = \cancel{\frac{9x^2-1}{2\sqrt{x}(1+3x^2)^2}}$$

⇒ <u>Regression</u> :- (R)

Suppose we are given n pairs of values of 2 variables x & y. If we fit a straight line to this data by taking x as independent variable & y as dependent variable, then the straight line obtained → <u>regression line of y on x</u> ≈ regression line of x on y. The reciprocal of its slope → <u>regression coefficient</u> of x on y.

→ <u>Eq for (R) lines</u> :-

let $y = a + bx$ —①

be the eq of (R) line of y on x.

a & b are determined by solving the normal eq obtained by the principle of least □'s.

$\Sigma y_i = na + b\Sigma x_i$ —②

$\Sigma x_i y_i = a\Sigma x_i + b\Sigma x_i^2$ —③

② ÷ n.

$$\frac{\Sigma y_i}{n} = a + \frac{b \Sigma x_i}{n}$$

$$\bar{y} = a + b\bar{x} \quad —④$$

$\bar{x}, \bar{y}$ are the means of x & y series

2.A)

④ - ⑥

$$y - \bar{y} = bx - b\bar{x}$$

$$y - \bar{y} = b(x - \bar{x}) \quad \text{—⑤}$$

② $\times \Sigma x_i$ ③ $\times n$.

$$\Sigma x_i y_i = n\cancel{a}\,\Sigma x_i + b\Sigma x_i^2 \quad \text{—⑥}$$

$$n\Sigma x_i y_i = n\cancel{a}\,\Sigma x_i + nb\Sigma x_i^2 \quad \text{—⑦}$$

⑦ - ⑥

$$n\Sigma x_i y_i - \Sigma x_i y_i = nb\Sigma x_i^2 - b\Sigma x_i^2.$$

$$= b\left(n\Sigma x_i^2 - \Sigma x_i^2\right)$$

$$b = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{n\Sigma x_i^2 - \Sigma x_i^2} \rightarrow (\div n^2)$$

$$= \frac{\frac{\cancel{n}\Sigma x_i y_i}{n^2} - \frac{\Sigma x_i \Sigma y_i}{n^2}}{\frac{\cancel{n}\Sigma x_i^2 - \Sigma x_i^2}{n^2}} = \frac{\frac{\Sigma x_i y_i}{n} - \frac{\Sigma x_i}{n}\frac{\Sigma y_i}{n}}{\frac{\Sigma x_i^2}{n} - \left(\frac{\Sigma x_i}{n}\right)^2}$$

$$b = \frac{\frac{\Sigma x_i y_i}{n} - \bar{x}\bar{y}}{\frac{\Sigma x_i^2}{n} - \bar{x}^2} = \frac{cov(x,y)}{\sigma_x^2} = \frac{P_{xy}}{\sigma_x^2} \quad \text{—⑦}$$

Sub ⑦ to ⑤

The ~~st~~ (R) eq of y on x as —

$$y - \bar{y} = \frac{P_{xy}}{\sigma_x^2}(x - \bar{x}) \quad \text{—⑧}$$

Simultaneously when x is depending on y, the (R) eq of x on y is obtained as —

$$(x - \bar{x}) = \frac{P_{xy}}{\sigma_y^2}(y - \bar{y}) \quad \text{—⑨}$$

let us denote $\frac{P_{xy}}{\sigma_x^2}$ as b_{xy} &

$\frac{P_{yx}}{\sigma_y^2}$ as b_{yx}

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

Here b_{yx} is → (R) coefficient of y on x and b_{xy} → (R) coefficient of x on y.

The (R) eq ~~of~~ of y on x —

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

The (R) eq of x on y —

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

Remarks :-

* Slope of (R) line of y on x, $\boxed{b_{xy} = \frac{r\sigma_y}{\sigma_x}}$

Slope of (R) line of x on y is reciprocal of $b_{xy} \to \frac{\sigma_y}{r\sigma_x}$.

* Since $\boxed{b_{yx} = r(\sigma_y / \sigma_x)}$ & σ_x & σ_y are +ve, it follows that r has same sign as that of b_{yx}

2.A)

* since $b_{xy} = r(\sigma_x/\sigma_y)$ we readily find that $(b_{yx})(b_{xy}) = r^2$.
* Since AM is always > GM for any 2 no... we have —

$$\frac{1}{2}(b_{yx} + b_{xy}) = \sqrt{b_{yx} \times b_{xy}} = |r|.$$

$\therefore$ AM of b_{xy} & b_{yx} is always > ~~coef~~ coefficient of (C).

* 2 lines of (R) always pass through the point $(\bar{x}, \bar{y})$
* (R) eq of y on x is need for estimating the value of y for a given value of x & the (R) eq of x on y is used for estimating x for a specified value of y.

correlation	Regression.
* (C) means the relationship b/w 2/more variables	(R) means act of returning to the avg value.
* (C) measures the degree of relationship b/w the variables	(R) measures the ntre of relationship b/w variables.
* There may be nonsense (C) b/w 2 variables	No such nonsense (R)
* very useful for further mathematical treatment.	used for further mathematical treatment.

i) Find the most likely price in Bombay corresponding to the price ₹=70 at calcutta. —

Avg price in calcutta 65, avg price at Bombay 67, SD at calcutta 2.5, SD at Bombay 3.5, (co) of correlation b/w the prices in 2 cities is 0.8.

a) $x \rightarrow$ price of calcutta

$y \rightarrow$ " of Bombay.

$\bar{x} = 65$ $\quad \bar{y} = 67,$ $\quad r = 0.8.$

$\sigma_x = 2.5$ $\quad \sigma_y = 3.5.$

$\therefore$ eq of line of (R) of y on x is,

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

$$y - 67 = \frac{0.8 \times 3.5}{2.5}(x - 65)$$

$$y - 67 = 1.12\,(x - 65)$$

when $x = 70$ $\quad \therefore y = 72.6.$

$(y-67) = 1.12(70-65)$

$y-67 = 1.12 \times 5$

$y-67 = 5.6$

$y = 5.6 + 67$

$y = 72.6$

thus most likely price in Bombay corresponding to the price of Rs. 70 at calcutta is Rs 72.6'

2.A)

2) If x_i & y_i are devi- of n pairs of values of x & y from their respective means (i.e if $x_i = x_i - \bar{x}$ & $y_i = y_i - \bar{y}$.

P. that,

a) $r = 1 - \frac{1}{2n} \Sigma\left(\frac{x_i}{\sigma_x} - \frac{y_i}{\sigma_y}\right)^2$

b) $r = -1 + \frac{1}{2n} \Sigma\left(\frac{x_i}{\sigma_x} - \frac{y_i}{\sigma_y}\right)^2$

hence deduce that $-1 \le r \le +1$.

A) a)

$$\Sigma\left(\frac{x_i}{\sigma_x} - \frac{y_i}{\sigma_y}\right)^2 = \Sigma\left[\frac{x_i^2}{\sigma_x^2} - \frac{2x_iy_i}{\sigma_x\sigma_y} + \frac{y_i^2}{\sigma_y^2}\right]$$

$(a-b)^2$

$\boxed{x_i^2 = n\sigma_x^2 \quad y_i^2 = n\sigma_y^2 \quad \frac{x_iy_i}{\sigma_x\sigma_y} = nr}$

$$= \frac{n\sigma_x^2}{\sigma_x^2} - 2nr + \frac{n\sigma_y^2}{\sigma_y^2}$$

$$\Sigma\left(\frac{x_i}{\sigma_x} - \frac{y_i}{\sigma_y}\right)^2 = 2n - 2nr = 2n(1-r)$$

$$2n(1-r) \to (2n - 2nr)$$

$$2nr = 2n - \Sigma\left(\frac{x_i}{\sigma_x} - \frac{y_i}{\sigma_y}\right)^2 \quad (\div 2n)$$

$$r = 1 - \frac{1}{2n} \Sigma\left(\frac{x_i}{\sigma_x} - \frac{y_i}{\sigma_y}\right)^2$$

b) $r = -1 + \frac{1}{2n} \Sigma\left(\frac{x_i}{\sigma_x} - \frac{y_i}{\sigma_y}\right)^2$

$$= -1 + \frac{1}{2n}(2n + 2nr) = r.$$

~~which is~~ since $\frac{1}{2n} \Sigma\left(\frac{x_i}{\sigma_x} - \frac{y_i}{\sigma_y}\right)^2$ & $\frac{1}{2n} \Sigma\left(\frac{x_i}{\sigma_x} - \frac{y_i}{\sigma_y}\right)^2$ are non -ve.

$\therefore r \le 1$ & $r \ge -1$

(i.e) $-1 \le r \le +1$

3) If r is (C) of correlation b/w n pairs of values of x & y & $\sigma_x, \sigma_y, \sigma_{x-y}$ are SD of x, y & (x-y), P. that.

$$r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x\sigma_y}$$

A) let $u_i = x_i - y_i$ $i = 1, 2, \ldots, n.$

$\bar{u} = \bar{x} - \bar{y}$ where $\bar{x}, \bar{y}, \bar{u}$ are mean of x, y, u.

$u_i - \bar{u} = (x_i - \bar{x}) - (y_i - \bar{y})$ (or)

$(u_i - \bar{u})^2 = (\underset{a}{x_i - \bar{x}})^2 - (\underset{b}{y_i - \bar{y}})$ $(a-b)^2$

$= (x_i - \bar{x})^2 - 2(x_i - \bar{x})(y_i - \bar{y}) + (y_i - \bar{y})^2$

(given Σ) →
$\cancel{\Sigma(u_i - \bar{u})^2 = \Sigma 2(x_i - \bar{x})^2 - \Sigma(y_i - \bar{y})^2 \Sigma(y_i - \bar{y})^2}$

$\cancel{\Sigma(u_i - \bar{u})^2 = (x_i - \bar{x})^2 + (y_i - y)^2 - 2(x_i - \bar{x})(y_i - y)}$

(given Σ) →
$\Sigma(u_i - \bar{u})^2 = \Sigma(x_i - \bar{x})^2 + \Sigma(y_i - y)^2 - 2\Sigma(x_i - \bar{x})(y_i - \bar{y})$.

(or) $n\sigma_u^2 = n\sigma_x^2 + n\sigma_y^2 - 2\Sigma(x_i - \bar{x})(y_i - \bar{y})$

(i.e) $\sigma_{x-y}^2 = \sigma_x^2 + \sigma_y^2 - 2\dfrac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n}$

$\sigma_{x-y}^2 = \sigma_x^2 + \sigma_y^2 - 2r\sigma_x\sigma_y$

$\Rightarrow r = \dfrac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x\sigma_y}$

Q) If $z = ax + by$ & r is (c) of correlation b/w x & y, s.that —

$\sigma_z^2 = a^2\sigma_x^2 + b^2\sigma_y^2 + 2abr\sigma_x\sigma_y$ — (1)

Deduce that,

$\sigma_{x-y}^2 = \sigma_x^2 + \sigma_y^2 - 2r\sigma_x\sigma_y$ — (2)

$\sigma_{x+y}^2 = \sigma_x^2 + \sigma_y^2 + 2r\sigma_x\sigma_y$. — (3)

s.that σ_{x+y} is > or < than σ_{x-y} according as r is +ve / -ve.

A) $z = ax + by$ $\bar{z} = a\bar{x} + b\bar{y}$

let $z_i = ax_i + by_i$

$z_i - \bar{z} = a(x_i - \bar{x}) + b(y_i - \bar{y})$

$(z_i - \bar{z})^2 = a^2(x_i - \bar{x})^2 + b^2(y_i - \bar{y})^2 + 2ab(x_i - \bar{x})(y_i - \bar{y})$

Σ to all → $\Sigma(z_i - \bar{z})^2 = a^2\Sigma(x_i - \bar{x})^2 + b^2\Sigma(y_i - \bar{y})^2 + 2ab\,\Sigma(x_i - \bar{x})(y_i - \bar{y})$

(or)

$\Sigma(z_i - \bar{z}) = n\sigma_z$

$n\sigma_z^2 = na^2\sigma_x^2 + nb^2\sigma_y^2 + 2ab\,\Sigma(x_i - \bar{x})(y_i - \bar{y})$

$\sigma_z^2 = a^2\sigma_x^2 + b^2\sigma_y^2 + 2ab\dfrac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n}$

$\sigma_z^2 = a^2\sigma_x^2 + b^2\sigma_y^2 + 2abr\sigma_x\sigma_y$

— (1) proved

$a = 1 \quad b = -1 \quad —(1)$

$$\sigma_x^2 + \sigma_y^2 - 2r\sigma_x\sigma_y$$

$a = -1 \quad b = 1 \quad —(2)$

$$\sigma_x^2 + \sigma_y^2 + 2r\sigma_x\sigma_y$$

$\sigma_{x+y} > < \sigma_{x-y}$

$$\sigma_x^2 + \sigma_y^2 + 2r\sigma_x\sigma_y \geq \sigma_x^2 + \sigma_y^2 - 2r\sigma_x\sigma_y$$

$$r > \text{ or } < 0.$$

Thus $\sigma_{x+y} > \text{ or } < \sigma_{x-y}$ according as r is +ve / -ve.

5) If θ is acute angle b/w 2 (R) lines relating the variables x & y s.t. that —

$$\boxed{\tan\theta = \frac{(1-r^2)}{r} \cdot \frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}}$$

Indicate the significance of the cases $r = 0$ & $r = \pm 1$.

A) $b_{yx} = \frac{r\sigma_y}{\sigma_x} \qquad b_{xy} = \frac{1}{r}\frac{\sigma_y}{\sigma_x}$

$$\boxed{\tan\theta = \frac{b_{xy} - b_{yx}}{1 + b_{yx} b_{xy}}}$$

Side

$$= \frac{\frac{1}{r}\frac{\sigma_y}{\sigma_x} - \frac{r\sigma_y}{\sigma_x}}{1 + \frac{r\sigma_y}{\sigma_x} \cdot \frac{1}{r}\frac{\sigma_y}{\sigma_x}}$$

$r \cdot \frac{1}{r}$

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

$$= \frac{\frac{1}{r}\frac{\sigma_y}{\sigma_x} - \frac{r\sigma_y}{\sigma_x}}{1 + \sigma_y^2/\sigma_x^2}$$

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

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

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

when $r = 0$, $\theta = \frac{\pi}{2}$

when $r = \pm 1$ $\quad \frac{1-1}{1} = 0$

$\theta = 0$

$\therefore$ the lines of (R) coincide