UNIT-5

CURVE FITTING, (REGRESSION) & CORRELATION

→ Introduction to Curve fitting:

1) Defⁿ of curve fitting
2) Best fit for the curve
*3) Principle of least square / method of least square (derivation). (SAQ) (2M)
4) Normal Equations of st. line (linear equation)
5) Fitting of st. line (4M)
6) Normal equations of parabola (2nd degree polynomial
7) Fitting of parabola
8) Normal equations of exponential curves.
9) Fitting of exponential curves.

→ Introduction to Correlation:

1) Defⁿ of correlation - properties*
2) Karl Pearson's ~~formula~~ coefficient - Problems.

Principle of least square: principle of least squares is probably the most systematic procedure to fit a unique curve through the given points.

Let $y = f(x)$ be the equation of the curve to be fitted to the given data (observed or experimental) points $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$

At $x = x_1$; the observed/experimental value of the ordinate is y_1 and the corresponding value on the fitting curve is N_1M_1 i.e. $f(x_1)$. The difference of the observed & expected value is

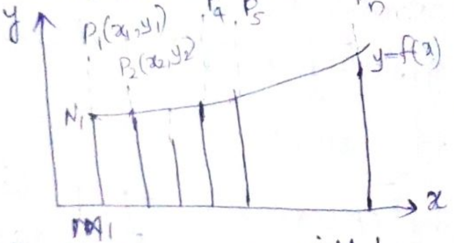
$$\boxed{P_1M_1 - N_1M_1 = P_1N_1 = e_1}$$

The difference is called the error.

$e_1 = y_1 - f(x_1)$

$e_2 = y_2 - f(x_2)$

$e_n = y_n - f(x_n)$



Some of the errors $e_1, e_2, \ldots e_n$ will be +ve and others -ve

In finding total errors, errors are added. In addition: some -ve & +ve errors may be cancelled and gives zero too, which lead to false result.

To avoid such situation; we may make all the errors +ve by squaring.

$$\boxed{Sum = S = e_1^2 + e_2^2 + \cdots + e_n^2}$$

The curve of the best fit is that for which the sum of the squares of errors (S) is minimum. This is called the principle of least squares.

*CURVE FITTING:

→ **Def^n of curve fitting**: Suppose (x_i, y_i); $i = 1, \ldots n$ be the given data of n pairs of values in terms of 2 variables. x being the independent variable. The curve fitting is to find an analytic expression of the form $y = f(x)$ which fits the given data.

→ **Best Fitting Curve**: Let the given data be represented by a set of n points (x_i, y_i), $i = 1, 2, 3, \ldots n$. Let $y = f(x)$ be an approximate curve which fits the given set of data.

Let $y_i = f(x)$ then y_i is called the expected value of y corresponding to $x = x_i$

The value of y_i is called the observed value of y corresponding to $x = x_i$

→ In general $y_i \neq y_i$ as the part (x_i, y_i) is called the error of estimate (or) the residue for y_i. For any given set of points the curve for which

$$R = R_1^2 + R_2^2 + \cdots + R_n^2$$
$$= \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

is minimum is called the best fitting curve (or) least square curve.

*Method of least squares:

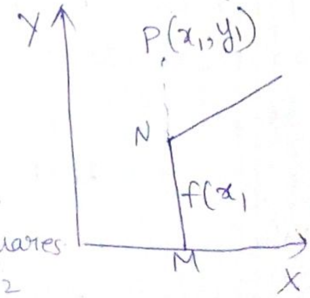
Let $y = a + bx$ be the straight line to be fitted to the given data points (x_1, y_1), $(x_2, y_2) \ldots \ldots (x_n, y_n)$.

Let $f(x_1)$ be the expected (theoretical) ordinate for x_1.

$PN = y_1$, $NM = f(x_1)$ $\quad (\because PN = e_1)$

$PN = PM - NM \Rightarrow e_1 = y_1 - f(x_1)$

$e_1 = y - (a + bx_1)$



On squaring

$e_1^2 = (y - a - bx_1)^2$

By the principle of least squares

$S = e_1^2 + e_2^2 + \cdots + e_n^2$

$S = (y_1 - a - bx_1)^2 + (y_2 - a - bx_2)^2 + \cdots + (y_n - a - bx_n)^2$

For S to be minimum

$$\frac{\partial S}{\partial a} = 2(y_1 - a - bx_1)(-1) + 2(y_2 - a - bx_2)(-1) + \cdots 2(y_n - a - bx_n)(-1) = 0$$

$(y_1 - a - bx_1) + (y_2 - a - bx_2) + \cdots + (y_n - a - bx_n) = 0$

$(y_1 + y_2 + \cdots + y_n) - na - b(x_1 + x_2 + \cdots x_n) = 0$

$$\boxed{\Sigma y = na + b\Sigma x} \quad \text{—} ①$$

$$\frac{\partial S}{\partial b} = 2(y_1 - a - bx_1)(-x_1) + 2(y_2 - a - bx_2)(-x_2) + \cdots + 2(y_n - a - bx_n)(-x_n) = 0$$

$= (x_1 y_1 - x_1 a - bx_1^2) + (x_2 y_2 - x_2 a - x_2^2) + \cdots + (x_n y_n - x_n a - x_n^2) = 0$

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

① & ② are called normal equations.

Solving ① & ②, we get values of a & b, putting a & b we get st. line equation

*NOTE:

The other form of st. line is $y = ax + b$

Its normal equations are $\Sigma y = a\Sigma x + nb$

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

* Fit a st. line by the method of least squares:

1) $y = ax+b$

$\Sigma y = a\Sigma x + nb$

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

2) $v = a_0 t + a_1$

$\Sigma v = a_0 \Sigma t + n a_1$

$\Sigma tv = a_0 \Sigma t^2 + a_1 \Sigma t$

3) $y = a + bx + cx^2$

$\Sigma y = na + b\Sigma x + c\Sigma x^2$

$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3$

$\Sigma x^2 y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4$

* $x = 1, 2, 3, 4, 5$

$y = 14, 27, 40, 55, 68$

$y = ax + b$

Consider $y = ax+b$.

The Normal equations are:

$\Sigma y = a\Sigma x + nb$

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

x	y	x^2	xy
1	14	1	14
2	27	4	54
3	40	9	120
4	55	16	220
5	68	25	340
$\Sigma x =$ 15	$\Sigma y =$ 204	$\Sigma x^2 =$ 55	748

$204 = 15a + 5b$

$748 = 55a + 15b$

$a = 13.6 \quad b = 0.$

$\boxed{y = 13.6x}$ → St line equation

27/05/2023

* Fit a st. line for the following data:

$x = 27, 45, 41, 19, 3, 39, 19, 49, 15, 31$

$y = 57, 64, 80, 46, 62, 72, 52, 77, 57, 68$

$y = a + bx.$

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

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

x	x^2	y	xy
27	729	57	1539
45	2025	64	2880
41	1681	80	3280
19	361	46	874
3	9	62	186
39	1521	72	2808
19	361	52	988
49	2401	77	3773
15	225	57	855
31	961	68	2108

$\Sigma x = 288$

$\Sigma x^2 = 10274$

$\Sigma y = 635$

$\Sigma xy = 19291$

$635 = 10a + 288b$

$19291 = 288a + 10274b$

$a = 48.90$

$b = 0.5$

$\boxed{y = 48.9 + 0.5x}$

Thus the least square st. line y on x $\boxed{y = 48.9 + 0.5x}$ is the required best fit for given data.

* Fit a st. line to the curve $y = a + bx$ passing through the points $(0,-1), (2,5), (5,12), (7,20)$

* The temperature T in °C and length L in mm of heated rod are given below $\boxed{L = a_0 + a_1 T}$. Find the best values for a_0 & a_1

T	L
20	800.3
30	800.4
40	800.6
50	800.7
60	800.9
70	801.0

$L_0 = a_0 + a_1 T$

$\Sigma L_0 = n a_0 + a_1 \Sigma T$

$\Sigma L_0 T = a_0 \Sigma T + a_1 \Sigma T^2$

T	L	T^2	LT
20	800.3	400	16006
30	800.4	900	24012
40	800.6	1600	32024
50	800.7	2500	40035
60	800.9	3600	48054
70	801.0	4900	56070

$\Sigma T = 270$

$\Sigma L = 4803.90$

$\Sigma T^2 = 13900$

$\Sigma LT = 216201$

$4803.90 = 6a_0 + 270T$

$216201 = 270a_0 + 13900a_1$

$a_0 = 799.9$

$a_1 = 0.01$

$\boxed{L = 799.9 + 0.01T}$

* $y = a + bx$

x	y	x^2	xy
0	-1	0	0
2	5	4	10
5	12	25	60
7	20	49	140

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

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

$\Sigma x = 14$ $\Sigma x^2 = 78$

$\Sigma y = 36$ $\Sigma xy = 210$

$36 = 4a + 14b$

$210 = 14a + 78b$

$a = -1.13$

$b = 2.8$

$\boxed{y = -1.13 + 2.8x}$

* $y = a + bx + cx^2$

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

(6) $y = ax^2 + bx + c$ — (1) Second degree polynomial is

Its normal eqns are — 3-constants

$\Sigma y = a \cdot \Sigma x^2 + b \cdot \Sigma x + nc$ — (2)

$\Sigma xy = a \Sigma x^3 + b \Sigma x^2 + c \Sigma x$ — (3)

$\Sigma x^2 y = a \Sigma x^4 + b \Sigma x^3 + c \Sigma x^2$ — (4)

$f(-1) = -2, \; f(0) = 1, \; f(1) = 2, \; f(2) = 4$

$f(x) = y$

x	y	x^2	x^3	x^4	xy	x^2y
-1	-2	1	-1	1	2	-2
0	1	0	0	0	0	0
1	2	1	1	1	2	2
2	4	4	8	16	8	16
$\boxed{\Sigma x = 2}$	$\boxed{\Sigma y = 5}$	$\boxed{\Sigma x^2 = 6}$	$\boxed{\Sigma x^3 = 8}$	$\boxed{\Sigma x^4 = 18}$	$\boxed{\Sigma xy = 12}$	$\boxed{\Sigma x^2 y = 16}$

$5 = a(6) + b(2) + 4c$

$12 = a(8) + b(6) + c(2)$

$16 = a(18) + b(8) + c(6)$

$a = -0.25, \quad b = 2.15, \quad c = 0.55$

$y = (-0.25)(x^2) + (2.15)(x) + 0.55$

Q) Find the parabola of the form $y = ax^2 + bx + c$ passing through the points $(-1,2)$, $(0,1)$, $(1,4)$

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

$$\Sigma y = a\Sigma x^2 + b\Sigma x + nc \quad \text{---} (2)$$

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

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

x	y	x^2	x^3	x^4	xy	x^2y
-1	2	1	-1	1	-2	2
0	1	0	0	0	0	0
1	4	1	1	1	4	4
$\Sigma x = 0$	$\Sigma y = 7$	$\Sigma x^2 = 2$	$\Sigma x^3 = 0$	$\Sigma x^4 = 2$	$\Sigma xy = 2$	$\Sigma x^2y = 6$

$\boxed{n = 3}$

$$7 = a(2) + b(0) + 3c$$

$$2 = a(0) + b(2) + c(0)$$

$$6 = a(2) + b(0) + c(2)$$

$a = 2$, $b = 1$, $c = 1$

~~$y = 2a^2$~~ $\boxed{y = 2x^2 + x + 1}$

Fit an exponential curve

$$y = ae^{bx} \quad (\text{OR}) \quad y = ab^x$$

Taking log on b/s.

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

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

$$Y = a_0 + a_1 x \quad \text{---} (2)$$

where $Y = \log y$

$X = x$

$a_0 = \log a$, $b = a_1$

$a = e^{a_0}$

eqn (2) of the form st. line.

Normal ~~line~~ eqn of (2) line.

$$\Sigma Y = na_0 + a_1 \Sigma x \quad \text{---} (3)$$

$$\Sigma xY = a_0 \Sigma x + a_1 \Sigma x^2 \quad \text{---} (4)$$

Solving (3) and (4)

we get a_0, a_1

Find a from a_0

$$a = e^{a_0}$$

Find b from a_1

$\boxed{b = a_1}$

Finally substitute a, b's in (2)

$y = (\ )e^{(\ )x}$

Fit an exponential ~~for~~ curve of the form $Y = ae^{bx}$ to ~~the~~ the following data by the method of least squares.

$Y = ae^{bx}$

x	y	$Y = \log_e y$	xy	x^2
1	10	2.3025	2.3025	1
5	15	2.7080	13.54	25
7	12	2.4849	17.3943	49
9	15	2.7080	24.372	81
12	21	3.0145	36.534	144
$\Sigma x = 34$	$\Sigma y = 73$	$\Sigma Y = 13.2479$	$\Sigma xy = 94.0028$	$\Sigma x^2 = 300$

$13.2479 = 5(a_0) + a_1(34)$

$94.0028 = a_0(34) + a_1(300)$

$a_0 = 2.2485 \quad a_1 = 0.0589$

$b = a_1$

$b = 0.0589$

$a = e^{a_0} \quad a \Rightarrow e^{2.2485}$

$= 9.4735$

~~$y = (9.4735)(0.0589)^x$~~

$y = (9.4735)\,e^{(0.0589)x}$

(2) $\boxed{Y = ae^{bx}}$ — (1)

sol) $\Sigma y = na_0 + a_1\Sigma x$ — (1)

$\Sigma xy = a_0\Sigma x + a_1\Sigma x^2$ — (2)

x	y	$Y = \log_e y$	xy	x^2
0.0	0.10	−2.3025	0	0
0.5	0.45	−0.7985	−0.3992	0.25
1.0	2.15	0.7654	0.7654	1
1.5	9.15	2.2137	3.3205	2.25
2.0	40.35	3.6975	7.395	4
2.5	180.75	5.1971	12.9927	6.25
7.5		8.772	24.0744	13.75

$8.772 = 6a_0 + a_1(7.5)$

$24.0744 = a_0(7.5) + a_1(13.75)$

$a_0 = -2.2835$

$a_1 = 2.9964$

$a = e^{-2.2835} = 0.1019 \quad y = (0.1019)e^{2.9964x}$

$b = 2.9964$

08/06/2023

* $y = ab^{x}$

Apply log on both sides

$\log_{10} y = \log_{10} a + x \log_{10} b$

$Y = A + BX$ — ②

It is of the straight line. Its normal eqns are

$\Sigma Y = nA + B\Sigma X$ — ③

$\Sigma XY = A\Sigma X + B\Sigma X^2$ — ④

from A find a

$A = \log_{10} a \quad B = \log_{10} b \quad X = x$

$a = 10^{A}, \; b = 10^{B}$

Finally substitute a & b in ①

~~$y = (\;)(\;)^{x}$~~

$\boxed{y = (\;)(\;)^{x}}$

* Fit an exponential curve to the following data. Of the form $y = ae^{bx}$.

x	y	$Y = \log_e y$	xy	x^2
77	2.4	0.8754	67.4058	5929
100	3.4	1.2237	122.37	10000
185	7.0	1.9459	359.9915	34225
239	11.1	2.4069	575.2491	57121
285	19.6	2.9755	848.0175	81225

$\Sigma x = 886$ $\quad \Sigma Y = 9.4274$ $\quad \Sigma x^2 = 188500$

$\Sigma XY = 1973.0339$

$9.4274 = 5A + 886B$.

$1973.03 = 886A + 188500B$.

$A = 0.1838$

$B = 9.602 \times 10^{-3}$

$= 0.009602$

$a = e^{A}$ ~~$b = 10^{B}$~~

~~$a = 10^{0.1838}$~~ ~~$b = 10^{0.009602}$~~

~~$a = 1.5268$~~ $\quad B = b = 0.0096$

$a = 1.2017$

$\boxed{y = 1.2017 e^{0.096x}}$

* $y = ab^x$

$X = x$	y	$Y = \log_{10} y$	XY	x^2
0.2	3.16	0.4996	0.0999	0.04
0.3	2.38	0.3765	0.1129	0.09
0.4	1.75	0.2430	0.0972	0.16
0.5	1.34	0.1271	0.0633	0.25
0.6	1.00	0	0	0.36
0.7	0.74	−0.1307	−0.0914	0.49

$\Sigma = 2.7$ $\quad \Sigma Y = 1.1155$

$\Sigma XY = 0.2821 \quad \Sigma x^2 = 1.39$

$1.1155 = 6A + 2.7B$

$0.2821 = 2.7A + 1.39B$

$A = 0.7513 \quad B = -1.2564$

$a = 10^{0.7513} \quad b = 10^{-1.25}$

$a = 5.6402 \quad b = 0.05541$

$\cancel{y = (5.6467)(0.05623)^x}$

$\boxed{y = (5.6402)(0.05541)^x}$

* $y = ab^x$

$X = x$

x	y	$Y = \log_{10} y$	XY	x^2
0	10			
1	21			
2	35			
3	59			
4	92			
5	200			
6	400			
7	610			

09/06/2023

CORRELATION:

If a change in one ~~variation~~ variable causes change in other variable then the relation between 2 variables is said to be correlated. else uncorrelated.

*→ Coefficient of correlation:

- It is a statistical analysis which measures and analyzes the degree between 2 variables and is denoted by r (new).

* NOTE:

→ Correlation coefficient lies between

~~$+1 \leq r \leq +1$~~ $\boxed{-1 \leq r \leq +1}$

→ If correlation coefficient $= -1$ then it is perfect and negative.

$r = +1 \rightarrow$ perfect & +ve

$r = 0 \rightarrow$ uncorrelated.

$r > 0 \rightarrow$ positive correlation

$r < 0 \rightarrow$ negative correlation.

, so r never exceeds 1.

Properties of co-relation:

1) → The maximum value of co-relation co-efficient is 1. or unity.

→ The co-efficient of r lies between -1 and $+1$

$-1 \leq r \leq +1$

$|r| \leq 1$.

2) 'r' is independent of change of origin and scale of measurement.

3) Two independent variables are un-correlated

If x and y are independent variables

$r(x, y) = 0$.

Types of correlations:-

→ correlation is classified into many types

1. +ve & −ve
2. simple (→ 2 variables) & multiple → relation b/w more than 2 variables
3. partial & total
4. linear & non-linear (power = 1)

→ There are 2 different methods for finding out the relationship b/w variables

(i) Graphical Method
- → Scatter diagram
- → Simple graph

(ii) Mathematical Method.
(coefficient)
- → Karl - Pearson's coefficient of correlation
- → Spearman's Rank Coefficient of correlation
- → Coefficient of concurrent deviation
- → Method of least squares.

* Karl - Pearson's Coefficient of Correlation:

If x and y are 2 variables then Karl - Pearson's Coefficient of Correlation is denoted by

$$r_{xy} = \frac{Cov(x,y)}{\sigma_x * \sigma_y}$$

Cov → Co-variance

$\sigma_x \rightarrow$ S.D. of x

$\sigma_y \rightarrow$ S.D of y.

→ Co-Variance (x,y): difference between 2 variables.

$\cancel{Cov(x,y) = \Sigma(x-\bar{x})(y-\bar{y})}$

$$\mathrm{cov}(x,y) = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\frac{\sum (x-\bar{x})^2}{n}} \cdot \sqrt{\frac{\sum (y-\bar{y})^2}{n}}}$$

$X = x - \bar{x}$ $\quad Y = y - \bar{y}$

$\bar{x} = \frac{1}{n}\sum x$ $\quad \bar{y} = \frac{1}{n}\sum y$

$\mathrm{cov}(x,y) = \cancel{\sum X \cdot \sum Y}$

$$\boxed{\mathrm{cov}(x,y) = \frac{\sum XY}{\sqrt{\sum X^2} \cdot \sqrt{\sum Y^2}}}$$

* Estimate coefficient of correlation b/w following data.

XY	x	y	X	Y	X^2	Y^2
10	12	14	2	5	4	25
1	9	8	-1	-1	1	1
6	8	6	-2	-3	4	9
0	10	9	0	0	0	0
2	11	11	1	2	1	4
9	13	12	3	3	9	9
18	7	3	-3	-6	9	36
			0	0		

$\sum x = 70$ $\quad \frac{\sum x}{n} = 10$

$\sum y = 63$ $\quad \frac{\sum y}{n} = 9$

$\sum X = 0$

$\sum Y = 0$

$\sum XY = 45$

$\sum X^2 = 28$

$\sum Y^2 = 84$

$$r = \frac{45}{\sqrt{28} \cdot \sqrt{72}} = \frac{45}{5.29 \times 9.46}$$

$$= \frac{45}{48.456}$$

$$= \underline{\underline{0.946}}$$

$$r = \frac{45}{\sqrt{28}\cdot\sqrt{72}} = \frac{45}{5.29 \times 9.48}$$

$$= \frac{45}{48.456}$$

$$= \underline{\underline{0.946}}$$

Propl: P.T the max value of correlation coeff is 1 (or) P.T the coeff of correlation lies b/w $-1 \le r \le +1$.

Proof: we have karl-pearson's coeff of correlation

$$r(x,y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}} \quad \text{— ①}$$

$$r(x,y) = \frac{\sum a_i b_i}{\sqrt{\sum a_i^2}\sqrt{\sum b_i^2}} \quad \text{— ②}$$

put $x-\bar{x} = a_i$; $y-\bar{y} = b_i$;

suppose we consider $a_i b_i$ are real no's

By Cauchy's schwartz inequality.

$$\sum(a_i b_i)^2 \le \sum a_i^2 \sum b_i^2 \Rightarrow \frac{\sum(a_i b_i)^2}{\sum a_i^2 \sum b_i^2} \le 1 \quad \text{— ③}$$

sub ② in ③

$r^2 \le 1$ i.e $-1 \le r \le 1$ OR $|r| \le 1$

Hence Proved.

Property 2: P.T. 2 independent variables are uncorrelated.

Proof: Let X, Y be two independent variables

$E(X,Y) = E(X)E(Y)$ — ①

By the def of $\mathrm{Cov}(X,Y)$.

$\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y)$ — ②

sub ① in ②

$\mathrm{Cov}(X,Y) = E(X)E(Y) - E(X)E(Y)$

$\mathrm{Cov}(X,Y) = 0$ — ③

By karl-pearson's co-eff of correlation

$$r(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_x \sigma_y} \quad \text{— ④}$$

sub ③ in ④

$$r(X,Y) = \frac{0}{\sigma_x \sigma_y} = 0$$

By the prop of correlation,

if $r = 0$, Then the variables are uncorrelated

∴ X, Y are uncorrelated.

Property 3: P.T the coeff of correlation is independent of change of origin and scale of measurements.

Proof: we have karl-pearson's coeff of correlation.

$$r(x,y) = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}} \quad \text{— ①}$$

$r(x,y)$ By changing the origin & scale of X, Y variables on putting

$\boxed{u_i = \frac{x_i - A}{h}}$ & $\boxed{v_i = \frac{y_i - B}{K}}$

A, B → change of origin
h, k → change of scale

$x_i = A + u_i h$
$y_i = B + v_i K$
$\bar{x} = A + h\bar{u}$
$\bar{y} = B + k\bar{v}$

$x_i - \bar{x} = h(u_i - \bar{u})$

$y_i - \bar{y} = k(v_i - \bar{v})$

$$r(x,y) = \frac{\Sigma(u_i-\bar{u})(v_i-\bar{v})\cancel{hk}}{\sqrt{\cancel{h^2}(u_i-\bar{u})^2}\cdot\sqrt{\cancel{k^2}(v_i-\bar{v})^2}}$$

$$r(x,y) = \frac{\Sigma(u_i-\bar{u})(v_i-\bar{v})}{\sqrt{\Sigma(u_i-\bar{u})^2}\cdot\sqrt{\Sigma(v_i-\bar{v})^2}}$$

Hence proved.

Property 4:

If $z = x - y$ & r is the correlation coefficient b/w x and y then

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

Proof:

Given that $z = x - y$ — (1)

Applying expectation on both sides

$$E(z) = E(x - y)$$

$$E(z) = E(x) - E(y) \quad \text{— (2)}$$

(By the property of $E(ax - by) = aE(x) - bE(y)$)

(1) − (2)

$$z - E(z) = [x - E(x)] - [y - E(y)] \quad \text{— (3)}$$

Squaring & again applying expectation on both sides.

$$E[z - E(z)]^2 = E\{[x - E(x)] - [y - E(y)]\}^2$$

$E^2 = E$

$$= E[x - E(x)]^2 + E[y - E(y)]^2$$

$$- 2E(x - E(x))(y - E(y)).$$

$$\sigma_z^2 = \sigma_x^2 + \sigma_y^2 - 2\,Cov(x, y) \qquad z = x - y$$

$$\sigma_z^2 = \sigma_x^2 + \sigma_y^2 - 2r\,\sigma_x \cdot \sigma_y$$

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