Correlation and Regression

Curve fitting

Two variables are said to be correlated if change in value of one variable appears to be related or linked with the change in the value of the other variable.

Ex 1: Pressure and volume of a gas is said to be correlated as an increase in pressure brings a decrease in volume

Ex 2: An increase in bank interest may lead to an increase in deposit.

Correlation is of two types:

- Positive Correlation or Direct Correlation
- Negative Correlation or indirect Correlation

Correlation is said to be positive or direct if an increase in the value of one variable is associated with an increase in the other variable also. In this case the both the variables change or move in the same direction.

Correlation is said to be negative or indirect if an increase in the value of one variable is associated with a decrease in the other variable also. In this case the both the variables change or move in the opposite direction.

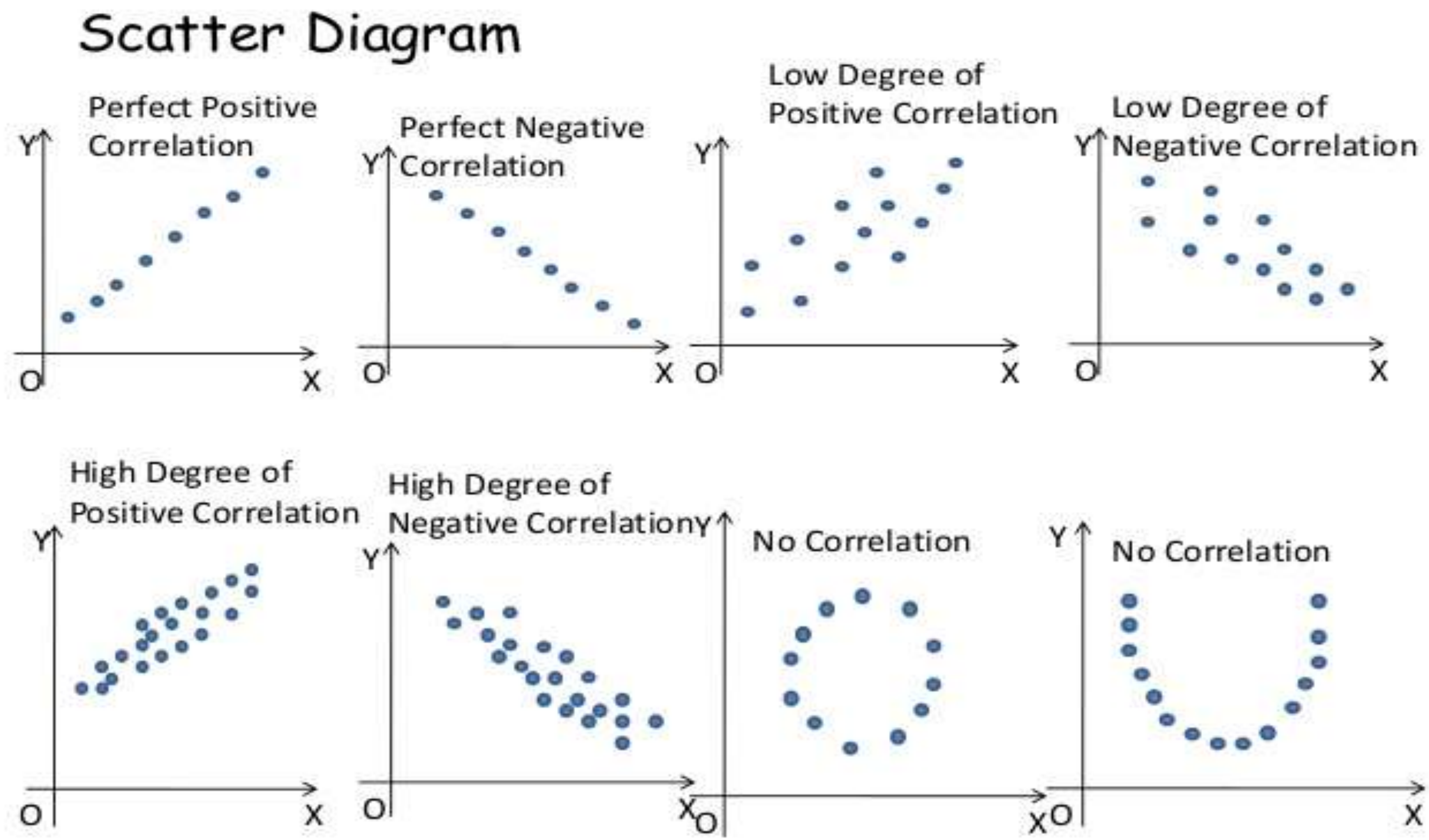
Correlation and Regression

Scatter Diagram: Let X and Y be two variables under consideration. Assume that data is collected from n units of the population regarding these two variables. Let $(x_1, y_1), (x_2, y_2), \ldots\ldots. (x_n, y_n)$ be the n pairs of observations. The diagram obtained by plotting the observations in a two dimensional plane (usually taking the variable X on the horizontal axis and Y on the vertical axis) is called the scatter diagram.

Uses of Scatter Diagram: The points in the scatter diagram can show the simultaneous variations in the values of the variables. The term scatter refers to the dispersion of dots on the graph. If the points in the scatter diagram are very dense, it indicates high degree of correlation, a widely scattered diagram indicate poor correlation. Scatter diagram can be used to identify

- whether there exist any relationship between the variables.
- whether an existing relationship is positive or negative.
- Whether the existing relationship (+ve or –ve) is perfect or strong or weak.
- identify whether the relationship is linear or non linear.

The following shows scatter diagram corresponding to different types of linear relationship between two variables.



Curve fitting: The scatter diagram presents an approximate relationship between the variables under consideration in a non-mathematical way. The exact functional relationship between the variables can be obtained by establishing a mathematical functional relationship between the variables using the given data.

Let y = f($a_{0,}\, a_{1,}\, a_{2,}$ ...., $a_{n,}$x) be the functional relationship between the dependent variable Y and independent variable X. Let $(x_1, y_1), (x_2, y_2), \ldots\ldots. (x_n, y_n)$ be the n pairs of observations. By curve fitting we mean the determination of the best values of $a_{0,}\, a_{1,}\, a_{2,}$ ...., a_n using the given set of observations.

In practice it is difficult to have an ideal situation that all the points falls on the curve. Hence best values of the $a_{0,}\, a_{1,}\, a_{2,}$ ...., a_n are those values for which maximum points in the scatter diagram lie on the curve.

Principle of least squares:

Let y = f($a_0, a_1, a_2, \ldots\ldots\ldots, a_n$, x) be the functional relationship between the dependent variable Y and independent variable X. Let $(x_1, y_1), (x_2, y_2), \ldots\ldots.(x_n, y_n)$ *be* the n pairs of observations. By curve fitting we mean the determination of the best values of $a_0, a_1, a_2, \ldots\ldots\ldots, a_n$ using the given set of observations.

When the value of the independent variable X = x_i, the observed value of Y is y_i and the corresponding value of Y estimated from the functional relationship is given by, $\widehat{y_i}$ = f($a_0, a_1, a_2, \ldots\ldots\ldots, a_n, x_i$).

The difference between y_i and $\widehat{y_i}$ is termed as the error e_i at the point (x_i, y_i) and is given by $e_i = y_i - \widehat{y_i} = \{y_i - f(a_0, a_1, a_2, \ldots\ldots\ldots, a_n, x_i)\}$.

The Principle of least squares states that the best values of the parameters or constants $a_0, a_1, a_2, \ldots\ldots\ldots, a_n$ are those values which minimise the sum of squares of errors.

By this Principle the best values of $a_0, a_1, a_2, \ldots\ldots\ldots, a_n$ are those values which minimize the sum S = $\sum e_i^2 = \sum\{y_i - f(a_0, a_1, a_2, \ldots\ldots\ldots, a_n, x_i)\}^2$

Procedure for fitting y = f($a_0, a_1, a_2, \ldots, a_n$, x) by L S principle:

Let y = f($a_0, a_1, a_2, \ldots, a_n$, x) be the functional relationship between the dependent variable Y and independent variable X. Let $(x_1, y_1), (x_2, y_2), \ldots\ldots (x_n, y_n)$ *be* the n pairs of observations.

When the value of the independent variable X = x_i, the observed value of Y is y_i and the corresponding value of Y estimated from the functional relationship is given by, $\widehat{y_i} = f(a_0, a_1, a_2, \ldots, a_n, x_i)$.

The difference between y_i and $\widehat{y_i}$ is termed as the error e_i at the point (x_i, y_i) and is given by $e_i = y_i - \widehat{y_i} = \{y_i - f(a_0, a_1, a_2, \ldots, a_n, x_i)\}$.

The Principle of least squares states that the best values of the parameters or constants $a_0, a_1, a_2, \ldots, a_n$ are those values which minimise the sum of squares of errors.

By this Principle the best values of $a_0, a_1, a_2, \ldots, a_n$ are those values which minimize the sum $S = \sum e_i^2 = \sum\{y_i - f(a_0, a_1, a_2, \ldots, a_n, x_i)\}^2$

S is a minimum when $\frac{\partial S}{\partial a_0} = 0, \frac{\partial S}{\partial a_1} = 0, \frac{\partial S}{\partial a_2} = 0, \ldots\ldots \frac{\partial S}{\partial a_n} = 0$

Correlation and Regression

Curve fitting

$$We\ have\ S = \sum e_i{}^2 = \sum\{y_i - f(a_0, a_1, a_2, \ldots\ldots\ldots, a_n, x_i)\}^2$$

$$\frac{\partial S}{\partial a_0} = 0 \Longrightarrow \frac{\partial}{\partial a_0}\sum\{y_i - f(a_0, a_1, a_2, \ldots\ldots\ldots, a_n, x_i)\}^2 = 0$$

$$\frac{\partial S}{\partial a_1} = 0 \Longrightarrow \frac{\partial}{\partial a_1}\sum\{y_i - f(a_0, a_1, a_2, \ldots\ldots\ldots, a_n, x_i)\}^2 = 0$$

$$\frac{\partial S}{\partial a_2} = 0 \Longrightarrow \frac{\partial}{\partial a_2}\sum\{y_i - f(a_0, a_1, a_2, \ldots\ldots\ldots, a_n, x_i)\}^2 = 0$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\frac{\partial S}{\partial a_n} = 0 \Longrightarrow \frac{\partial}{\partial a_3}\sum\{y_i - f(a_0, a_1, a_2, \ldots\ldots\ldots, a_n, x_i)\}^2 = 0$$

The above set of (n+1) equations are called the normal equations. Solving the normal equations, we get the best values of $a_0, a_1, a_2, \ldots\ldots\ldots, a_n$.

To fit a straight line of the form *$y = ax + b$ to a given data set.*

Let $(x_1, y_1), (x_2, y_2), \ldots\ldots. (x_n, y_n)$ *be* the n pairs of observations. By curve fitting we mean the determination of the best values of a and b using the given set of observations.

When the value of the independent variable X = x_i, the observed value of Y is y_i and the corresponding value of Y estimated from the functional relationship is given by, $\widehat{y_i} = a\, x_i + b$

The difference between y_i and $\widehat{y_i}$ is termed as the error e_i at the point (x_i, y_i) and is given by $e_i = y_i - \widehat{y_i} = \{y_i - (ax_i + b)\}$.

By the Least Squares Principle, best values of a *and* b are those values which minimize the sum of squares of errors given by,

$$S = \sum {e_i}^2 = \sum\{y_i - (ax_i + b)\}^2$$

----------------------*Contd in next slide*

Correlation and Regression

Curve fitting

To fit $y = ax + b$ ------------Contd from previous slide

$S = \sum e_i^2 = \sum\{y_i - (ax_i + b)\}^2$

S is a minimum when $\frac{\partial S}{\partial a} = 0$ and $\frac{\partial S}{\partial b} = 0$

$\frac{\partial S}{\partial a} = 0 \Longrightarrow 2\sum\{y_i - (ax_i + b)\}^{2-1}(0 - x_i - 0) = 0$

$-2\sum(y_i - ax_i - b)\, x_i = 0 \Longrightarrow \sum x_i y_i = a\sum x_i^2 + bx_i \ldots\ldots(1)$

$\frac{\partial S}{\partial b} = 0 \Longrightarrow 2\sum\{y_i - (ax_i + b)\}^{2-1}(0 - 0 - 1) = 0$

$-2\sum(y_i - ax_i - b) = 0 \Longrightarrow \sum y_i = a\sum x_i + nb \ldots\ldots\ldots(2)$

The normal equations are given by,

$\sum y_i = a\sum x_i + nb \ldots\ldots(2)$

$\sum x_i y_i = a\sum x_i^2 + bx_i \ldots\quad(1)$

Solving (1) and (2) we get a and b.

x_i	y_i	$x_i y_i$	x_i^2
x_1	y_1	$x_1 y_1$	x_1^2
x_2	y_2	$x_2 y_2$	x_2^2
....	....	....	....
x_n	y_n	$x_n y_n$	x_n^2
Σx_i	Σy_i	$\Sigma x_i y_i$	Σx_i^2

Qn 1: Fit a straight line to the data

X	0	1	2	3	4
y	1	1.8	3.3	4.5	6.3

Let the straight line to be fitted be

$y = ax + b$

The normal equations are given by

$\sum y_i = a \sum x_i + nb \ldots\ldots(1)$

$\sum x_i y_i = a \sum x_i^2 + bx_i \ldots \quad (2)$

Substituting from table in

(1) and (2) we get,

x	y	xy	x^2
0	1.00	0.00	0.0
1	1.80	1.80	1.0
2	3.30	6.60	4.0
3	4.50	13.50	9.0
4	6.30	25.20	16.0
10	**16.9**	**47.1**	**30**

$16.9 = \; 5\,a + 10\,b \ldots\ldots..(3)$

$47.1 = 10\,a + 30\,b \ldots\ldots.(4)$

Solving (3) and (4) we get

a = 0.72 and **b = 1.33**

The fitted straight line is y = 0.72 *x* + 1.33

To fit a curve of the form $y = ax^2 + bx + c$ *to a given data set.*

Let $(x_1, y_1), (x_2, y_2), \ldots\ldots. (x_n, y_n)$ *be* the n pairs of observations. By curve fitting we mean the determination of the best values of a and b using the given set of observations.

When the value of the independent variable X = x_i, the observed value of Y is y_i and the corresponding value of Y estimated from the functional relationship is given by, $\widehat{y_i} = a\,x_i^{\,2} + bx_i + c$

The difference between y_i and $\widehat{y_i}$ is termed as the error e_i at the point (x_i, y_i) and is given by $e_i = y_i - \widehat{y_i} = \{y_i - (a\,x_i^{\,2} + bx_i + c)\}$.

By the Least Squares Principle, best values of a *and* b are those values which minimize the sum of squares of errors given by,

$$S = \sum e_i^{\,2} = \sum\{y_i - (a\,x_i^{\,2} + bx_i + c)\}^2$$

----------------------*Contd in next slide*

Correlation and Regression

Curve fitting

To fit $y = y = ax^2 + b\,x + c$ -------Contd from previous slide

$S = \sum e_i^{\,2} = \sum\{y_i - (a\,x_i^{\,2} + bx_i + c)\}^2$

S is a minimum when $\frac{\partial S}{\partial a} = 0$, $\frac{\partial S}{\partial b} = 0$ and $\frac{\partial S}{\partial c} = 0$

$\frac{\partial S}{\partial a} = 0 \longrightarrow 2\sum\{y_i - (a\,x_i^{\,2} + bx_i + c)\}^{2-1}(0 - x_i^{\,2} - 0 - 0) = 0$

$-2\sum(y_i - a\,x_i^{\,2} - bx_i - c)\,x_i^{\,2} = 0 \longrightarrow \sum x_i^{\,2} y_i = a\sum x_i^{\,4} + b\,\Sigma x_i^{\,3} + c\,\Sigma x_i^{\,2} \;\ldots\ldots\ldots (1)$

$\frac{\partial S}{\partial b} = 0 \longrightarrow 2\sum\{y_i - (a\,x_i^{\,2} + bx_i + c)\}^{2-1}(0 - 0 - x_i - 0) = 0$

$-2\sum(y_i - a\,x_i^{\,2} - bx_i - c)\,x_i = 0 \longrightarrow \sum x_i y_i = a\sum x_i^{\,3} + b\,\Sigma x_i^{\,2} + c\,\Sigma x_i \;\ldots\ldots\ldots (2)$

$\frac{\partial S}{\partial c} = 0 \longrightarrow 2\sum\{y_i - (a\,x_i^{\,2} + bx_i + c)\}^{2-1}(0 - 0 - 0 - 1) = 0$

$-2\sum(y_i - a\,x_i^{\,2} - bx_i - c) = 0 \longrightarrow \sum y_i = a\sum x_i^{\,2} + b\,\Sigma\,x_i + nc \;\ldots\ldots\ldots\ldots (3)$

The normal equations are given by,

$\sum y_i = a\sum x_i^{\,2} + b\,\Sigma\,x_i + nc \ldots\ldots (3)$

$\sum x_i y_i = a\sum x_i^{\,3} + b\,\Sigma x_i^{\,2} + c\,\Sigma\,x_i \;\ldots (2)$

$\sum x_i^{\,2} y_i = a\sum x_i^{\,4} + b\,\Sigma x_i^{\,3} + c\,\Sigma x_i^{\,2} \;\ldots (1)$

Solving (1), (2) and (3) we get a, b and c.

x_i	y_i	$x_i y_i$	$x_i^{\,2}$	$x_i^{\,2} y_i$	$x_i^{\,3}$	$x_i^{\,4}$
x_1	y_1	$x_1 y_1$	$x_1^{\,2}$	$x_1^{\,2} y_1$	$x_1^{\,3}$	$x_1^{\,4}$
x_2	y_2	$x_2 y_2$	$x_2^{\,2}$	$x_2^{\,2} y_2$	$x_2^{\,3}$	$x_2^{\,4}$
…	….	….	……	……..	……	…….
Σx_i	Σy_i	$\Sigma x_i y_i$	$\Sigma x_i^{\,2}$	$\Sigma x_i^{\,2} y_i$	$\Sigma x_i^{\,3}$	$\Sigma x_i^{\,4}$

Correlation and Regression

Curve fitting

Qn 2: Fit a parabola to the data

X	1.0	1.5	2.0	2.5	3.0	3.5	4.0
y	1.1	1.3	1.6	2.0	2.7	3.4	4.1

Let the parabolat o be fitted be

$$y = ax^2 + b\,x + c$$

The normal equations are given by

$\sum y_i = a \sum x_i^{\,2} + b\, \Sigma\, x_i + nc\ldots\ldots(1)$

$\sum x_i y_i = a \sum x_i^{\,3} + b\, \Sigma x_i^{\,2} + c\, \Sigma\, x_i\ \ldots(2)$

$\sum x_i^{\,2} y_i = a \sum x_i^{\,4} + b\, \Sigma x_i^{\,3} + c\, \Sigma x_i^{\,2}\ \ldots(3)$

From the given data set

$n = 7, \sum x = 17.5,\ \sum y = 16.2,$

$\sum xy = 47.65, \sum x^2 = 50.75$

$\sum x^2 y = 154.475, \sum x^3 = 161.875$

$\sum x^4 = 548.1875$

Substituting from table in (1), (2) and (3) and solving them we get **a = 0.24**

b = -0.20

c = 1.00

The fitted parabola is y = $\mathbf{0.24\,x^2 - 0.20\,x + 1}$

Correlation and Regression

Curve fitting

Normal Equations for fitting various curves

To Fit $Y = \alpha + \beta x$

Normal Equations are

$\Sigma y = n\alpha + \beta\, \Sigma x$

$\Sigma xy = \alpha\, \Sigma x + \beta\, \Sigma x^2$

To fit $Y = ax^2 + bx + c$

The normal equations are

$\sum y_i = a \sum x_i^{\,2} + b\, \Sigma\, x_i + nc$

$\sum x_i y_i = a \sum x_i^{\,3} + b\, \Sigma x_i^{\,2} + c\, \Sigma\, x_i$

$\sum x_i^{\,2} y_i = a \sum x_i^{\,4} + b\, \Sigma x_i^{\,3} + c\, \Sigma x_i^{\,2}$

To Fit $x = \alpha + \beta y$

Normal Equations are

$\Sigma x = n\alpha + \beta\, \Sigma y$

$\Sigma xy = \alpha\, \Sigma y + \beta\, \Sigma y^2$

To fit $x = \alpha + \beta y + \gamma y^2$

The normal equations are

$\Sigma x = n\alpha + \beta \Sigma y + \gamma \Sigma y^2$

$\Sigma xy = \alpha \Sigma y + \beta\, \Sigma y^2 + \gamma \Sigma y^3$

$\Sigma x y^2 = \alpha \Sigma y^2 + \beta \Sigma y^3 + \gamma \Sigma y^4$

To fit curves of the form (1) $y = ax^b$ (2) $y = ab^x$ (3) $y = ae^{bx}$

(1) To fit $y = ax^b$

To fit curve of the form $y = ax^b$,
We first apply a transformation to convert the given equation into a linear equation.

Taking log on both sides,

$\log y = \log a + b \log x$ ...(1)

Let $Y = \log y$, $A = \log a$

$B = b$ and $X = \log x$

Then (1) gives $Y = A + BX$ ...(2)

The normal equations for fitting the linear equation (2)are

$\Sigma Y = nA + B\Sigma X$ ....(3)

$\Sigma XY = A\Sigma X + B\Sigma X^2$ ....(4)

Solving (3) and (4) we get A and B.

Then **$a = \text{Antilog}(A)$, $b = B$**

x	y	X = log x	Y = log y	XY	X^2
x_1	y_1	$X_1 = \log x_1$	$Y_1 = \log y_1$	X_1Y_1	${X_1}^2$
x_2	y_2	$X_2 = \log x_2$	$Y_2 = \log y_2$	X_2Y_2	${X_2}^2$
.....	.....	....	....	....	....
x_n	y_n	$X_n = \log x_n$	$Y_n = \log y_n$	X_nY_n	${X_n}^2$
----	----	ΣX	ΣY	ΣXY	ΣX^2

(2) To fit $y = ab^x$

Taking log on both sides,

$\log y = \log a + x \log b$ ...(1)

Let $Y = \log y$, $A = \log a$

$B = \log b$ and $X = x$

Then (1) gives $Y = A + BX$ ...(2)

The normal equations for fitting the linear equation (2)are

$$\Sigma Y = nA + B\Sigma X \quad \text{..........(3)}$$

$$\Sigma XY = A\Sigma X + B\Sigma X^2 \quad \text{.........(4)}$$

Solving (3) and (4) we get A and B.

Then **a = Antilog(A), b = B**

To fit curve of the form $y = ab^x$, We first apply a transformation to convert the given equation into a linear equation.

x	y	X = x	Y = log y	XY	X^2
x_1	y_1	$X_1 = x_1$	$Y_1 = \log y_1$	X_1Y_1	$X_1{}^2$
x_2	y_2	$X_2 = x_2$	$Y_2 = \log y_2$	X_2Y_2	$X_2{}^2$
.....	.....	....	....	....	....
x_n	y_n	$X_n = x_n$	$Y_n = \log y_n$	X_nY_n	$X_n{}^2$
----	----	ΣX	ΣY	ΣXY	ΣX^2

(2) To fit $y = ae^{bx}$

Taking log on both sides,

$\text{Log } y = \log a + bx\,.log_{10}e$

$\text{Log } y = \log a + b\,log_{10}e\,.\,x$ ...(1)

Let $Y = \log y$, $A = \log a$, $X = x$

$B = b.\,log_{10}\,e = 0.4343b$

Then (1) gives $Y = A + BX$ ...(2)

The normal equations for fitting the linear equation (2)are

$\mathbf{\Sigma Y = nA + B\Sigma X}$ ....(3)

$\mathbf{\Sigma XY = A\Sigma X + B\Sigma X^2}$ ....(4)

Solving (3) and (4) we get A and B.

Then $\mathbf{a = Antilog(A)}$, $\mathbf{b = B/0.4343}$

To fit curve of the form $y = ae^{bx}$,
We first apply a transformation to convert the given equation into a linear equation.

x	y	X = x	Y = log y	XY	X^2
x_1	y_1	$X_1 = x_1$	$Y_1 = \log y_1$	X_1Y_1	$X_1{}^2$
x_2	y_2	$X_2 = x_2$	$Y_2 = \log y_2$	X_2Y_2	$X_2{}^2$
.....	.....	....	....	....	....
x_n	y_n	$X_n = x_n$	$Y_n = \log y_n$	X_nY_n	$X_n{}^2$
----	----	ΣX	ΣY	ΣXY	ΣX^2

Correlation and Regression

Curve fitting

Qn 3: Fit the curve $y = ax^b$ for

x	1	2	3	4	5	6
y	2.98	4.26	5.21	6.10	6.80	7.50

Taking log on both sides,

$\text{Log } y = \log a + b \log x \ldots(1)$

Let $Y = \log y$, $A = \log a$

$B = b$ and $X = \log x$

Then (1) gives $Y = A + BX \ldots(2)$

The normal equations for fitting the linear equation (2)are

$\Sigma Y = nA + B\Sigma X \quad \ldots\ldots\ldots(3)$

$\Sigma XY = A\Sigma X + B\Sigma X^2 \quad \ldots\ldots\ldots(4)$

x	y	X = log x	Y = log y	X Y	X^2
1	2.98	0.000	0.474	0.000	0.000
2	4.26	0.301	0.629	0.189	0.091
3	5.21	0.477	0.717	0.342	0.228
4	6.10	0.602	0.785	0.473	0.362
5	6.80	0.699	0.833	0.582	0.489
6	7.50	0.778	0.875	0.681	0.606
		2.857	**4.313**	**2.267**	**1.775**

Substituting from table in (3) and (4)

$4.313 = 6A + 2.857B \quad \ldots\ldots \quad \ldots\ldots\ldots(5)$

$2.267 = 2.857A + 1.775B \quad \ldots\ldots\ldots(6)$

$(5) \text{ x } 2.857 \rightarrow 12.322 = 17.142A + 8.162 \text{ B}\ldots\ldots(7)$

$(6) \text{ x } 6 \quad \rightarrow 13.602 = 17.142\text{A} + 10.65 \text{ B}\ldots\ldots(8)$

$(8) - (7) \quad \rightarrow \quad 1.28 = 2.488 \text{ B} \quad \rightarrow \text{B} = 0.51$

Putting B in (5), $6\text{ A} = 4.313 - 1.457 \quad \rightarrow \text{A} = 0.48$

$\therefore b = B = \mathbf{0.51}$

$a = \text{Antilog}(A)$

$= \text{Antilog}(0.48) = \mathbf{3.020}$

$\therefore$ The fitted curve is given by,

$\mathbf{y = 3.020\, x^{0.51} \cong 3x^{0.5} = 3\sqrt{x}}$

Correlation and Regression

Curve fitting

Qn 4: Fit the curve $y = ab^x$ for

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

Taking log on both sides,

$\text{Log } y = \log a + x \log b \ldots(1)$

Let $Y = \log y$, $A = \log a$

$B = \log b$ and $X = x$

Then (1) gives $Y = A + BX \ldots(2)$

The normal equations for fitting the linear equation (2)are

$\Sigma Y = nA + B\Sigma X \quad \ldots\ldots\ldots(3)$

$\Sigma XY = A\Sigma X + B\Sigma X^2 \quad \ldots\ldots\ldots(4)$

x	y	X = x	Y = log y	X Y	X^2
2	144.00	2.00	2.16	4.32	4.0
3	172.80	3.00	2.24	6.71	9.0
4	207.40	4.00	2.32	9.27	16.0
5	248.80	5.00	2.40	11.98	25.0
6	298.50	6.00	2.47	14.85	36.0
		20.00	11.58	47.13	90.0

Substituting from table in (3) and (4)

$11.58 = 5A + 20B \quad \ldots\ldots \ldots\ldots\ldots(5)$

$47.13 = 20A + 90B \quad \ldots\ldots\ldots\ldots\ldots(6)$

$(5) \times 20 \rightarrow 231.60 = 100A + 400B \ldots(7)$

$(6) \times 5 \rightarrow 235.65 = 100A + 450B \ldots(8)$

$(8) - (7) \rightarrow 4.05 = 50B \rightarrow B = 0.09$

Putting B in (5), $5A = 11.58 - 1.8 \rightarrow A = 1.96$

$\therefore b = \text{Antilog}(B) = \text{Antilog}(0.09) = \mathbf{1.23}$

$a = \text{Antilog}(A) = \text{Antilog}(1.96) = \mathbf{91.20}$

$\therefore$ The fitted curve is given by,

$\mathbf{y = 91.20\,(1.23)^x}$

Correlation and Regression

Curve fitting

Qn 5: Fit the curve $y = ae^{bx}$ for

x	1	2	3	4	5	6
y	1.6	4.5	13.8	40.2	125.3	300

Taking log on both sides,

Log $y = \log a + bx .log_{10}e$

Log $y = \log a + b log_{10}e . x$ ...(1)

Let $Y = \log y$, $A = \log a$, $X = x$

$B = b. log_{10} e = 0.4343b$

Then (1) gives $Y = A + BX$ ...(2)

The normal equations for fitting the linear equation (2)are

$\Sigma Y = nA + B\Sigma X$ ....(3)

$\Sigma XY = A\Sigma X + B\Sigma X^2$ ....(4)

Substituting from table in (3) and (4)

x	y	X = x	Y = log y	X Y	X^2
1	1.60	1.00	0.20	0.20	1.0
2	4.50	2.00	0.65	1.31	4.0
3	13.80	3.00	1.14	3.42	9.0
4	40.20	4.00	1.60	6.42	16.0
5	125.30	5.00	2.10	10.49	25.0
6	300	6.00	2.48	14.86	36.0
		21.00	8.18	36.70	91.0

$8.18 = \quad 6A + 21B$ ....(5)

$36.70 = 21A + 91B$ ....(6)

(5) x 21 $\rightarrow 171.78 = 126A + 441B$.....(7)

(6) x 6 $\rightarrow 220.20 = 126A + 546B$.....(8)

(8) – (7) $\rightarrow$ $48.42 = 105B \rightarrow B = 0.46$

Putting B in (5), $6A = 8.18 - 9.66 \rightarrow A = -0.25$

$\therefore b = \frac{B}{0.4343} = \frac{0.46}{.4343} = \mathbf{1.06}$

$a = \text{Antilog}(A) = \text{Antilog}(-0.25)$

$= \text{Antilog}(\bar{1}.75) = \mathbf{0.5623}$

The fitted curve is given by,

$\mathbf{y = 0.5623\ (e)^{1.06x}}$

Coefficient of correlation

Coefficient of correlation is a numerical measure of the degree of linear relationship between two variables.

Karl Pearson's product moment correlation coefficient (usually denoted by r or r_{xy}) is the ratio of covariance between the variables to the product of standard deviations of the variables.

$$r = \frac{Covariance\ between\ the\ variables\ x\ and\ y}{product\ of\ standard\ deviations\ of\ x\ and\ y} = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

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

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

x	y	xy	x^2	y^2
----	----	----	----	----
----	----	----	----	----
Σx	Σy	Σxy	Σx^2	Σy^2

Correlation and Regression

Show that Coefficient of correlation is not affected by change of origin and scale

Let r_{xy} denote the correlation coefficient between two variables X and Y.

$$\text{Then } r_{xy} = \frac{Cov(x,y)}{\sigma_x \sigma_y} = \frac{\frac{1}{n}\sum\{(x-\bar{x})(y-\bar{y})\}}{\sqrt{\frac{1}{n}\sum(x-\bar{x})^2}\sqrt{\frac{1}{n}\sum(y-\bar{y})^2}}$$

Consider the transformations of the form U = $\frac{X-a}{b}$ and V = $\frac{Y-c}{d}$

Then $\bar{u} = \frac{\bar{x}-a}{b}$ and $\bar{v} = \frac{\bar{y}-c}{d}$. Let r_{uv} be the correlation coefficient between the transformed variables U and V. Then we have,

$$r_{uv} = \frac{Cov(u,v)}{\sigma_u \sigma_v} = \frac{\frac{1}{n}\sum\{(u-\bar{u})(v-\bar{v})\}}{\sqrt{\frac{1}{n}\sum(u-\bar{u})^2}\sqrt{\frac{1}{n}\sum(v-\bar{v})^2}}$$

$$= \frac{\frac{1}{n}\sum\left\{\left(\frac{x-a}{b} - \frac{\bar{x}-a}{b}\right)\left(\frac{y-c}{d} - \frac{\bar{y}-c}{d}\right)\right\}}{\sqrt{\frac{1}{n}\sum\left(\frac{x-a}{b} - \frac{\bar{x}-a}{b}\right)^2}\sqrt{\frac{1}{n}\sum\left(\frac{y-c}{d} - \frac{\bar{y}-c}{d}\right)^2}} = \frac{\frac{1}{bd}\,\frac{1}{n}\sum\{(x-\bar{x})(y-\bar{y})\}}{\frac{1}{bd}\sqrt{\frac{1}{n}\sum(x-\bar{x})^2}\sqrt{\frac{1}{n}\sum(y-\bar{y})^2}} = r_{xy}$$

$\therefore r_{uv} = r_{xy}$. Hence r is independent of change of origin and scale

Qn 6: Calculate r from the following details in a data sheet.

$$n = 50, \sum x = 75, \sum y = 80, \sum x^2 = 130, \sum y^2 = 140, \sum x\, y = 120$$

$$r = \frac{\frac{\sum xy}{n} - \left(\frac{\sum x}{n}\right)\left(\frac{\sum y}{n}\right)}{\sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2}\sqrt{\frac{\sum y^2}{n} - \left(\frac{\sum y}{n}\right)^2}} = \frac{\frac{120}{50} - \left(\frac{75}{50}\right)\left(\frac{80}{50}\right)}{\sqrt{\frac{130}{50} - \left(\frac{75}{50}\right)^2}\sqrt{\frac{140}{50} - \left(\frac{80}{50}\right)^2}} = 0 \text{ (why?)}$$

Qn 7: In the calculation of the above data sheet, one pair (1.5,2) was wrongly taken as (2.5, 1). Calculate the correct correlation coefficient

Correct $\sum x = 75 - 2.5 + 1.5 = 74$

Correct $\sum y = 80 - 1 + 2 = 81$

Correct $\sum x^2 = 130 - (2.5)^2 + (1.5)^2 = 126$

Correct $\sum y^2 = 140 - (1)^2 + (2)^2 = 143$

Correct $\sum x\, y = 120 - 2.5\text{x}1 + 1.5\text{x}2 = 120.5$

$$\text{Correct } r = \frac{\frac{\sum xy}{n} - \left(\frac{\sum x}{n}\right)\left(\frac{\sum y}{n}\right)}{\sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2}\sqrt{\frac{\sum y^2}{n} - \left(\frac{\sum y}{n}\right)^2}} = \frac{\frac{120.5}{50} - \left(\frac{74}{50}\right)\left(\frac{81}{50}\right)}{\sqrt{\frac{126}{50} - \left(\frac{74}{50}\right)^2}\sqrt{\frac{143}{50} - \left(\frac{81}{50}\right)^2}} = \textbf{H.W}$$

Qn 8: Calculate n from the following details

r = 0.8, $\sum x^2 = 90$, $\sum x\,y = 60$, $\sigma_y = 2.5$ (x and y are deviations from mean)

We have r = $\dfrac{\sum\{(\boldsymbol{x}-\overline{\boldsymbol{x}})(\boldsymbol{y}-\overline{\boldsymbol{y}})\}}{\sqrt{\sum(\boldsymbol{x}-\overline{\boldsymbol{x}})^2}\sqrt{\sum(\boldsymbol{y}-\overline{\boldsymbol{y}})^2}}$

When x and y are deviations from mean, the above formula becomes,

$$r = \frac{\sum \boldsymbol{x}\,\boldsymbol{y}}{\sqrt{\sum(\boldsymbol{x})^2}\sqrt{\sum \boldsymbol{y}^2}} \longrightarrow r^2 = \frac{(\sum x\,y)^2}{(\sum x^2)(\sum y^2)} = \frac{(\sum x\,y)^2}{(\sum x^2)\,n\,\sigma_y{}^2}$$

$$(0.8)^2 = \frac{(60)^2}{(90)\,n(2.5)^2} \longrightarrow n = \frac{3600}{90 \text{ x } 6.25 \text{ x } 0.64} = \mathbf{10}$$

Qn 9: Calculate σ_y from the following details

r = 0.28, Cov (x, y) = 7, V(X) = 9

$$r = \frac{Cov(x,y)}{\sigma_x\sigma_y} = \frac{Cov(x,y)}{\sqrt{V(X)}\sqrt{V(Y)}} \longrightarrow 0.28 = \frac{7}{\sqrt{9}\sqrt{V(Y)}}$$

$$\sigma_y = \sqrt{V(Y)} = \frac{7}{\sqrt{9} \text{ x } 0.28} = \mathbf{8.33}$$

Qn 10: Calculate the correlation coefficient r from following data:

Ht of fathers (X):	65	66	67	67	68	69	70	72
Ht of sons (Y):	67	68	65	68	72	72	69	71

$$r = \frac{Cov(x,y)}{\sigma_x \sigma_y} = \frac{\frac{\sum xy}{n} - \left(\frac{\sum x}{n}\right)\left(\frac{\sum y}{n}\right)}{\sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2}\sqrt{\frac{\sum y^2}{n} - \left(\frac{\sum y}{n}\right)^2}}$$

$$\mathrm{Cov(x,y)} = \frac{\sum xy}{n} - \left(\frac{\sum x}{n}\right)\left(\frac{\sum y}{n}\right) = \frac{37560}{8} - \left(\frac{544}{8}\right)\left(\frac{552}{8}\right)$$

$$= 4695 - 68 \times 69 = 4695 - 4692 = \mathbf{3}$$

$$\sigma_x = \sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2} = \sqrt{\frac{37028}{8} - \left(\frac{544}{8}\right)^2} = \sqrt{4.5} = \mathbf{2.12}$$

$$\sigma_x = \sqrt{\frac{\sum y^2}{n} - \left(\frac{\sum y}{n}\right)^2} = \sqrt{\frac{38132}{8} - \left(\frac{552}{8}\right)^2} = \sqrt{5.5} = \mathbf{2.35}$$

$$r = \frac{\mathrm{Cov(x,y)}}{\sigma_x \sigma_y} = \frac{3}{2.12 \times 2.35} = \mathbf{0.603}$$

X	Y	XY	X^2	Y^2
65	67	4355	4225	4489
66	68	4488	4356	4624
67	65	4355	4489	4225
67	68	4556	4489	4624
68	72	4896	4624	5184
69	72	4968	4761	5184
70	69	4830	4900	4761
72	71	5112	5184	5041
544	**552**	**37560**	**37028**	**38132**

Qn 11: Calculate the correlation coefficient r from following data:

Ht of fathers (X):	65	66	67	67	68	69	70	72
Ht of sons (Y):	67	68	65	68	72	72	69	71

$$r = \frac{Cov(x,y)}{\sigma_x \sigma_y} = \frac{\frac{1}{n}\sum\{(x-\bar{x})(y-\bar{y})\}}{\sqrt{\frac{1}{n}\sum(x-\bar{x})^2}\sqrt{\frac{1}{n}\sum(y-\bar{y})^2}}$$

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

Let U = (X - $\bar{X}$) and V = (Y - $\bar{Y}$)

$$\bar{x} = \frac{\sum x}{n} = \frac{544}{8} = \mathbf{68},\ \bar{y} = \frac{\sum y}{n} = \frac{552}{8} = \mathbf{69}$$

$$\mathbf{r} = \frac{\sum\{(x-\bar{x})(y-\bar{y})\}}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}}$$

$$= \frac{\sum uv}{\sqrt{\sum(u)^2}\sqrt{\sum(v)^2}} = \frac{24}{\sqrt{36}\sqrt{44}} = \mathbf{0.603}$$

X	Y	U	V	UV	U^2	V^2
65	67	-3	-2	6	9	4
66	68	-2	-1	2	4	1
67	65	-1	-4	4	1	16
67	68	-1	-1	1	1	1
68	72	0	3	0	0	9
69	72	1	3	3	1	9
70	69	2	0	0	4	0
72	71	4	2	8	16	4
544	552	0	0	**24**	**36**	**44**

Note: This method is convenient when the means $\bar{x}$ *and* $\bar{y}$ are integers.

Regression Analysis

Regression Analysis is a mathematical measure of the average relationship between two or more correlated variables.

If there are only two variables under consideration then one is taken as the independent variable and the other as dependent variable, and regression means an average relationship between them. Regression explains the average change in dependent variable with a change in the independent variable.

If the variables in a bivariate distribution are correlated, the points in the scatter diagram will show a tendency to cluster around some curve called the *curve of regression*. The mathematical equation of the regression curve is called *regression equation*. If the points in the scatter diagram are clustered around a straight line it is called the line of regression and we say that there is a linear regression between the variables. In this case the regression curve is a polynomial of degree one. When the degree of the polynomial corresponding to any regression curve is more than one, we call the regression to be curvilinear.

In practice, the regression analysis deals with estimation of the value of dependent variable for some particular value of the independent variable. The estimate is called regression estimate and equation of the line used for estimation is called the regression equation.

Two Regression Lines (What? How? Why?)

When we have two variables under consideration say X and Y, we can have two regression equations, one for estimating the value of dependent variable Y for a particular value of independent variable X and the other for estimating the value of dependent variable X for a particular value of independent variable Y.

When the variable Y is taken as dependent variable and X is taken as the independent variable (for estimating Y for a given value of X) the regression equation is called the regression equation of y on x.

When the variable X is taken as dependent variable and Y is taken as the independent variable (for estimating X for a given value of Y) the regression equation is called the regression equation of x on y.

The regression equation of y on x is obtained by minimising the sum of squares of errors parallel to y axis and the regression equation of x on y is obtained by minimising the sum of squares of errors parallel to x axis.

When all the points in the scatter diagram are exactly on a straight line, the error at any point is zero and hence the two regression lines will become one and the same or they coincides. When all the points in the scatter diagram are not exactly on a straight line, the two procedures using L.S principle will give two different equations.

Standard form of the two Regression Lines.

Assume that the regression equation used for estimating y when x is known (regression equation of y on x) is given by, y = a + bx ………………..(1)

Let $(x_1, y_1), (x_2, y_2), \ldots\ldots. (x_n, y_n)$ be the n pairs of observations on the variables X and Y. When the value of the independent variable X is say x_i, the observed value of the dependent variable Y is y_i and the corresponding estimated value of Y is given by $\widehat{y_i} = a + b\,x_i$. The difference between y_i and $\widehat{y_i}$ is the error denoted by e_i.

$\therefore e_i = y_i - \widehat{y_i} = y_i - (a + b\,x_i) = (y_i - a - b\,x_i)$

The sum of squares of errors (S) is given by, $S = \sum {e_i}^2 = \sum (y_i - a - b\,x_i)^2$ ……..(2)

By the Principle of Least Squares, the best

estimates of a and b are those values which

minimise S. For this $\frac{\partial S}{\partial a} = 0$ and $\frac{\partial S}{\partial b} = 0$

The above two equations will give the

following normal equations,

$\Sigma y = na + b\,\Sigma x$ ………….(3)

$\Sigma xy = a\,\Sigma x + b\,\Sigma x^2$ ……...(4)

Solving the normal equations, we get values of a and b.

$\Sigma y = na + b\,\Sigma x$ ………….(3)

$\Sigma xy = a\,\Sigma x + b\,\Sigma x^2$ ………..(4)

(3) x (Σx) $\longrightarrow$ $(\Sigma x)(\Sigma y) = n\,a\,(\Sigma x) + b\,(\Sigma x)\,(\Sigma x)$

$(\Sigma x)(\Sigma y) = n\,a\,(\Sigma x) + b\,(\Sigma x)^2$ ……………..(5)

(4) x n $\longrightarrow$ $n\,\Sigma xy = n\,a\,\Sigma x + n\,b\,\Sigma x^2$ ……………… (6)

(6) – (5) $\longrightarrow$ $n\,\Sigma xy - (\Sigma x)(\Sigma y) = n\,b\,\Sigma x^2 - b\,(\Sigma x)^2$

$= b\,\{n\,\Sigma x^2 - (\Sigma x)^2\}$

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

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

$$= \frac{\frac{\sum xy}{n} - \frac{\sum x}{n}\frac{\sum y}{n}}{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2} = \frac{\mathbf{Cov(x,y)}}{\boldsymbol{\sigma_x}^2} \quad \text{……………………….(7)}$$

Correlation and Regression

Regression Analysis

(1) gives, $\bar{y} = a + b\,\bar{x}$

$$\therefore a = \bar{y} - b\bar{x} = \bar{y} - \frac{Cov(x,y)}{{\sigma_x}^2}\,\bar{x}$$

$$\therefore a = \left\{\bar{y} - \frac{Cov(x,y)}{{\sigma_x}^2}\,\bar{x}\right\} \ldots\ldots\ldots\ldots\ldots\ldots(8)$$

Using (7) and (8) in (1), we get $y = \left\{\bar{y} - \frac{Cov(x,y)}{{\sigma_x}^2}\,\bar{x}\right\} + \frac{Cov(x,y)}{{\sigma_x}^2}\,x$

ie, $y - \bar{y} = \frac{Cov(x,y)}{{\sigma_x}^2}(x - \bar{x})$ ……………*Standard form 1*

ie, $y - \bar{y} = \frac{r\sigma_y}{\sigma_x}(x - \bar{x})$ ……………*Standard form 2*

ie, $y - \bar{y} = b_{yx}\ (x - \bar{x})$ ……………*Standard form 3,*

where $b_{yx} = \frac{Cov(x,y)}{{\sigma_x}^2} = \frac{r\sigma_y}{\sigma_x}$ is called the regression coefficient of y on x

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

$$Cov(x, y) = r\ \sigma_x\sigma_y$$

Note: *From Standard form 1 above we get* $y = \frac{Cov(x,y)}{{\sigma_x}^2}\,x + \left(\bar{y} - \frac{Cov(x,y)}{{\sigma_x}^2}\,\bar{x}\right)$, *which is of the form* $y = a + bx$

Note: *Slope of the regression line of y on x is given by* $\mathbf{b} = \frac{Cov(x,y)}{{\sigma_x}^2} = \frac{r\sigma_y}{\sigma_x} = b_{yx}$

Correlation and Regression

Similarly if the regression line of x on y can be taken as $x = c + dy$

As in the case of the regression equation of y on x, the standard form of the regression line of x on y can be obtained as

ie, $x - \bar{x} = \frac{\mathrm{Cov}(x,y)}{\sigma_y{}^2}(y - \bar{y})$ ……………*Standard form 1*

ie, $x - \bar{x} = \frac{r\sigma_x}{\sigma_y}(y - \bar{y})$ ……………*Standard form 2*

ie, $x - \bar{x} = b_{xy}\ (y - \bar{y})$ ……………*Standard form 3,*

where $b_{xy} = \frac{\mathrm{Cov}(x,y)}{\sigma_y{}^2} = \frac{r\sigma_x}{\sigma_y}$ is called the regression coefficient of x on y

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

$$\mathrm{Cov}(x, y) = r\ \sigma_x\sigma_y$$

Note: *From Standard form 1 above we get* $(y - \bar{y}) = \frac{\sigma_y{}^2}{Cov(x,y)}(x - \bar{x})$

ie, $y = \frac{\sigma_y{}^2}{Cov(x,y)}x + \left(\bar{y} - \frac{\sigma_y{}^2}{Cov(x,y)}\bar{x}\right)$ *which is of the form* $y = mx + c$

Note: *Slope of the regression line of x on y is given by* $\mathbf{m} = \frac{\sigma_y{}^2}{Cov(x,y)} = \frac{\boldsymbol{\sigma_y}}{\boldsymbol{r\sigma_x}} = \boldsymbol{b_{xy}}$

Comparison of Regression line of *y on x* and regression line of *x on y*

Regression line of y on x	Regression line of x on y
Used to estimate y when x is known Y is the dependent variable Obtained by minimising the sum of squares of errors parallel to the y axis	Used to estimate x when y is known X is the dependent variable Obtained by minimising the sum of squares of errors parallel to the x axis
The R.L of y on x can be y = ax +b	The R.L of y on x can be x = cy +d
Standard form of R.L of y on x are : • $y - \bar{y} = \frac{\mathrm{Cov}(x,y)}{\sigma_x^2}(x - \bar{x})$ • $y - \bar{y} = \frac{r\sigma_y}{\sigma_x}(x - \bar{x})$ • $y - \bar{y} = b_{yx}\ (x - \bar{x})$	Standard form of R.L of x on y are : • $x - \bar{x} = \frac{\mathrm{Cov}(x,y)}{\sigma_y^2}(y - \bar{y})$ • $x - \bar{x} = \frac{r\sigma_x}{\sigma_y}(y - \bar{y})$ • $x - \bar{x} = b_{xy}\ (y - \bar{y})$
$b_{yx} = \frac{\mathrm{Cov}(x,y)}{\sigma_x^2} = \frac{r\sigma_y}{\sigma_x}$ is called the regression coefficient of y on x, which measures the change in y for unit change in x.	$b_{xy} = \frac{\mathrm{Cov}(x,y)}{\sigma_y^2} = \frac{r\sigma_x}{\sigma_y}$ is called the regression coefficient of x on y, which measures the change in x for unit change in y.

Remarks about Correlation and Regression

Correlation	Regression
Identify the nature and Measure the degree of relationship between the variables.	Measure the average relationship between the variables
Correlation coefficient is a relative measure	Regression is an absolute measure of relationship
Correlation coefficient is the signed Geometric mean of regression coefficients r = $\pm\sqrt{b_{yx}\, b_{xy}}$	There are two regression lines and they are not mutually reversible
If both regression coefficients are positive, the sign of correlation coefficient is positive. If both are negative, r is also negative	Both regression coefficients are of the same sign (either both positive or bot negative)
Correlation coefficient is not affected by change of origin and scale	Regression coefficients give the slope of regression lines. They measure the average change in dependent variable when independent variable changes by one unit.
Correlation coefficient is a symmetrical function between x and y	Regression coefficients are affected by change of scale.

P.T the correlation coefficient is the G.M of regression coefficients

We have the regression coefficients given by,

$$b_{yx} = \frac{\text{Cov(x,y)}}{\sigma_x{}^2} = \frac{r\sigma_y}{\sigma_x} \text{ and } b_{xy} = \frac{\text{Cov(x,y)}}{\sigma_y{}^2} = \frac{r\sigma_x}{\sigma_y}$$

$$\therefore b_{yx}.\, b_{xy} = \frac{r\sigma_y}{\sigma_x} . \frac{r\sigma_x}{\sigma_y} = r^2 \longrightarrow \text{r} = \pm\sqrt{b_{yx}.\, b_{xy}}$$

$$= \pm \text{ GM of regression coefficients}$$

If both the regression coefficients are positive, the value of r is positive.

If both the regression coefficients are negative, the value of r is negative.

The regression coefficients, $b_{yx} = \frac{\text{Cov(x,y)}}{\sigma_x{}^2}$ and $b_{xy} = \frac{\text{Cov(x,y)}}{\sigma_y{}^2}$ *can not be of opposite signs since sign of cov(x,y) decides the sign of regression coefficients.*

$\therefore$ r $= +\sqrt{b_{yx}.\, b_{xy}}$ if both b_{yx} and b_{xy} are positive

r $= -\sqrt{b_{yx}.\, b_{xy}}$ if both b_{yx} and b_{xy} are negative

Obtain the angle between the two regression lines

The regression line of y on x is given by, $y - \bar{y} = \frac{r\sigma_y}{\sigma_x}(x - \bar{x})$

The above equation can be written as $y = \frac{\text{Cov}(x,y)}{\sigma_x^2}\, x + \left(\bar{y} - \frac{Cov(x,y)}{\sigma_x^2}\bar{x}\right)$,

which is of the form $y = a + bx$

Then the slope of the regression line of y on x $(m_1) = \frac{\boldsymbol{Cov(x,y)}}{\boldsymbol{\sigma_x^2}} = \frac{\boldsymbol{r\sigma_y}}{\boldsymbol{\sigma_x}}$

The regression line of x on y is given by, $x - \bar{x} = \frac{r\sigma_x}{\sigma_y}(y - \bar{y})$

The above equation can be written as $y = \frac{\sigma_y^2}{Cov(x,y)}\, x + \left(\bar{y} - \frac{\sigma_y^2}{Cov(x,y)}\, \bar{x}\right)$,

which is of the form $y = mx + c$

Then the slope of the regression line of x on y $(m_2) = \frac{\sigma_y^2}{Cov(x,y)} = \frac{\boldsymbol{\sigma_y}}{\boldsymbol{r\sigma_x}}$

Let θ be the angle between the two regression lines.

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

Angle between the regression lines - Contd

$$\tan\theta = \pm\frac{m_2-m_1}{1+m_1m_2} = \pm\frac{\frac{\sigma_y}{r\sigma_x}-\frac{r\sigma_y}{\sigma_x}}{1+\frac{r\sigma_y\,\sigma_y}{\sigma_x\,r\sigma_x}} = \pm\frac{\frac{\sigma_y-r^2\,\sigma_y}{r\,\sigma_x}}{1+\frac{\sigma_y{}^2}{\sigma_x{}^2}} = \pm\left(\frac{1-r^2}{r}\right)\frac{\frac{\sigma_y}{\sigma_x}}{\frac{\sigma_x{}^2+\sigma_y{}^2}{\sigma_x{}^2}}$$

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

Positive or negative sign is taken according as the angle is acute or obtuse.

If θ is an acute angle, $\theta = tan^{-1}\left[\left(\frac{1-r^2}{|r|}\right)\left(\frac{\sigma_x\sigma_y}{\sigma_x{}^2+\sigma_y{}^2}\right)\right]$

If θ is an obtuse angle, $\theta = tan^{-1}\left[\left(\frac{r^2-1}{|r|}\right)\left(\frac{\sigma_x\sigma_y}{\sigma_x{}^2+\sigma_y{}^2}\right)\right]$

Note 1: When r = 0, tan θ = ∞, giving θ = 90, i.e., both the regression lines are perpendicular to each other

Note 1: When r = ±1, tan θ = 0, giving θ = 0, i.e., both the regression lines coincides with each other. In this case there is a perfect correlation (either positive when r = 1 or negative when r = -1) between the variables involved.

Standard error of estimate of Y and Standard error of estimate of X

Let $(x_1, y_1), (x_2, y_2), \ldots\ldots. (x_n, y_n)$ be the n pairs of observations.

The regression line of y on x is given by, $y - \bar{y} = \frac{r\sigma_y}{\sigma_x}(x - \bar{x})$

Corresponding to the point (x_i, y_i), the value of Y is observed as y_i and estimated as, $\hat{y}_i = \bar{y} + \frac{r\sigma_y}{\sigma_x}(x - \bar{x})$

The standard error of estimate of y (denoted by S_y) is the square root of the arithmetic mean of the squares of deviations between y_i and $\hat{y}_i$.

$$\therefore\ S_y{}^2 = \frac{1}{n}\sum(y_i - \hat{y}_i.\ )^2 = \frac{1}{n}\sum\left\{y_i - \left[\bar{y} + \frac{r\sigma_y}{\sigma_x}(x - \bar{x})\right]\right\}^2$$

$$= \frac{1}{n}\sum\left\{(y_i - \bar{y}) - \frac{r\sigma_y}{\sigma_x}(x - \bar{x})\right\}^2$$

$$= \frac{1}{n}\sum\left\{(y_i - \bar{y})^2 + \left(\frac{r\sigma_y}{\sigma_x}(x - \bar{x})\right)^2 - 2(y_i - \bar{y})\frac{r\sigma_y}{\sigma_x}(x - \bar{x})\right\}$$

$$= \frac{1}{n}\sum(y_i - \bar{y})^2 + r^2\frac{\sigma_y{}^2}{\sigma_x{}^2}\frac{1}{n}\sum(x_i - \bar{x})^2 - 2\frac{r\sigma_y}{\sigma_x}\frac{1}{n}\sum(y_i - \bar{y})(x - \bar{x})$$

Standard error of estimate of Y and Standard error of estimate of X

$$= \frac{1}{n}\sum(y_i - \bar{y})^2 + r^2\frac{{\sigma_y}^2}{{\sigma_x}^2}\frac{1}{n}\sum(x_i - \bar{x})^2 - 2\frac{r\sigma_y}{\sigma_x}\frac{1}{n}\sum(y_i - \bar{y})(x - \bar{x})$$

$$= {\sigma_y}^2 + r^2\frac{{\sigma_y}^2}{{\sigma_x}^2}{\sigma_x}^2 - 2\frac{r\sigma_y}{\sigma_x}\,\mathrm{Cov(x,y)}$$

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

$$= {\sigma_y}^2 - r^2{\sigma_y}^2$$

$$= (1 - r^2){\sigma_y}^2$$

$\therefore$ *Standard Error of estimate of y*, $S_y = \sqrt{(1 - r^2)}\,\sigma_y$

Similarly Standard Error of estimate of x, $S_x = \sqrt{(1 - r^2)}\,\sigma_x$

Note: *To show that -1 $\leq$ r $\leq$ 1*

Being a perfect square, ${S_y}^2 \geq 0$ always

i.e,. $(1 - r^2){\sigma_y}^2 \geq 0 \longrightarrow (1 - r^2) \geq 0 \longrightarrow r^2 \leq 1 \longrightarrow -1 \leq r \leq 1$

To find the mean of variables and identification of the regression lines

To find the mean of variables: The two regression lines are given by

$$y - \bar{y} = \frac{r\sigma_y}{\sigma_x}(x - \bar{x}) \ \ldots\ldots\ldots\ldots(1)$$

$$x - \bar{x} = \frac{r\sigma_x}{\sigma_y}(y - \bar{y}) \ \ldots\ldots\ldots\ldots(2)$$

The point ($\bar{x}$, $\bar{y}$) satisfies both the equations (1) and (2). Hence ($\bar{x}$, $\bar{y}$) is the point of intersection of (1) and (2), which can be obtained by solving the regression lines. Therefore, the mean of the variables are obtained by solving the regression lines. The value of x will give $\bar{x}$ and value of y will give $\bar{y}$.

To identify the regression lines: Assume any one the lines to be the regression line of y on x and express it in the form y = ax + b.

Assume the other line to be the regression line of x on y and express it in the form x = cy +d.

If the assumptions are correct, the value of ac is nothing but r^2 (Why?)

But $r^2 \leq 1$ (Why?).

$\therefore$ The assumptions are correct if ac ≤ 1

$\therefore$ If ac ≤ 1, assumptions are correct, a and c will give the regression coefficients.

Question No.1: A computer while calculating the regression coefficients between two variables x and y from 25 pairs of observations obtained the following results:

$$n = 25, \sum x = 125, \sum y = 100, \sum x^2 = 650, \sum y^2 = 460, \sum xy = 508.$$

It was however, discovered in checking that it had copied down two pairs of observations as (6,14) and (8,6) while the correct pairs were (8,12) and ((6,8).

Obtain (1) Regression coefficients (2) Regression lines (3) r

(4) Value of X when Y = 3 (5) value of Y for value of X in (4)

Answer: We have, $\sum x = 125, \sum y = 100, \sum x^2 = 650, \sum y^2 = 460, \sum xy = 508$

Correct $\sum x = 125 - (8+6) + (8+6) = 125$

Correct $\sum y = 100 - (12+8) + (14+6) = 100$

Correct $\sum x^2 = 650 - (8^2 + 6^2) + (6^2 + 8^2) = 650$

Correct $\sum y^2 = 460 - (14^2 + 6^2) + (12^2 + 8^2) = 436$

Correct $\sum xy = 508 - (6x14 + 8x6) + (8x12 + 6x8) = 520$

$$\text{Cov(x,y)} = \frac{\sum xy}{n} - \frac{\sum x}{n}\frac{\sum y}{n} = \frac{520}{25} - \frac{125}{25}\frac{100}{25} = \frac{25\ x\ 520 - 125\ x\ 100}{25\ x\ 25} = \frac{125\ x\ 104 - 125\ x\ 100}{25\ x\ 25} = \mathbf{\frac{4}{5}}$$

$$\sigma_x = \sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2} = \sqrt{\frac{650}{25} - \left(\frac{125}{25}\right)^2} = \sqrt{\frac{25\ x\ 650 - 125\ x\ 125}{25\ x\ 25}} = \sqrt{\frac{125\ x\ 130 - 125\ x\ 125}{25\ x\ 25}} = \mathbf{1}$$

$$\sigma_y = \sqrt{\frac{\sum y^2}{n} - \left(\frac{\sum y}{n}\right)^2} = \sqrt{\frac{436}{25} - \left(\frac{100}{25}\right)^2} = \sqrt{\frac{25 \text{ x } 4 \text{ x } 109 - 100 \text{ x } 100}{25\ x\ 25}} = \sqrt{\frac{100\text{x } 109 - 100 \text{ x } 100}{25 \text{ x } 25}} = \mathbf{\frac{6}{5}}$$

Correlation and Regression

(1) $To find\ the\ regression\ coefficients$

Regression coefficient of y on x, $b_{yx} = \frac{Cov(x,y)}{\sigma_x{}^2} = \frac{4/5}{1} = \frac{4}{5}$

Regression coefficient of x on y, $b_{xy} = \frac{Cov(x,y)}{\sigma_y{}^2} = \frac{4/5}{(6/5)^2} = \frac{4}{5} \cdot \frac{25}{36} = \frac{5}{9}$

(2) $To find\ the\ regression\ lines$

Regression line of y on x is given by, $y - \bar{y} = b_{yx}(x - \bar{x})$

$$\left(y - \frac{100}{25}\right) = \frac{4}{5}\left(x - \frac{125}{25}\right) \longrightarrow y - 4 = 0.8\,(x - 5) \longrightarrow \mathbf{y = 0.8\,x}$$

Regression line of y on x is given by, $y - \bar{y} = b_{yx}(x - \bar{x})$

$$\left(x - \frac{125}{25}\right) = \frac{5}{9}\left(y - \frac{100}{25}\right) \longrightarrow x - 5 = \frac{5}{9}\,(y - 4) \longrightarrow \mathbf{9x = 5y + 25}$$

(3) $To find\ the\ correlation\ coefficient$, $r = \frac{Cov(x,y)}{\sigma_x\,\sigma_y} = \frac{\frac{4}{5}}{1\ x\ \frac{6}{5}} = \frac{2}{3}$

(4) Value of x when y =3 is given by, 9x = 5 x 3 +24 = 39 $\longrightarrow \mathbf{x = \frac{13}{3} = 4.33}$

(5) Value of y when y =13/3 is given by, y = 0.8 x 13/3 $\longrightarrow \mathbf{Y = \frac{10.4}{3} = 3.46}$

Question No.2: The two regression lines are $x + 2y - 5 = 0$ and $2x + 3y - 8 = 0$ and variance of y is 12.

Calculate (1) the means (2) r (3) y when x = 3 (4) x when y=2 (5) $\sigma_x^{\,2}$

(1) To calculate the means

The means of the variables are obtained by solving the given regression lines

x + 2y – 5 =0 or x + 2y = 5 ………..(1)

2x +3y – 8 = 0 or 2x +3y = 8 ………(2)

Solving (1) and (2) we get, **$\bar{x}$ = 1and $\bar{y}$ =2**

(2) To find the value of r

$$r = \pm\sqrt{byx \cdot bxy}$$

First we have to identify the regression lines as follows:

Let the regression line of y on x be, x + 2y – 5 =0

Then, y = - $\frac{1}{2}$ x + $\frac{5}{2}$, which is of the form y = ax + b where a = - $\frac{1}{2}$

Let the regression line of x on y be, 2x +3y – 8 = 0

Then, x = - $\frac{3}{2}$ y + 4, which is of the form x = cy + d where c = - $\frac{3}{2}$

$\therefore ac = \left(-\frac{1}{2}\right)\left(-\frac{3}{2}\right) = \frac{3}{4} < 1$, which indicates that the assumptions are correct

$byx = \left(-\frac{1}{2}\right)$ *and* $bxy = \left(-\frac{3}{2}\right)$

$\therefore r^2 = byx\,.\,bxy = \left(-\frac{1}{2}\right)\left(-\frac{3}{2}\right) = \frac{3}{4}$

$\therefore \mathbf{r = -\left(\frac{\sqrt{3}}{2}\right)}$

(3) To find y when x = 3 (4) x when y= 2

The regression line of y on x is given by, $y = -\frac{1}{2}x + \frac{5}{2}$

$\therefore$ When x = 3, the value of y is given by, $y = -\frac{1}{2}.3 + \frac{5}{2} = -\frac{3}{2} + + \frac{5}{2} = 1$

(4) To find x when y = 2

The regression line of x on y is given by, $x = -\frac{3}{2}y + 4$

$\therefore$ When y = 2, the value of x is given by, $x = -\frac{3}{2}.2 + 4 = -3 + 4 = 1$

(5) To find ${\sigma_x}^2$ We have $byx = \frac{r\sigma_y}{\sigma_x}$ $\longrightarrow$ ${\sigma_x}^2 = \frac{(r\sigma_y)^2}{(byx)^2} = \frac{r^2{\sigma_y}^2}{(byx)^2} = \frac{\frac{3}{4}.12}{\frac{1}{4}} = \mathbf{36}$

Question No.3: The two regression lines are given by the equation $ax+by+c = 0$. Show that the correlation between them is -1 if signs of a and b are alike and +1 if they are different.

Answer: Let the R.L of y on x be $ax + by + c = 0$, *which can be put in the form* $y = -\left(\frac{a}{b}\right)x - \frac{c}{b}$

Similarly, the R.L of x on y can be given as, $x = -\left(\frac{b}{a}\right)x - \frac{c}{a}$

$\therefore\ r^2 = byx\,.\,bxy = \left(-\frac{a}{b}\right)\left(-\frac{b}{a}\right) = 1$

$\therefore r = \pm 1$

$r = +1$ *iff* $\left(-\frac{a}{b}\right)$ *and* $\left(-\frac{b}{a}\right)$ *are both positive which is possible only when a and b have different signs.*

$r = -1$ *iff* $\left(-\frac{a}{b}\right)$ *and* $\left(-\frac{b}{a}\right)$ *are both negative which is possible only when a and b have the same sign.*

Question No.4: Calculate the two regression lines from the following data.

X:	24	40	36	45	55	30	50	43	53	44
Y:	50	78	72	100	100	54	116	88	102	90

Answer: From the above data set of n=10 pairs of observations, we can calculate :-

$\sum X = 420$, $\sum Y = 850$, $\sum X^2 = 18516$ $\sum Y^2 = 76388$, $\sum XY = 37482$

$$\bar{x} = \frac{\sum x}{n} = \frac{420}{10} = \mathbf{42}, \qquad \bar{y} = \frac{\sum y}{n} = \frac{850}{10} = \mathbf{85}$$

$$\text{Cov(x,y)} = \frac{\sum xy}{n} - \left(\frac{\sum x}{n}\right)\left(\frac{\sum y}{n}\right) = \frac{37482}{10} - \left(\frac{420}{10}\right)\left(\frac{850}{10}\right) = 3748.2 - 42\text{x}85 = \mathbf{178.2}$$

$$\sigma_x = \sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2} = \sqrt{\frac{18516}{10} - \left(\frac{420}{10}\right)^2} = \sqrt{1851.6 - 1764} = \sqrt{87.6} = \mathbf{9.36}$$

$$\sigma_y = \sqrt{\frac{\sum y^2}{n} - \left(\frac{\sum y}{n}\right)^2} = \sqrt{\frac{76388}{10} - \left(\frac{850}{10}\right)^2} = \sqrt{7638.8 - 7225} = \sqrt{413.8} = \mathbf{20.34}$$

R.L of y on x: $(y - \bar{y}) = \frac{\text{Cov(x,y)}}{\sigma_x^2}(x - \bar{x})$

$$(y - 85) = \frac{178.2}{87.6}(x - 42)$$

$$\mathbf{y = 2.03\,x - 0.44}$$

R.L of x on y: $(x - \bar{x}) = \frac{\text{Cov(x,y)}}{\sigma_y^2}(y - \bar{y})$

$$(x - 42) = \frac{178.2}{413.8}(y - 85)$$

$$\mathbf{x = 0.43y + 5.45}$$

Rank correlation: Rank correlation is the simple correlation coefficient between the ranks of two sets of observations where each set corresponds to a characteristic. For example, the score given by two judges to n participants in a dance competition. In fact there is no definite scale to measure the dance performance. But the judges give a score out of some total mark say K. In such a situation, we can rank the participants in the order or ranks. The correlation coefficient calculated using the ranks is called rank correlation coefficient.

Rank correlation coefficient is meaningful when the relative positions or ranks of individual objects are more meaningful or reliable or easy to explain than their actual measurements. The rank correlation measures the intensity of correlation between two sets of rankings, each set corresponds to one characteristic.

Spearman's Rank Correlation Coefficient: It is the Karl Pearsons product moment correlation coefficient between the ranks of two sets of observations.

Let $(x_1, y_1), (x_2, y_2), \ldots\ldots. (x_n, y_n)$ be the n pairs of observations on two characteristics. For example these pairs can be the score of n participants in two stage programmes say painting and fancy dress.

When we rank the n observations in each of these two sets we get n ranks from 1 to n.

Let the rank of (x_i, y_i) be (X_i, Y_i) where X_i and Y_i are some numbers from 1 to n.

Spearman's Rank Correlation Coefficient – Contd

Since X_i and Y_i are some numbers from 1 to n, we get

$$\bar{X}=\frac{X_1+X_2+\dots\dots\dots+X_n}{n}=\frac{1}{n}.\frac{n(n+1)}{2}=\frac{n+1}{2},\ \text{Similarly}\ \bar{Y}=\frac{Y_1+Y_2+\dots\dots\dots+Y_n}{n}=\frac{1}{n}.\frac{n(n+1)}{2}=\frac{n+1}{2}$$

$$\sigma_X{}^2=\frac{\sum X_i{}^2}{n}-\left(\frac{\sum X_i}{n}\right)^2=\frac{1}{n}.\frac{n(n+1)(2n+1)}{6}-\left(\frac{1}{n}.\frac{n(n+1)}{2}\right)^2=\frac{(n+1)(2n+1)}{6}-\left(\frac{(n+1)}{2}\right)^2=\frac{n^2-1}{12}$$

$$\sigma_Y{}^2=\frac{\sum Y_i{}^2}{n}-\left(\frac{\sum Y_i}{n}\right)^2=\frac{1}{n}.\frac{n(n+1)(2n+1)}{6}-\left(\frac{1}{n}.\frac{n(n+1)}{2}\right)^2=\frac{(n+1)(2n+1)}{6}-\left(\frac{(n+1)}{2}\right)^2=\frac{n^2-1}{12}$$

Let $d_i=(X_i-Y_i)=\left[\left\{X_i-\frac{n+1}{2}\right\}-\left\{Y_i-\frac{n+1}{2}\right\}\right]=[\{X_i-\bar{X}\}-\{Y_i-\bar{Y}\}]$

$$\sum \boldsymbol{d_i}^2=\sum[\{X_i-\bar{X}\}-\{Y_i-\bar{Y}\}]^2$$

$$=\sum(X_i-\bar{X})^2+\sum(Y_i-\bar{Y})^2-2\sum(X_i-\bar{X})(Y_i-\bar{Y})$$

$$=n\sigma_X{}^2+n\sigma_Y{}^2-2\,n\,\mathrm{Cov}\,(X,Y)=n\sigma_X{}^2+n\sigma_Y{}^2-2\,n\,r\,\sigma_X\sigma_Y$$

$$=n\left(\frac{n^2-1}{12}\right)+n\left(\frac{n^2-1}{12}\right)-n\,2\,r\sqrt{\left(\frac{n^2-1}{12}\right)}\sqrt{\left(\frac{n^2-1}{12}\right)}=n\left(\frac{n^2-1}{6}\right)(1-r)$$

$$\therefore 1-r=\frac{6\sum d_i{}^2}{n(n^2-1)} \longrightarrow \mathbf{r=1-\frac{6\sum d_i{}^2}{n(n^2-1)}}$$

Rank Correlation Coefficient when there are repeated ranks

Rank Correlation coefficient is calculated using the formula, $\mathbf{r = 1 - \frac{6\sum d_i^2}{n(n^2-1)}}$.

While deriving the above formula, it was assumed that no two individuals have the same value in either of the group. If there are repeated observations, all of them are assigned the same rank which is the average of the ranks which they would have been assigned if they were different. (Eg,. After the largest and second largest observations, let there be three repeated observations. If these three were different they would be given the ranks 3, 4 and 5. Average of these ranks is 4 and all the three observations are given the rank 4). When there are repeated ranks a correction factor is used for calculating the Spearman's rank correlation coefficient.

The formula for calculating the Spearman's rank correlation coefficient is given by,

$$r = 1 - \frac{6\sum d_i^2 + \frac{1}{12}\{(m_1^3 - m_1) + (m_2^3 - m_2) + (m_3^3 - m_3) + \ldots\ldots\ldots\ldots\}}{n(n^2-1)},$$ where m_i is the number of times a particular rank is repeated.

Correlation and Regression

Regression Analysis

Calculate the Rank Correlation Coefficient from the following data of marks obtained by 8 students in Mathematics and Physiscs

Physics:	15	20	27	13	45	60	20	75
Mathematics:	50	30	55	30	25	10	30	70

Answer: r = $\mathbf{1} - \frac{6\sum d_i{}^2 + \frac{1}{12}\{(m_1{}^3 - m_1) + (m_2{}^3 - m_2) + (m_3{}^3 - m_3) + \ldots\ldots\ldots\ldots\}}{n(n^2-1)}$

Note: Two students have got equal marks (20) for physics. If mark of these two students were slightly different, they would have got theRanks 5 and 6. So they were given the rank 5.5.

3 students got equal marks (30) in Mathematics. All of them are given the average of ranks 4,5,6. Which is same as 5.

There fore m takes two values 2 and 3.

Marks in Physics	Marks in Maths	Rank in Physics (X)	Rank in Maths (Y)	d = x-y	d^2
15	50	7	3	4	16
20	30	5.5	5	.5	.25
27	55	4	2	2	4
13	30	8	5	3	9
45	25	3	7	-4	16
60	10	2	8	-6	36
20	30	5.5	5	.5	.25
75	70	1	1	0	0
					81.5

$$r = \mathbf{1} - \frac{6\sum d_i{}^2 + \frac{1}{12}\{(m_1{}^3 - m_1) + (m_2{}^3 - m_2)\}}{n(n^2-1)}$$

$$= \mathbf{1} - \frac{6\,x\,81.5 + \frac{1}{12}\{(2^3 - 2) + (3^3 - 3)\}}{8(8^2-1)} = \mathbf{0.02}$$