STATISTICAL METHODS

Bivariate analysis:

It is one of the statistical analysis where two variables are observed. One variable here is dependent while the other is independent. These variables are usually denoted by X and Y. So, here we analyse the changes occured between the two variables

CURVE FITTING

The Method of Least Squares:

The method of Least Squares states that the best representative cure is that for which the sum of the squares of the residuals is a minimum.

Let the set of data points be (x_i , y_i), i = 1,2,3,….,n.

Suppose the curve y = f(x) is fitted to this data. Let the observed value at $x = x_i$ is y_i and the corresponding value on the curve is $f(x_i)$. Let e_i is the error of approximation at $x = x_i$, then we have $e_i = y_i - f(x_i)$ ------ (1)

Consider $S = [y_1 - f(x_1)]^2 + [y_2 - f(x_2)]^2 + \ldots\ldots\ldots + [y_n - f(x_n)]^2$

$= [e_1]^2 + [e_2]^2 + \ldots\ldots\ldots + [e_n]^2$, the method of least squares consists of minimizing S.

Fitting of a Straight Line:

Let the straight line to be fitted is y = a + bx --------- (1).

Suppose equ(1) is fitted to the set of points $(x_1, y_1), (x_2, y_2), \ldots\ldots\ldots\ldots (x_n, y_n)$.

Then by least squares method

$$S = [y_1 - (a + bx_1)]^2 + [y_2 - (a + bx_2)]^2 + \ldots\ldots\ldots + [y_n - (a + bx_n)]^2 = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$ our aim is to minimize S. For minimum value of S, $\frac{\partial S}{\partial a} = 0$ *and* $\frac{\partial S}{\partial b} = 0$

$$\frac{\partial S}{\partial a} = 0 \Rightarrow \sum_{i=1}^{n} 2[y_i - (a + bx_i)](-1) = 0 \Rightarrow \sum_{i=1}^{n} [y_i - (a + bx_i)] = 0 \Rightarrow \sum_{i=1}^{n} y_i - a\sum_{i=0}^{n} (1) - b\sum_{i=1}^{n} x_i = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i \ \ldots\ldots\ldots (2)$$

$$\frac{\partial S}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n} 2[y_i - (a + bx_i)](-x_i) = 0 \Rightarrow \sum_{i=1}^{n} [x_i y_i - (ax_i + bx_i^2)] = 0 \Rightarrow \sum_{i=1}^{n} x_i y_i = a\sum_{i=0}^{n} (x_i) + b\sum_{i=1}^{n} x_i^2$$

…(3).

Equation's (2) & (3) are normal equations for equation (1). Since x_i , y_i are known, solving (2)& (3) for a and b and substituting in equation (1), we get the required st.Line.

Note: Divide equation(2) with n, we get $\Rightarrow \frac{1}{n}\sum_{i=1}^{n} y_i = a + \frac{b}{n}\sum_{i=1}^{n} x_i \Rightarrow \bar{y} = a + b\bar{x}$ ----- (4)

comparing (1) & (4), show's that the Fitted line passing through the centroid $(\bar{x}, \bar{y})$ of the given points.

Problem: Fit a straight line to the following data by the method of least squares.

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

Solution: Let the straight line to be fitted is y = a + bx ------- (1)

Normal equations of equation (1) are

$\sum y = na + b\sum x$ -------- (2) and $\sum xy = a\sum x + b\sum x^2$ ---------- (3)

Table for calculation

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

Substituting the tabular values in equations (2) and (3), we get

16.9 = 5a + 10b and 47.1 = 10a + 30b Solving these two equations, we get a = 0.72 & b = 1.33

Substituting the values of a and b in equation (1), we y = 0.72 + 1.33x is the required fit.

Home work Problems:

1. Find the values of a and b for the law y = a + bx by the method of least squares which best fits the following data.

x	100	120	140	160	180	200
y	0.45	0.55	0.60	0.70	0.80	0.85

2. Certain experimental values of x and y are given below. Fit a straight line to the following data by the method of least squares.

X	0	2	5	7
y	-1	5	12	20

3. If V and R are related by a relation of the form $R = a + bV^2$, find a and b by the method of method of least squares.

V	10	20	30	40	50
R	8	10	15	21	30

CURVE FITTING FOR NON-LINEAR CURVES

FITTING OF A PARABOLA:

Let $y = a + bx + cx^2$ -------(1) be the equation of the parabola to be fitted to the data (x_i , y_i), $i = 1,2,3,\ldots\ldots,n$.

By the method of least squares $S = \sum_{i=1}^{n}(y_i - a - bx_i - cx_i^2)$, we have to determine a, b, c such that S is minimum. For minimum $\frac{\partial S}{\partial a} = \frac{\partial S}{\partial b} = \frac{\partial S}{\partial c} = 0$

$$\frac{\partial S}{\partial a} = 0 \Rightarrow \sum_{i=1}^{n} 2(y_i - a - bx_i - cx_i^2)(-1) = 0 \Rightarrow \sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} 1 - b\sum_{i=1}^{n} x_i - c\sum_{i=1}^{n} x_i^2 = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2 \quad \text{---------- (2)}$$

$$\frac{\partial S}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n} 2(y_i - a - bx_i - cx_i^2)(-x_i) = 0 \Rightarrow \sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2 - c\sum_{i=1}^{n} x_i^3 = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3 \quad \text{---------- (3)}$$

$$\frac{\partial S}{\partial c} = 0 \Rightarrow \sum_{i=1}^{n} 2(y_i - a - bx_i - cx_i^2)(-x_i^2) = 0 \Rightarrow \sum_{i=1}^{n} x_i^2 y_i - a\sum_{i=1}^{n} x_i^2 - b\sum_{i=1}^{n} x_i^3 - c\sum_{i=1}^{n} x_i^4 = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i^2 y_i = a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4 \quad \text{---------- (4)}$$

Equations (2), (3), & (4) are 3 equations in three unknowns a, b, c. these are called normal equations for a parabola. Solving these and placing in equation (1), we get the required parabola.

Problem: Fit a Parabola to the following data by the method of least squares.

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

Solution: From given data n = 5

Let the equation of the parabola to be fitted is $y = a + bx + cx^2$ -------- (1), its normal equations are

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2 \quad \text{------- (2),}$$

$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3$ ------ (3) and $\sum_{i=1}^{n} x_i^2 y_i = a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4$ ------ (4)

Table for calculations

x	0	1	2	3	4	$\sum x = 10$
y	1	1.8	1.3	2.5	6.3	$\sum y = 12.9$
xy	0	1.8	2.6	7.5	25.2	$\sum xy = 37.1$
x^2	0	1	4	9	16	$\sum x^2 = 30$
x^3	0	1	8	27	64	$\sum x^3 = 100$
x^2y	0	1.8	5.2	22.5	100.8	$\sum x^2 y = 130.3$
x^4	0	1	16	81	256	$\sum x^4 = 354$

Substituting tabular values in equation's (2), (3), and (4), we get

12.9 = 5a + 10b + 30c, 37.1 = 10a + 30b + 100c and 130.3 = 30a + 100b + 354c

Solving these equations using calculator, we get

A = 1.42, b = -1.07, c = 0.55

With these required parabola is $y = 1.42 - 1.07x + 0.55x^2$

Home work Problems:

1) By the method of least squares, fit a second degree curve, $y = a + bx + cx^2$ to the following data

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

2) By the method of least squares, fit a second degree curve, $y = a + bx + cx^2$ to the following data

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

FITTING OF THE CURVE OF THE FORM $y = ae^{bx}$

Let the curve to be fitted is $y = ae^{bx}$ -------- (1).

Take log on both sides,

we get $\log_e^y = \log_e^a + bx$ this is of the form Y = A + bx ---- (2) where $Y = \log y$ *and* $A = \log a$

Equation (2) is a straight line, its normal equations are

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

Solving equations (3) & (4) for A and b.

From A find a and substitute a and b in equation (1) then we get required fit.

Problem: Fit an exponential curve $y = ae^{bx}$ to the following data by the method of least squares.

x	2	3	4	5	6
y	144	172.8	207.4	248.8	298.6

Solution: From the given data n =5, let the given curve be $y = ae^{bx}$ ------ (1)

Take log on both sides, we get $\log_e^y = \log_e^a + bx$ this is of the form Y = A + bx ---- (2) where

$$Y = \log y \text{ and } A = \log a$$

Equation (2) is a straight line, its normal equations are

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

Table for calculation

x	2	3	4	5	6	$\sum x = 20$
y	144	172.8	207.4	248.8	298.6	
$Y = \log_e^y$	4.97	5.15	5.33	5.52	5.69	$\sum Y = 26.66$
xY	9.94	15.45	21.32	27.60	34.14	$\sum xY = 108.45$
x^2	4	9	16	25	36	$\sum x^2 = 90$

Substituting tabular values in equations (3) & (4), we get

26.66 = 5A +20b and 108.45 = 20A +90b

Solving these equations, we get A = 4.608, b = 0.181 implies $A = \log_e^a = 4.608, b = 0.181$

$\Rightarrow a = e^{4.608} = 100.28$ (approx.), b = 0.181

Placing the values of A and b in equation (1), we get the required curve is $y = 100.28x^{0.181x}$

Home work Problems:

1) Fit an exponential curve $y = ae^{bx}$ to the following data by the method of least squares.

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

2) Using the method of least squares determine the constants a and b such that $y = ae^{bx}$ fits the following data.

x	0	1	2	3	4	5	6	7	8
y	20	30	52	77	135	211	326	550	1052

FITTING OF THE CURVE OF THE FORM $y = ax^b$

Let the curve to be fitted is $y = ax^b$ -------- (1). Take log on both sides,

we get $\log y = \log a + b \log x$ this is of the form Y = A + bX ---- (2) where

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

Equation (2) is a straight line , its normal equations are

$$\sum Y = nA + b\sum X \text{ ------ (3) and } \sum XY = A\sum X + b\sum X^2 \text{ ------(4)}$$

Solving equations (3) & (4) for A and b. From A find a and substitute a and b in equation (1) then we get required fit.

Problem: Fit a curve $y = ax^b$ to the following data by the method of least squares.

x	2	4	7	10	20	40	60	80
y	43	25	18	13	8	5	3	2

Solution: From the given data n =8, let the given curve be $y = ax^b$ ------ (1)

Take log on both sides, we get $\log y = \log a + b \log x$ this is of the form Y = A + bX ---- (2)

where $Y = \log y,\ A = \log a,\ and\ X = \log x$

Equation (2) is a straight line , its normal equations are

$$\sum Y = nA + b\sum X \text{ ------ (3) and } \sum XY = A\sum X + b\sum X^2 \text{ ------(4)}$$

Table for calculation

x	2	4	7	10	20	40	60	80	
y	43	25	18	13	8	5	3	2	
$X = \log_e^x$	0.693	1.386	1.946	2.303	2.996	3.689	4.094	4.382	$\sum X = 21.216$
$Y = \log_e^y$	3.76	3.21	2.89	2.56	2.07	1.609	1.099	0.693	$\sum Y = 0.693$

	1	9	0	5	9				
XY	2.606	4.462	5.624	5.907	6.229	5.936	4.499	3.037	$\sum XY = 38.3$
X^2	0.480	1.921	3.787	5.304	3.976	13.609	16.761	19.202	$\sum X^2 = 70.04$

Substituting tabular values in equations (3) & (4), we get

17.915 = 8A +21.216b and 38.3 = 21.216A +70.4b

Solving these equations, we get A = 4.09, b = -0.7 implies $A = \log_e^a = 4.09, b = -0.7$

$\Rightarrow a = e^{4.09} = 60$ (appro), b = -0.7

Placing the values of A and b in equation (1), we get the required curve is $y = 60x^{-0.7}$

Home work problems:

1) Fit a power function $y = ax^b$ to the following data by the method of least squares.

x	1	2	3	4	5
y	0.5	2	4.5	8	12.5

FITTING OF THE CURVE OF THE FORM $y = ab^x$

Let the curve to be fitted is $y = ab^x$ -------- (1). Take log on both sides,

we get $\log_e^y = \log_e^a + x\log_e^b$ this is of the form Y = A + Bx ---- (2)

where $Y = \log y, A = \log a \text{ and } B = \log_e^b$

Equation (2) is a straight line , its normal equations are

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

Solving equations (3) & (4) for A and B. From A,B find a,b and substitute a and b in equation (1) then we get required fit.

Problem: Obtain a relation of the form $y = ab^x$ for the following data by the method of least squares

x	2	3	4	5	6
y	8.3	15.4	33.1	65.2	127.4

Solution: From given data n = 5. Let the curve to be fitted is $y = ab^x$ -------- (1). Take log on both sides, we get $\log_e^y = \log_e^a + x\log_e^b$ this is of the form Y = A + Bx ---- (2)

where $Y = \log y, A = \log a \text{ and } B = \log_e^b$

Equation (2) is a straight line, its normal equations are

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

Table for calculation

x	2	3	4	5	6	$\sum x = 20$
y	8.3	15.4	33.1	65.2	127.4	
$Y = \log_e^y$	0.9191	1.1875	1.5198	1.8142	2.1052	$\sum Y = 7.5455$
xY	1.8382	3.5625	6.0792	9.0710	12.631	$\sum xY = 33.1819$
x^2	4	9	16	25	36	$\sum x^2 = 90$

Substituting tabular values in equations (3) & (4), we get

7.5455 = 5A +20B and 33.1819 = 20A +90B

Solving these equations, we get A = 0.31, B = 1.3 implies $A = \log_e^a = 0.31, B = \log b = 1.3$

$\Rightarrow a = e^{0.31} = 2.04$ and $b = e^{1.3} = 1.995$

Placing the values of a and b in equation (1), we get the required curve is $y = 2.04.(1.995)^x$

Home work Problems:

1) Fit an exponential curve $y = ab^x$ to the following data by the method of least squares.

x	0	1	2	3	4	5	6	7
y	10	21	35	59	92	200	400	610

2) Using the method of least squares determine the constants a and b such that $y = ab^x$ fits the following data.

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

CORRELATION

INTRODUCTION

Correlation refers to the relationship of two or more variables. We know that there exists relationship between the heights of a Father and a son, wage and price index. The study of the relation is called correlation. It measures the closeness of the relationship between the variables.

DEFINITION

Correlation is a statistical analysis which measures and analyses the degree or extent to which two variables fluctuate with reference to each other.

The correlation expresses the relationship or interdependence of two sets of variables upon each other. One variable may be called the subject (independent) and the other relative (dependent).

TYPES OF CORRELATION

Correlation is classified into many types.

1. Positive and negative
2. Simple and multiple
3. Partial and total
4. Linear and non – linear

1. Positive and Negative Correlation:

Positive and Negative correlation depend upon the direction of change of the variables. If two variables tend to move together in the same direction i.e. an increase in the value of one variable is accompanied by an increase in the value of the other variable; or a decrease in the value of one variable is accompanied by a decrease in the value of the other variable, then the correlation is called **positive or direct correlation.** Height and weight, rainfall and yield of crops, price and supply are examples of positive correlation.

If two variables tend to move together in opposite directions so that an increase or decrease in the values of one variable is accompanied by a decrease or increase in the value of the other variable, then the correlation is called negative or inverse correlation.

METHODS OF STUDYING CORRELATION

There are 2 different methods for finding out the relationship between variables. They are (1) Graphic methods (2) Mathematical method.

1.**Graphic method**

Scatter diagram or scatter gram

2. Mathematical method is

a) Karl' Person's coefficient of correlation
b) Spearman's Rank coefficient of correlation
c) Coefficient of concurrent deviation
d) Method of least squares

SCATTER DIAGRAM OR SCATTERGRAM

The scatter diagram is a chart obtained by plotting two variables to find out whether there is any relationship between them. In this diagram X variables are plotted on the horizontal axis and Y variables are plotted on the vertical axis. Thus we can know the scatter or concentration of various points. Various scatter diagrams are briefly shown here.

ADVANTAGES OF SCATTER DIAGRAM

1. Scatter diagram is a simple, attractive method to find out the nature of correlation .
2. It is easy to understand.
3. A rough idea is got at a glance whether it is positive or negative correlation.

COEFFICIENT OF CORRELATION

Correlation is a statistical technique used for analyzing the behavior of two or more variables. Its analysis deals with the association, between two or more variables. Statistical measures of correlation relates to co variation between series but not of function or casual relationship.

KARL PEARSON'S COEFICIENT OF CORRELATION

Karl Pearson (1867-1936) a British Biometrician and Statistician suggested a mathematical method for measuring the magnitude of linear relationship between two variables. This is known as Pearson Coefficient of correlation. It is denoted by 'r'. This method is most widely used. It is also called **Product- Moment correlation coefficient.**

These are several formulae to calculate r. They are

$$(1) r = \frac{\text{covariance of } xy}{\sigma_x \times \sigma_y} \qquad (2) r = \frac{\sum xy}{N\sigma_x\sigma_y} \qquad (3) r = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}}$$

$$(4) r = \frac{(\sum xy \times N) - (\sum x \times Y)}{\sqrt{(\sum x^2 \times N - (\sum x)^2).(\sum y^2 \times N - (\sum y)^2)}}$$

$X = (x - \overline{X}), Y = (y - \overline{Y})$ *where* $\overline{X}, \overline{Y}$ are means of the series x and y.

σ_x = standard deviation of series x.

σ_y = standard deviation of series y.

We will be using formula (3) more. We now list some of properties of r.

PROPERTIES OF CORRELATION COEFFICIENT

Property 1: Show that the maximum value of rank correlation coefficient is 1.

(OR)

The coefficient of correlation lies between – 1 and + 1. Symbolically, $-1 \le r \le +1 \quad or \mid r \mid \le 1$

Proof: Let x and y be deviations of X and Y series from their mean.

Let σx and σy be their respective standard deviations.

$$Let = \sum\left(\frac{x}{\sigma_x} + \frac{y}{\sigma_y}\right)^2 = \sum\left[\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} + \frac{2xy}{\sigma_x\sigma_y}\right] = \frac{\sum x^2}{\sigma_x^2} + \frac{\sum y^2}{\sigma_y^2} + \frac{2\sum xy}{\sigma_x\sigma_y} - - - (1)$$

But $\frac{\sum x^2}{\sigma_x^2}$ = N. Similarly $\frac{\sum y^2}{\sigma_y^2}$ =N -------- (2)

Again, $r = \frac{\sum xy}{N\sigma_x\sigma_y} \Rightarrow Nr = \frac{\sum xy}{\sigma_x\sigma_y}$

From (1), (2) and (3)

$$\Rightarrow \sum\left(\frac{x}{\sigma_x}+\frac{y}{\sigma_y}\right)^2 = N+N+2Nr$$

$$= 2N+2Nr = 2N(1+r), 1+r \geq 0 \Rightarrow r \geq -1$$

But $\left(\frac{x}{\sigma_x}+\frac{y}{\sigma_y}\right)^2$ is the sum of squares of real quantities and as such it can not be negative; at the most it can be zero= $2N(1+r) \geq 0$

Hence r cannot be less than – 1; at the most it can be -1.

Similarly by expanding $\sum\left(\frac{x}{\sigma_x}+\frac{y}{\sigma_y}\right)^2$ it can be shown that $\sum\left(\frac{x}{\sigma_x}-\frac{y}{\sigma_y}\right)^2 = 2N(1-r)$

This again cannot be negative; at the most it can be zero

$$\Rightarrow 2N(1+r) \geq 0$$

$$\Rightarrow 1-r \geq 0$$

$$\Rightarrow 1 \geq r$$

$$\Rightarrow r \leq 1$$

Hence r cannot be greater than +1; at the most it can be +1.

Hence $-1 \leq r \leq 1$.

Note: Limits for correlation coefficient are $-1 \leq r \leq 1$.

Hence correlation coefficient can not exceed one numerically.

If r =1 correlation is perfect and positive.

If r = – 1 correlation is perfect and negative. If r = 0, then there is no relationship between the variables.

Property II:

The coefficient of correlation is independent of the change of origin and scale of measurements.

Proof: We change the origin and scale of both the variables x and y.

Let $u_i = \frac{x_i - A}{h}$ *and* $v_i = \frac{y_i - B}{k}$, where the constants A and B refer to the change of origin, and the constant h and k refer to the change of scale.

Also $A \geq 0, B \geq 0, h>0$, k>0

$$\therefore x_i = A+hu_i \Rightarrow \bar{x} = A+h\bar{u} \Rightarrow x-\bar{x} = h(u_i-\bar{u}) \ldots (1)$$

$$y_i = B+kv_i \Rightarrow \bar{y} = B+k\bar{v} \Rightarrow y-\bar{y} = k(v_i-\bar{v}) \ldots\ldots (2)$$

The coefficient of correlation between X and Y is

$$r_{xy} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2 \times \sum(y_i-\bar{y})^2}} \text{------} (3)$$

From (1), (2) and (3)

$$r_{xy} = \frac{\sum h(u_i-\bar{u})k(v_i-\bar{v})}{\sqrt{\sum h^2(u_i-\bar{u})^2 \times \sum k^2(v_i-\bar{v})^2}}$$

$$= \frac{\sum(u_i-\bar{u})(v_i-\bar{v})}{\sqrt{\sum(u_i-\bar{u})^2 \times \sum(v_i-\bar{v})^2}} = r_{uv}$$

Thus the correlation between X and Y is equal to correlation between u and v, where u and v are variables obtained by change of origin and scale of the variables x and y. Hence the coefficient of correlation is independent of the change of origin and scale of measurement.

Property III: If X, Y are random variables and a, b, c, d are any numbers such that $a \neq 0, c \neq 0$ Then

$$r(ax+b, cY+d) = \frac{ac}{|ac|} r(X,Y)$$

Property IV: Two independent variables are uncorrelated. That is if X and Y are independent variables then r (X, Y) = 0.

Proof: If X and Y are independent then $\sum (X-\bar{x})(Y-\bar{y})$ *or* Cor (X, Y) will be zero and hence $r_{XY} = 0$. Thus if X and Y are independent variates. Then they are uncorrelated.

Note: The converse of the above is not true, i.e., two uncorrelated variables may not be independent as the following examples.

e.g.1

X	-3	–2	–1	1	2	3	$\sum X = 0$
Y	9	4	1	1	4	9	$\sum Y = 28$
XY	–27	–8	– 1	1	8	27	$\sum X$ Y=0

$$\bar{X} = \frac{1}{n}\sum X = 0, \text{Cov } (X,Y) = \frac{1}{n}\sum XY - \bar{X}\bar{Y} = 0$$

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

Thus X and Y are uncorrelated. But on examination, we can see that X and Y are connected by the relation $Y = X^2$ and hence are not independent.

EX.2. Consider the following data:

x	1	2	3	4	5	6	7
y	9	4	1	0	1	4	9

Here $\sum x = 28, \ \sum y = 28, \ \sum xy = 112$

$$\therefore cov \quad (x,y) = \frac{1}{n}\left[\sum xy - \frac{(\sum x)(\sum y)}{n}\right] = \frac{1}{7}\left[112 - \frac{28\times 28}{7}\right] = 0$$

Thus, $r_{xy} = 0$. But the variables x and y are related by the relation $y = (x-4)^2$ Hence the converse is not true.

Example: Calculate coefficient of correlation from the following data

x	12	9	8	10	11	13	17
y	14	8	6	9	11	12	3

Solution: In both series items are in small number.

So there is no need to take deviations.

We use the formula $r = \frac{Cov(X,Y)}{\sigma_{(X)}\sigma_{(Y)}}$ or

$$r = \frac{(\sum xy \times N) - (\sum x \times Y)}{\sqrt{(\sum x^2 \times N - (\sum x)^2).(\sum y^2 \times N - (\sum y)^2)}}$$

Computation of coefficient of correlation

X	Y	X^2	Y^2	XY
12	14	144	196	168
9	8	81	64	72
8	6	64	36	48
10	9	100	81	90
11	11	121	121	121
13	12	169	144	156
7	3	49	9	21
$\sum X$ = 70	$\sum Y$ = 63	$\sum X^2$ = 728	$\sum Y^2$ = 651	$\sum X$ Y= 676

$$r = \frac{(\sum XY \times N) - (\sum X \times Y)}{\sqrt{(\sum X^2 \times N - (\sum X)^2).(\sum Y^2 \times N - (\sum Y)^2)}}$$

Here N = 7

$$r = \frac{(676 \times 7) - (70 \times 63)}{\sqrt{(728 \times 7 - (70)^2}\sqrt{651 \times 7 - (63)^2}}$$

$$= \frac{4732 - 4410}{\sqrt{5096 - 4900}\sqrt{4557 - 3969}}$$

$$= \frac{322}{\sqrt{196 \times 588}} = \frac{322}{339.48} = +0.95$$

Example 2 : Find if there is any significant correlation between the heights and weights given below.

Height in inches	57	59	62	63	64	65	55	58	57
Weight in lbs	113	117	126	126	130	129	111	116	112

Solution :

Computation of coefficient of correlation

Height in inches x	Deviation from Mean (60) $X = x - \bar{x}$	Square of deviations X^2	Weight in lbs y	Deviations from Mean $Y = y - \bar{y}$	Square of deviations Y^2	Product of deviations of X and Y series (XY)
57	−3	9	113	−7	49	21
59	−1	1	117	−3	9	3
62	2	4	126	6	36	12
63	3	9	126	6	36	18
64	4	16	130	10	100	40
65	5	25	129	9	81	45
55	−5	25	111	−9	81	45
58	−2	4	116	−4	16	8
57	−3	9	112	−8	64	24
540	0	102	1080	0	472	216

$$\text{Coefficient of correlation } r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \times \Sigma Y^2}}$$

$$\therefore \quad r = \frac{216}{\sqrt{102 \times 471}} = 0.98$$

WHEN DEVIATIONS ARE TAKEN FROM AN ASSUMED MEAN

When actual mean is not a whole number, but a fraction or when the series is large, the calculation by direct method will involve a lot of time. To avoid such tedious calculation, we can use the assumed mean method.

$$\therefore \textit{Formula} = \frac{\sum XY - \frac{\Sigma X \Sigma Y}{N}}{\sqrt{\left[\sum X^2 - \frac{(\Sigma X)^2}{N}\right]\sqrt{\sum Y^2 - \frac{(\Sigma Y)^2}{N}}}}$$

Where

X = Deviation of the items of x- series from an assumed mean i.e. X = (x -A)

Y = Deviation of the items of x- series from an assumed mean i.e. X = (x -A)

N = Number of items

$\sum X$ Y= the total of the product of the deviations of x and y-series from their assumed mean.

$\sum X^2$ = The total of the squares of the deviations of x- series from an assumed mean.

$\sum Y^2$ = The total of the squares of the deviations of y – series from an assumed mean.

$\sum X \sum Y$ = Total of the deviations of x – series from assumed mean.

$\sum Y$ = Total of the deviations of y – series from assumed mean.

EXAMPLE: Calculate Karl Pearson's correlation coefficient for the following paired data.

X	28	41	40	38	35	33	40	32	36	33
Y	23	34	33	34	30	26	28	31	36	38

Solution: Computation of correlation coefficient

X	Deviation from assumed mean	Square of deviation		Deviation from assumed mean	Square of deviation	Product of deviation
	x =x – 35 $(X = x - \bar{x})$	X^2	y	Y = y -31 $(Y = y - \bar{y})$	Y^2	X Y
28	- 7	49	23	-8	64	56
41	+ 6	36	34	+ 3	9	18
40	+ 5	25	33	+2	4	10
38	+3	9	34	+3	9	9
35	0	0	30	-1	1	0
33	-2	4	26	-5	25	10
40	+5	25	28	-3	9	-15
32	-3	9	31	0	0	0
35	+1	1	36	+5	25	5
33	-2	4	38	+7	49	-4
$\sum x = 355$ $\sum X = 6$		$\sum X^2 =$ 162	$\sum y = 313$	$\sum Y = 3$	$\sum Y^2 =$ 195	$\sum X$ Y= 79

N = 10, Take $\bar{x} = 35$ and $\bar{y} = 31$

Applying to the above data, the formula

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left[\sum X^2 - \frac{(\sum X)^2}{N}\right] \times \left[\sum Y^2 - \frac{(\sum Y)^2}{N}\right]}}$$

$$= \frac{7 - \frac{6 \times 3}{10}}{\sqrt{(162 - \frac{6^2}{10})(195 - \frac{3^2}{10})}}$$

$$= \frac{77.2}{\sqrt{1584}\sqrt{194.9}} = 0.45$$

REGRESSION

INTRODUCTION

The study of correlation measures the direction and strength of the relationship between two variables. In correlation we can estimate the value of one variable, when the value of the other variable is given. Since price and supply are correlated, we can find out the expected amount of supply for a given price or we can find out the required price to get a given amount of supply.

In regression, we can estimate the value of one variable with the value of the other variable which is known . The statistical method which helps us to estimate the unknown value of one variable from the known value of the related variable is called regression. The line described in the average relationship between two variables is known as line of regression. Now a day we are using the term estimating line instead of regression line.

USES

1.It is used to estimate the relation between two economic variables like Income and Expenditure.

2.It is highly valuble tool in Economic and Business.

3.Widey used for prediction purpose.

4.We can calculate coefficient of correlation and coefficient of determination with the help of the regression coefficient.

5. It is useful in statistical estimation of demand curves, supply curves, production function, cost function and consumption function etc.

COMPAEISON BETWEEN CORRELATION AND REGRESSION

The correlation coefficient is a measure of degree of covariability between two variables, while the regression establishes a functional relation between dependent and independent variables, so that the former can be prodected for a given value of the later. In correlation, both the variables x and y are random variables, where as in Regression, x is a random variable andy is fixed variable. The coefficient of correlation is a relative measure where as Regression coefficient is an absolute figure.

METHODS OF STUDYING REGRESSION

We have two methods for studying regression

1) Graphic method 2) Algebraic method.

1. Graphic Method :

In this method, the points representing the pairs of values of the variables are plotted on a graph. The independent variable is taken on X- axis and the dependent variables on

Y- axis. These points from a scatter diagram. A regression line is drawn between these points by free hand.

e.g. Fit a regression line on the scatter diagram for the following data.

X	Y
65	68
67	68
62	66
70	68
67	67
69	68
71	70

2. Algebraic Method:

Regression Line: A regression line is a straight line fitted to the data by the method of least squares. It indicates the best possible mean value of one variable corresponding to the mean

value of the other . There are always two regression lines con structed for the relationship between two variables X and Y. Thus one regression line shows the regression of X upon Y and the other shows the regression of Y on X.

REGRESSION EQUATION

Regression equation is an algebraic expression of the regression line. It can be classified into Regression equation, Regression coefficient, individual observation and group discussion.

The standard Form of the Regression equation is Y = a + b X where a, b are called constants . 'a' indicates the value of Y when X =0,It is called Y- interpet. 'b' indicates the value of slope of the regression line and gives a measure of change of Y for a unit change in X. It is also called the Regression coefficient of Y on X. Thus if we know the value of a and b we can easily compute the value of Y for any given value of X. The values of a and b are found with the help of the following Normal equations.

Regression equation of Y on X :

Regression equation X on Y:

Normal equations are

$$\sum X = Na + b\sum Y$$

$$\sum XY = a\sum Y + b\sum Y^2$$

Example 1: Determine the equation of a straight line which best fits the data.

X:	10	12	12	16	17	20	25
Y:	10	22	24	27	29	33	37

Solution: Straight line is Y = a + bx

The two normal equations are $\sum Y = b\sum X + Na$

$$\sum XY = b\sum X^2 + a\sum X$$

X	X^2	Y	XY
10	100	10	100
12	144	22	264
13	169	24	312
16	256	27	432
17	289	29	493
20	400	33	660
25	625	37	925
$\sum X = 113$	$\sum X^2 = 1938$	$\sum Y$ =182	$\sum XY = 3186$

Substituting the values, we get

113 b +7 a = 182 ……(1)

1983b +113a=3186 ……..(2)

Solving the equation of the straight lines Y= a + b X

:. Y = 0.82 + 1.56 X

This is called the regression equation of Y on X.

DEVITATION TAKEN FROM ARITHMETIC MEAN OF X AN Y

This method is easier and simpler than the previous method to find the values of a and b. We can find out the deviations of X and Y series from their respective means.

Regression equation of X on Y

$$X - \overline{X} = r\frac{\sigma_x}{\sigma_y}(Y - \overline{Y})$$

where $\overline{X}$ = Mean of X series, $\overline{Y}$ = Mean of Y series

the regression coefficient of X on Y = $r\frac{\sigma_x}{\sigma_y} = \frac{\sum XY}{\sum Y^2} = b_{XY}$

Regression equation of Y on X

$$Y - \bar{Y} = r\frac{\sigma_x}{\sigma_y}(X - \bar{X})$$

The regression coefficient of Y on X = $b_{YX} = \frac{\sum XY}{\sum X^2} = r\frac{\sigma_y}{\sigma_x}$

Thus $r^2 = b_{XY} \times b_{YX}$

EXAMPLE: If θ is the angle between two regression lines and S.D. of Y is twice the S.D.of X and r = 0.25, find $\tan\theta$.

Solution: Given $\sigma_y = 2\sigma_x$. and r =0.25

If θ is the angle between two regression lines, then

$$\tan\theta = \frac{1-r^2}{r}\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2} = \frac{1-(0.25)^2}{0.25}\frac{\sigma_x 2\sigma_x}{\sigma_x^2+4\sigma_x^2}$$

$$= \frac{1-0.0625}{0.25}.\frac{3}{5} = 3.75\times\frac{2}{5} = 1.5$$

Problems:

Example 8 : The following data, based on 450 students, are given for marks in statistics are Economics at a certain examination.

Mean marks in statistics = 40

Mean marks in Economics = 8

S. D. of marks in statistics = 12

Variance of marks (Economics) = 256

Sum of the products of deviations of marks from this respective mean 42075. Give the equations of the two lines of regression and estimate the average marks in Economics of candidates who obtained 50 marks in statistics.

Solution : Given $\overline{X}$ = mean of marks in statistics = 40

$\overline{Y}$ = mean of marks in Economics = 8

σ_Y = S. D. of marks in statistics = 12

σ_Y = S. D. of marks in Economics = 16

Co.eft. of correlation $r = \frac{\sum XY}{N\sigma_X\sigma_Y} = \frac{42075}{450\times12\times16} = 0.49$

Regression equation of X on Y, $X - \bar{X} = \frac{r\sigma_X}{\sigma_Y}(Y - \bar{Y})$

$$\Rightarrow X = 0.37Y + 22.24$$

Regression equation of Y on X, $Y - \bar{Y} = \frac{r\sigma_X}{\sigma_Y}(X - \bar{X})$

$$\Rightarrow Y = 0.65X + 22$$

when $X = 50$ then $Y = 54.5$

Example 6 : Calculate the regression equations of Y on X from the data given below, taking deviations from actual means of X and Y.

Price (RS.)	10	12	13	12	16	15
Amount Demanded	40	38	43	45	37	43

Estimate the likely demand when the price is Rs. 20.

Solution : Calculation of Regression equation

x	$(x-13) = X$	X^2	y	$(y-41) = Y$	Y^2	XY
10	–3	9	40	–1	1	3
12	–1	1	38	–3	9	3
13	0	0	43	2	4	0
12	–1	1	45	4	16	–4
16	3	9	37	–4	16	–12
15	2	4	43	2	4	4

Regression equation of Y on X is $Y - \bar{Y} = r\frac{\sigma_y}{\sigma_x}(X - \bar{X})$ .... (1)

Now $r\frac{\sigma_y}{\sigma_x} = \frac{\Sigma XY}{\Sigma X^2} = -0.25$

$Y - 41 = -0.25(X - 13)$ [by (1)] $\Rightarrow Y = -0.25X + 44.25$

When X is 20; Y = 39.25

When the price is Rs. 20, the likely demand is 39.25.

Example 3 : Find the most likely production corresponding to a rainfall 40 from the following data: **[JNTU (H) Nov. 2010 (Set No.2)]**

	Rain fall (X)	Production (Y)
Average	30	500 Kgs
Standard deviation	5	100 Kgs
Coefficient of correlation	0.8	

Solution : We have to calculate the value of Y when X = 40.

So we have to find the regression equation of Y on X.

Mean of X series, $\bar{X} = 30$; Mean of Y series, $\bar{Y} = 500$

σ of X series, $\sigma_x = 5$; σ of Y series, $\sigma_y = 100$

Regression of Y on X

$$Y - \bar{Y} = r \cdot \frac{\sigma_x}{\sigma_y}(X - \bar{X})$$

$$\Rightarrow Y - 500 = (0.8)\frac{5}{100}(X - 30)$$

When $X = 40$, $Y - 500 = \frac{4}{100}(40 - 30) = \frac{40}{100}$

$$\Rightarrow \quad Y = 500 + \frac{4}{10} = 500.4$$

Hence the expected value of Y is 500.4 kg.

Example 6 : Calculate the regression equations of Y on X from the data given below, taking deviations from actual means of X and Y.

Price (RS.)	10	12	13	12	16	15
Amount Demanded	40	38	43	45	37	43

Estimate the likely demand when the price is Rs. 20.

Solution : Calculation of Regression equation

x	$(x-13)=X$	X^2	y	$(y-41)=Y$	Y^2	XY
10	–3	9	40	–1	1	3
12	–1	1	38	–3	9	3
13	0	0	43	2	4	0
12	–1	1	45	4	16	–4
16	3	9	37	–4	16	–12
15	2	4	43	2	4	4

Regression equation of Y on X is $Y-\bar{Y}=r\frac{\sigma_y}{\sigma_x}(X-\bar{X})$ .... (1)

Now $r\frac{\sigma_y}{\sigma_x}=\frac{\sum XY}{\sum X^2}=-0.25$

$Y-41=-0.25(X-13)$ [by (1)] $\Rightarrow$ $Y=-0.25X+44.25$

When X is 20; Y = 39.25

When the price is Rs. 20, the likely demand is 39.25.

Example 6 : Calculate the regression equations of Y on X from the data given below, taking deviations from actual means of X and Y.

Price (RS.)	10	12	13	12	16	15
Amount Demanded	40	38	43	45	37	43

Estimate the likely demand when the price is Rs. 20.

Solution : Calculation of Regression equation

x	$(x-13)=X$	X^2	y	$(y-41)=Y$	Y^2	XY
10	–3	9	40	–1	1	3
12	–1	1	38	–3	9	3
13	0	0	43	2	4	0
12	–1	1	45	4	16	–4
16	3	9	37	–4	16	–12
15	2	4	43	2	4	4

Regression equation of Y on X is $Y-\bar{Y}=r\frac{\sigma_y}{\sigma_x}(X-\bar{X})$ .... (1)

Now $r\frac{\sigma_y}{\sigma_x}=\frac{\Sigma XY}{\Sigma X^2}=-0.25$

$Y-41=-0.25\,(X-13)$ [by (1)] $\Rightarrow$ $Y=-0.25X+44.25$

When X is 20; Y = 39.25

When the price is Rs. 20, the likely demand is 39.25.

Example 4 : Find the mean values of the variable X and Y and correlation coefficient from the following regression equations.

$$2Y - X - 50 = 0$$

$$3Y - 2X - 10 = 0$$

Solution : $2Y - X = 50$ ... (1)

$3Y - 2X = 10$ .... (2)

$4Y - 2X = 100$ .. (3)

Solving (2) and (3) $Y = 90$.

Substituting the value of Y in (1), we get $X = 130$.

$\therefore \bar{x} = 130$ and $\bar{y} = 90$

Rewriting (1) and (2)

$$Y = \frac{1}{2}X + 25 \text{ and } X = \frac{3}{2}Y - 5,$$

$r = \frac{\sigma_X}{\sigma_Y} = \frac{3}{2}$ and $r = \frac{\sigma_Y}{\sigma_X} = \frac{1}{2}$

$\therefore r^2 = \frac{3}{4}$

$\Rightarrow r = 0.866$

Example 6 : If $X = 2Y + 3$ and $Y = kX + 6$ are the regression lines of X on Y and Y on X respectively.

(*a*) Show that $0 \le k \le \frac{1}{2}$ (*b*) If $k = \frac{1}{8}$ find r and $(\bar{x}, \bar{y})$ **[JNTU (H) May 2011 (Set No.1)]**

Solution : (*a*) Given $X = 2Y + 3$ is the regression line of X on Y.

$$\therefore \frac{r\sigma_X}{\sigma_Y} = 2$$

The regression line of Y on X is $Y = kX + 6$

$$\therefore \frac{r\sigma_Y}{\sigma_X} = k$$

Multiplying, These 2 equations we get $r^2 = 2k$

We have $0 \le r^2 \le 1 \Rightarrow 0 \le 2k \le 1$

$$\Rightarrow 0 \le k \le \frac{1}{2}$$

(*b*) If $k = \frac{1}{8}$ then $y = \frac{1}{8}x + 6$

$$r^2 = 2k = \frac{2}{8} = \frac{1}{4}$$

$r = \frac{1}{2}$ since both regressive coefficients are postive, we take positive values for *r*.

$$\therefore r = \frac{1}{2}$$

Since the regression lines pass through $(\bar{x}, \bar{y})$ we have

$\bar{x} = 2\bar{y} + 3$ .... (1) and $\bar{y} = \frac{1}{8}x + 6$ .... (1)

From (1) $\bar{y} = \frac{1}{8}(2\bar{y} + 3) + 6$

$$= \frac{2\bar{y} + 3 + 48}{8}$$

$8\bar{y} = 2\bar{y} + 51 \Rightarrow 6\bar{y} = 51$

$$\Rightarrow \bar{y} = \frac{51}{6} \text{ and } \bar{x} = 2\bar{y} + 3 = \frac{102}{6} + 3 = \frac{120}{6} = 20$$

Example 9 : The equations of two regression lines are $7x-16y+9=0$ and $5y-4x-3=0$. Find the coefficient of correlation and the means of x and y.

(JNTU (H) May 2013)

Solution : Given equations are,

$7x-16y+9=0$ ... (1) $\qquad 5y-4x-3=0$ ... (2)

(1)$\times 4$ gives $28x-64y+36=0$

(2)$\times 7$ gives $-28x+35y-21=0$

Adding, $-29y+15=0 \Rightarrow y=\frac{15}{29}=0.5172$

$(1) \Rightarrow 7x=16y-9=\frac{240}{29}-9=\frac{240-261}{29}=\frac{-21}{29}$

$\therefore \quad x=\frac{-3}{29}=0.1034$

Since the regression line passes through $(\bar{x},\bar{y})$ we have

$\therefore \bar{x}=$ mean $=0.1034$

$\bar{y}=$ mean $=0.5172$

From (1), $x=\frac{16}{7}y-\frac{9}{7}$

From (2), $y=\frac{4}{5}x+\frac{3}{5}$

$\therefore r\frac{\sigma_x}{\sigma_y}=\frac{16}{7}$ and $r\frac{\sigma_y}{\sigma_x}=\frac{4}{5}$

Multiplying these two equations, we get

$r^2=\frac{16}{7}\cdot\frac{4}{5}=\frac{64}{35} \Rightarrow r=\frac{8}{\sqrt{35}}$.

----------THE END----------