PARUL UNIVERSITY
FACULTY OF ENGINEERING AND TECHNOLOGY
DEPARTMENT OF APPLIED SCIENCE AND HUMANITIES
4th SEMESTER B. TECH PROGRAMME
PROBABILITY, STATISTICS AND NUMERICAL METHODS (303191251)
ACADEMIC YEAR 2025-2026
UNIT: 1 CORRELATION, REGRESSION AND CURVE FITTING

CORRELATION

Correlation is the relationship that exists between two or more variables. Two variables are said to be correlated if a change in one variable affects a change in the other variable.

EXAMPLE:

1. Relationship between heights and weights.
2. Relationship between price and demand of commodity.
3. Relationship between rainfall and yield of crops.

Types Of Correlations

1. Positive and Negative correlations.
2. Simple and Multiple correlations.
3. Partial and Total correlations.
4. Linear and Non-linear correlations.

POSITIVE AND NEGATIVE CORRELATIONS

POSITIVE CORRELATIONS (Same direction)

If both the variables vary in the same direction, the correlation is said to be positive. In other words, if the value of one variable increases, the value of the other variable also increases. Same decreases.

Height (cm)	120	130	135	140	145
Weight(kg)	50	55	60	65	70

NEGATIVE CORRELATIONS (Opposite direction)

If both the variables vary in the opposite direction, the correlation is said to be negative. In other words, if the value of one variable increases, the value of other variable decreases.

Height (cm)	120	130	135	140	145
Weight(kg)	70	65	60	55	50

SIMPLE AND MULTIPLE CORRELATIONS

1. Simple Correlation: -

When only two variables are studied, the relationship is described as simple correlation, e.g., the quantity of money and price level, demand and price, etc.

2. Multiple Correlation: -

When more than two variables are studied, the relationship is described as multiple correlation, e.g., relationship of price, demand, and supply of a commodity.

PARTIAL AND TOTAL CORRELATIONS

1. Partial Correlation

When more than two variables are studied excluding some other variables, the relationship is termed as partial correlation.

2. Total Correlation

When more than two variables are studied without excluding any variables, the relationship is termed total correlation.

Linear and Nonlinear Correlations

1 . Linear Correlation:

If the ratio of change between two variables is constant, the correlation is said to be linear. If such variables are plotted on a graph paper, a straight line is obtained.

X	5	10	15	20	25	30
Y	2	4	6	8	10	12

2. Nonlinear Correlation:

If the ratio of change between two variables is not constant, the correlation is said to nonlinear. The graph of a nonlinear or curvilinear relationship will be a curve.

X	15	22	25	30	35	40
Y	4	5	8	9	10	12

Method of correlation:

There are two different methods.

1. Graphic methods.
2. Mathematical methods.

Graphic methods:

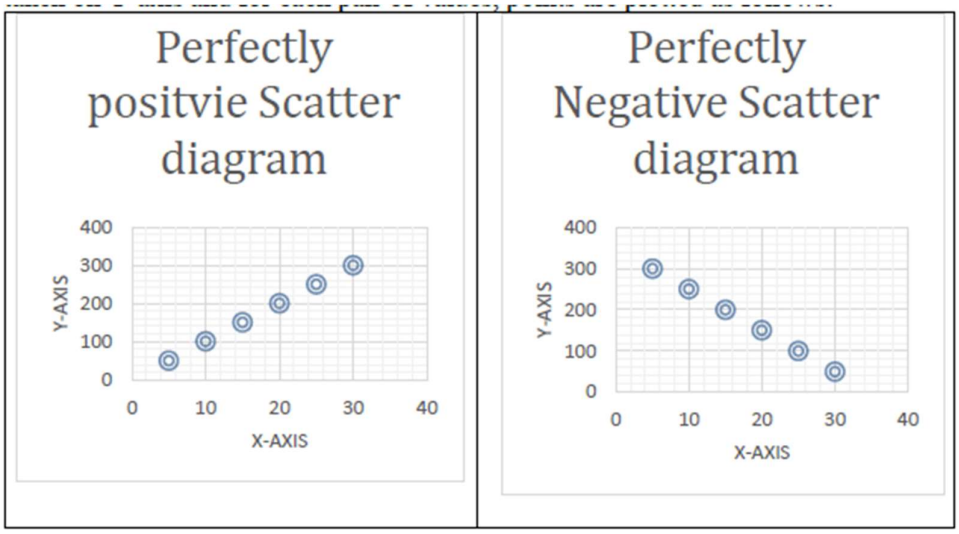
1. Scatter diagram.
2. Simple graph.

Mathematical methods:

1. Karl Pearson's coefficient of correlation.
2. Spearman's rank coefficient of correlation.

Scatter diagram:

This is a very simple method studying the relationship between two variables. In this method one variable is taken on X-axis and the other variable is taken on Y-axis and for each pair of values, points are plotted as follows:



Example 1:

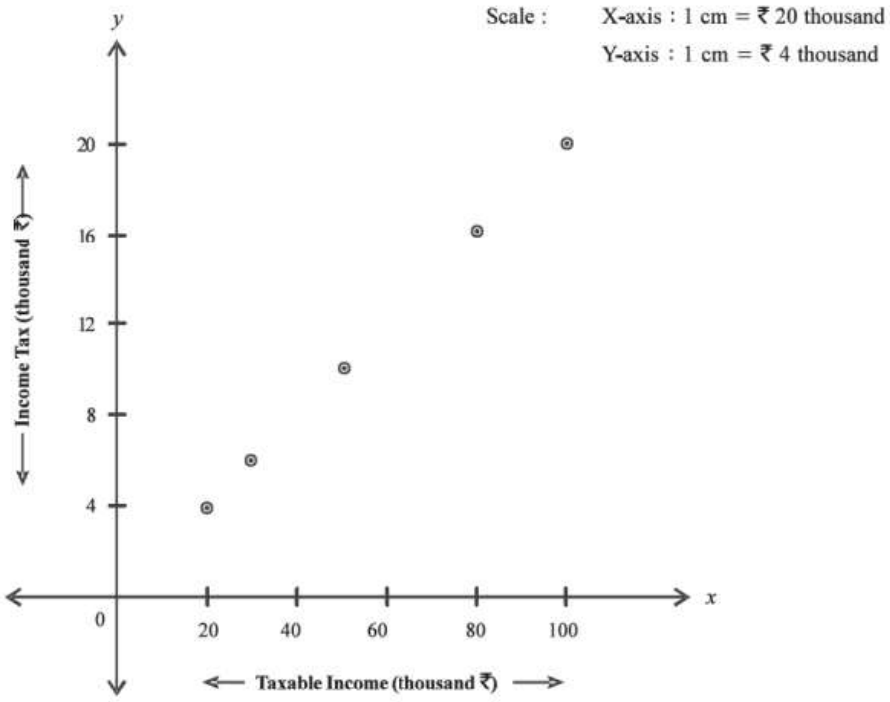
After standard deduction from total income, 20% income tax is imposed on the remaining income. The information regarding the taxable income and the tax to be paid is given below for five persons.

Person	1	2	3	4	5
Taxable Income (thousand ₹) x	50	30	80	20	100
Income Tax (thousand ₹) y	10	6	16	4	20

Draw a scatter diagram from this information and discuss about the correlation.

Solution:

The following scatter diagram is obtained by plotting the points corresponding to the ordered pairs (50,10), (30,6), (80,16), (20,4) and (100,20) of x and y.



We can see that all the points lie on the same line in the scatter diagram. We can also see that as the values of variable X change, the values of variable Y also change in the same direction with a constant proportion. Hence, we can see that there is a perfect positive correlation between two variables X and Y.

Example: 2

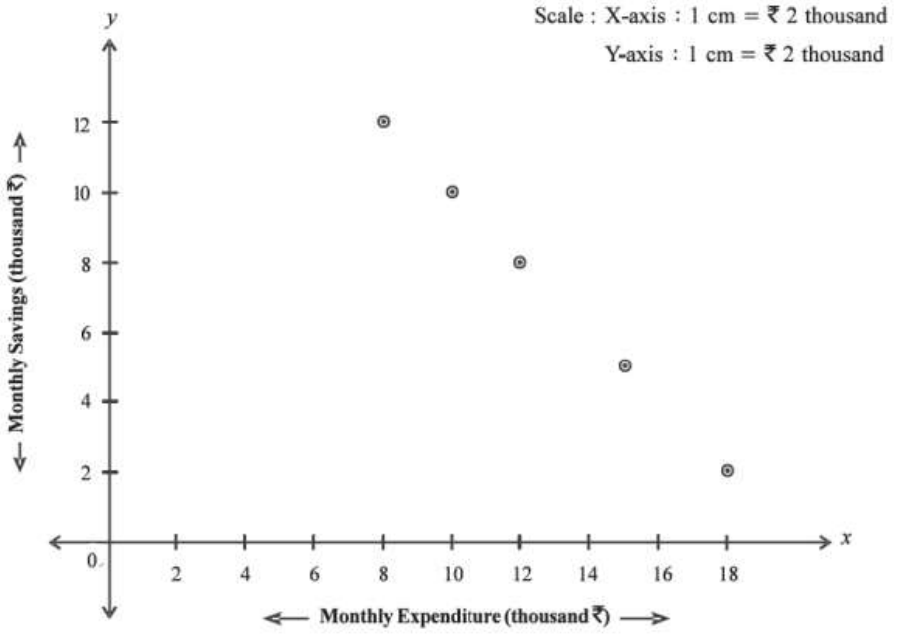
To know the relation between monthly expenditure and monthly savings for middle class families, the information regarding expenditure and savings for 5 families is given below. (The monthly income of each family is ₹ 20,000)

Monthly Expenditure (thousand ₹) x	15	18	8	10	12
Monthly Savings (thousand ₹) y	5	2	12	10	8

Draw a scatter diagram indicating the relation between monthly expenditure and monthly savings from this information and discuss about their correlation.

Solution:

The following scatter diagram is obtained by plotting the points of ordered pairs (15,5), (18,2), (8,12), (10,10), (12,8) of X and Y on the graph paper.



We can see that all the points lie on the same line in the scatter diagram. We can also see that as the values of variable X change, the values of variable Y also change in the opposite direction with a constant proportion. Hence, we can see that there is a perfect negative correlation between X and Y.

Exercise

1. A ball pen making company wants to know the relation between the price (in ₹) and supply (in thousand units) of its most selling Gel Pen. The following information is collected for it: Draw a scatter diagram and interpret it.

Price (in ₹)	14	16	12	11	15	13	17
Monthly Supply	32	50	20	12	45	30	53

2. The following information is collected to study the relationship between the minimum day temperature and sale of woollen cloths during a particular day of winter for six different cities.

Minimum day temperature (Celsius)	12	20	8	5	15	24
Sale of woollen cloths (thousand units)	35	10	45	70	20	8

Draw a scatter diagram from this information and interpret it.

Karl Pearson's Coefficient of Correlation

The coefficient of correlation is the measure of correlation between two random variables X and Y, and is denoted by r.

$$r = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where $\text{cov}(X,Y)$ is covariance of variables X and Y,

σ_X is the standard deviation of variable X,

and σ_Y is the standard deviation of variable Y.

This expression is known as Karl Pearson's coefficient of correlation or Karl Pearson's product-moment coefficient of correlation.

$$cov(X,Y) = \frac{1}{n}\sum(x-\bar{x})(y-\bar{y})$$

$$\sigma_X = \sqrt{\frac{\sum(x-\bar{x})^2}{n}}, \qquad \sigma_Y = \sqrt{\frac{\sum(y-\bar{y})^2}{n}}$$

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

The above expression can be further modified.

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

Properties of Coefficient of Correlation

1. The coefficient of correlation lies between -1 and 1, i.e., $-1 \leq r \leq 1$.
2. Correlation coefficient is independent of change of origin and change of scale.
3. Two independent variables are uncorrelated.

Example : 1

Calculate the correlation coefficient between x and y using the following data:

x	2	4	5	6	8	11
y	18	12	10	8	7	5

Solution:

$n = 6$

x	y	x^2	y^2	xy
2	18	4	324	36
4	12	16	144	48
5	10	25	100	50
6	8	36	64	48
8	7	64	49	56
11	5	121	25	55
$\Sigma x = 36$	$\Sigma y = 60$	$\Sigma x^2 = 266$	Σy^2=706	Σxy=293

$$r = \frac{\sum xy - \frac{\Sigma x \Sigma y}{n}}{\sqrt{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}\sqrt{\Sigma y^2 - \frac{(\Sigma y)^2}{n}}} = \frac{293 - \frac{(36)(60)}{6}}{\sqrt{266 - \frac{(36)^2}{6}}\sqrt{706 - \frac{(60)^2}{6}}} = \mathbf{-0.9203}$$

Example : 2

Calculate the correlation coefficient between the following data:

x	5	9	13	17	21
y	12	20	25	33	35

(Summer 2023)

Solution:

$$n = 5$$

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{65}{5} = 13, \qquad \bar{y} = \frac{\Sigma y_i}{n} = \frac{125}{5} = 25$$

x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
5	12	–8	–13	64	169	104
9	20	–4	–5	16	25	20
13	25	0	0	0	0	0
17	33	4	8	16	64	32
21	35	8	10	64	100	80
$\sum x$ =65	$\sum y$ =125	$\sum(x-\bar{x}) = 0$	$\sum(y-\bar{y}) = 0$	$\sum(x-\bar{x})^2 = 160$	$\sum(y-\bar{y})^2 = 358$	$\sum(x-\bar{x})(y-\bar{y}) = 236$

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}} = \frac{236}{\sqrt{160}\sqrt{358}} = \mathbf{0.986}$$

Example : 3

Calculate the correlation coefficient between for the following values of demand and the corresponding price of a commodity:

Demand in quintals	65	66	67	67	68	69	70	72
Price in rupees per kg	67	68	65	68	72	72	69	71

Solution

Let the demand in quintal be denoted by x and the price in rupees per kg be denoted by y.

$$n = 8$$

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{544}{8} = 68$$

$$\bar{y} = \frac{\Sigma y_i}{n} = \frac{552}{8} = 69$$

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

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}} = \frac{24}{\sqrt{36}\sqrt{44}} = \mathbf{0.603}$$

Example : 4

Given n = 10, σ_x = 5.4, σ_y = 6.2, and sum of the product of deviations from the mean of x and y is 66. Find the correlation coefficient.

Solution

n = 10, σ_x = 5.4, σ_y = 6.2

$$\sum(x-\bar{x})(y-\bar{y}) = 66$$

$$cov(X,Y) = \frac{1}{n}\sum(x-\bar{x})(y-\bar{y}) = \frac{66}{10} = 6.6$$

$$r = \frac{\text{cov}(X,Y)}{\sigma_X\,\sigma_Y} = \frac{6.6}{5.4 \times 6.2} = \mathbf{0.197}$$

Exercise:

1. Find the Pearson's Correlation Coefficient of the following data:

x	100	101	102	102	100	99	97	98	96	95
y	98	99	99	97	95	92	95	94	90	91

2. Calculate Karl Pearson's coefficient of correlation for the data given below:

x	10	14	18	22	26	30	10
y	18	12	24	6	30	36	18

3. Find the Pearson's Correlation Coefficient of the following data:

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

(Winter 2022-23)

4. Find Coefficient of Correlation between the following data:

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

(Winter 2023-24)

5. Calculate Karl Pearson's coefficient of correlation for the data given below:

x	17	19	21	26	20
y	23	27	25	26	27

6. Given n = 10, σ_x = 10.8, σ_y = 12.4, and sum of the product of deviations from the mean of x and y is 132. Find the correlation coefficient.

Spearman's Rank correlation coefficient:

Spearman's rank correlation coefficient, often denoted by the symbol ***ρ* (rho),** is a non-parametric measure of statistical dependence between two variables.

Here's a brief explanation of how Spearman's rank correlation coefficient is calculated.

Rank the data: For each variable, rank the data from lowest to highest, assigning a rank to each value. If there are ties, assign each tied value the average of the ranks it would have received if there were no tie.

Calculate the differences between ranks: For each pair of data points, find the difference between their ranks.

Spearman's Rank correlation coefficient:

Calculated by following formula: $\boldsymbol{r = 1 - \frac{6\sum d^2}{n(n^2-1)}}$

Where n = number of pair

In case finding out **rank correlation coefficient** when the observations are paired, the above formula can be written as:

$$r = 1 - \frac{6\left\{\sum d^2 + \frac{m}{12}(m^2-1) + \frac{m}{12}(m^2-1) + \ldots\ldots\ldots\ldots\ldots\ldots\right\}}{n(n^2-1)}$$

In $\sum d^2$, $\frac{m}{12(m^2-1)}$ is added where m is the number of times an item is repeated.

The value of $\boldsymbol{\rho}$ lies between -1 and 1. A positive value indicates a positive monotonic relationship, while a negative value indicates a negative monotonic relationship. A value of 0 indicates no monotonic relationship.

Example: 1

Two judges have given ranks to 10 students for their honesty. Find the rank correlation coefficient of the following data:

1ST Judge	3	5	8	4	7	10	2	1	6	9
2nd Judge	6	4	9	8	1	2	3	10	5	7

Solution

Rank given by 1st judge	Rank given by 2nd judge	Difference in ranks d	d^2
3	6	-3	9
5	4	1	1
8	9	-1	1
4	8	-4	16
7	1	6	36
10	2	8	64
2	3	-1	1
1	10	-9	81
6	5	1	1
9	7	2	4
			$\sum d^2$=214

$$r = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6(214)}{10(100-1)} = 1 - \frac{1284}{990} = 1 - 1.30 = \boldsymbol{-0.3}$$

Example: 2

Ten students got the following percentage of marks in mathematics and physics.

(x)maths	8	36	98	25	75	82	92	62	65	35
(y)physics	84	51	91	60	68	62	86	58	35	49

Find the rank correlation coefficient.

Solution

$n = 10$

x	y	Rank in maths (x)	Rank in physics (y)	$d = x - y$	d^2
8	84	10	3	7	49
36	51	7	8	–1	1
98	91	1	1	0	0
25	60	9	6	3	9
75	68	4	4	0	0
82	62	3	5	–2	4
92	86	2	2	0	0
62	58	6	7	–1	1
65	35	5	10	–5	25
35	49	8	9	–1	1
				$\sum d = 0$	$\sum d^2 = 90$

$$r = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6(90)}{10(100-1)} = \mathbf{0.455}$$

Example: 3

Find the Coefficient of rank correlation of the following data: **(Summer 2022-23)**

x	35	40	42	43	40	53	54	49	41	55
y	102	101	97	98	38	101	97	92	95	95

Solution

$n = 10$

x	y	Rank in (x)	Rank in (y)	$d = x - y$	d^2
35	102	10	1	9	81
40	101	8.5	2.5	6	36
42	97	6	5.5	0.5	0.25
43	98	5	4	1	1
40	38	8.5	10	-1.5	2.25
53	101	3	2.5	0.5	0.25
54	97	2	5.5	-3.5	12.25
49	92	4	9	-5	25
41	95	7	7.5	-0.5	0.25
55	95	1	7.5	-6.5	42.25
					$\sum d^2 = 200.50$

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

$$= 1 - \frac{6\{200.50 + 0.5 + 0.5 + 0.5 + 0.5\}}{990}$$

$$= \mathbf{-0.227}$$

Exercise:

1.Compute Spearman's rank correlation coefficient from the following data:

x	18	20	34	52	12
y	39	23	35	52	12

2.Obtain the rank correlation coefficient from the following data.

x	10	12	18	18	15	40
y	12	18	25	25	50	25

(Summer 2023-24)

REGRESSION:

By studying the correlation, we can know the existence, degree and direction of relationship between two variables but we cannot answer the question of the type if there is a certain amount of change in one variable, what will be the corresponding change in the other variable. The above type of question can be answered if we can establish a quantitative relationship between two related variables. The statistical tool by which it is possible to predict or estimate the unknown values of one variable from known values of another variable is called regression. A line of regression is a straight line.

LINES OF REGRESSION

If the variables, which are highly correlated, are plotted on a graph then the points lie in a narrow strip. If all the points in the scatter diagram cluster around a straight line, the line is called the line of regression. The line of regression is the line of best fit and is obtained by the principle of least squares.

Line of Regression of y on x:

It is the line which gives the best estimate for the values of y for any given values of x. The regression equation of y on x is given by

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

It is also written as

$$y = a + bx$$

Line of regression of x on y:

It is the line which gives the best estimate for the values of x for any given values of y. The regression equation for x on y is given by

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

It is also written as

$$x = a + by$$

where $\bar{x}$ and $\bar{y}$ are means of x series and y series respectively, σ_x and σ_y are standard deviations of x series and y series respectively, r is the correlation coefficient between x and y.

REGRESSION COEFFICIENTS

The slope b of the line of regression of y on x is also called the coefficient of regression of y on x. It represents the increment in the value of y corresponding to a unit change in the value of x.

b_{yx} = Regression coefficient of y on x $= r\frac{\sigma_y}{\sigma_x}$

Similarly, the slope b of the line of regression of x on y is called the coefficient of regression of x on y. It represents the increment in the value of x corresponding to a unit change in the value of y.

$$b_{xy} = \text{Regression coefficient of } x \text{ on } y = r\frac{\sigma_x}{\sigma_y}$$

Expressions for Regression Coefficients:

(a) $$b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$$

and

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(y-\bar{y})^2}$$

(b) $$b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

and

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum y^2 - \frac{(\sum y)^2}{n}}$$

Properties of Regression Coefficient:

(1) The coefficient of correlation is the geometric mean of the coefficients of regression, i.e., $r = \sqrt{b_{yx}b_{xy}}$.
(2) If one of the regression coefficients is greater than one, the other must be less than one.
(3) The arithmetic mean of regression coefficients is greater than or equal to the coefficient of correlation.
(4) Regression Coefficients are independent of the change of origin but not of scale.

Example: 1

The following data regarding the heights (y) and weights (x) of 100 college students are given:

$$\Sigma x = 15000, \quad \Sigma x^2 = 2272500, \quad \Sigma xy = 1022250$$
$$\Sigma y = 6800, \quad \Sigma y^2 = 463025.$$

Find the coefficient of correlation between height and weight and also the equation of regression of height and weight.

Solution:

$$n = 100$$

$$b_{yx} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

$$= \frac{1022250 - \frac{(15000)(6800)}{100}}{2272500 - \frac{(15000)^2}{100}}$$

$$= 0.1$$

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

$$= \frac{1022250 - \frac{(15000)(6800)}{100}}{463025 - \frac{(6800)^2}{100}}$$

$$= 3.6$$

$$r = \sqrt{b_{xy}\, b_{yx}} = \sqrt{(0.1)(3.6)} = 0.6$$

$$\bar{x} = \frac{\sum x}{n} = \frac{15000}{100} = 150$$

$$\bar{y} = \frac{\sum y}{n} = \frac{6800}{100} = 68$$

The equation of the line of regression of y on x is;

$$(y - \bar{y}) = b_{yx}(x - \bar{x})$$
$$(y - 68) = 0.1(x - 150)$$
$$\boldsymbol{y = 0.1x + 53}$$

The equation of the line of regression of x on y is;

$$(x - \bar{x}) = b_{xy}(y - \bar{y})$$
$$x - 150 = 3.6(y - 68)$$
$$\boldsymbol{x = 3.6y - 94.8}$$

Example: 2

Find the regression coefficients b_{yx} and b_{xy} and hence, find the correlationcoefficient between x and y for the following data:

x	4	2	3	4	2
y	2	3	2	4	4

Solution

$n = 5$

x	y	x^2	y^2	xy
4	2	16	4	8
2	3	4	9	6
3	2	9	4	6
4	4	16	16	16
2	4	4	16	8
$\sum x = 15$	$\sum y = 15$	$\sum x^2 = 49$	$\sum y^2 = 49$	$\sum xy = 44$

$$b_{yx} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{44 - \frac{(15)(15)}{5}}{49 - \frac{(15)^2}{5}} = -0.25$$

$$b_{xy} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum y^2 - \frac{(\sum y)^2}{n}} = \frac{44 - \frac{(15)(15)}{5}}{49 - \frac{(15)^2}{5}} = -0.25$$

$$r = \sqrt{b_{xy}\, b_{yx}} = \sqrt{(-0.25)(-0.25)} = \mathbf{0.25}$$

Exercise

1.Find the regression coefficient of y on x for the following data:

$\boldsymbol{x}$	1	2	3	4	5
$\boldsymbol{y}$	160	180	140	180	200

2. Find the equation of regression lines from the following data and also estimate y for $x = 1$ and x for $y = 4$.

$\boldsymbol{x}$	3	2	-1	6	4	-2	5	7
$\boldsymbol{y}$	5	13	12	-1	2	20	0	-3

3.Find the equation of regression lines and the correlation coefficient from the following data:

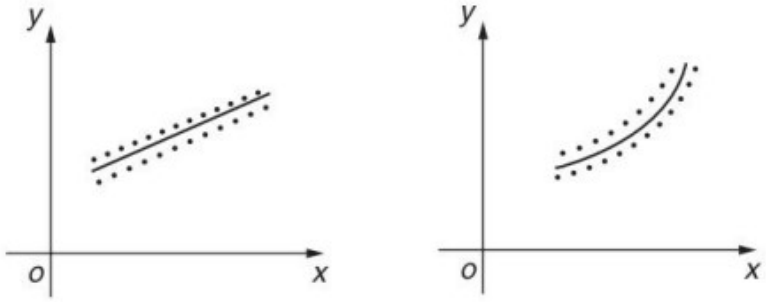
x	28	41	40	38	35	33	46	32	36	33
y	30	34	31	34	30	26	28	31	26	31

4.The following information is obtained for two variables x and y. Find regression equation of y on x. n=10;$\sum x = 130; \sum x^2 = 2288; \sum xy = 3467$.

CURVE FITTING

Curve fitting is the process of finding the 'best-fit' curve for a given set of data. It is the representation of the relationship between two variables by means of an algebraic equation. On the basis of this mathematical equation, predictions can be made in many statistical problems.

Suppose a set of n points of values $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ of the two variables x and y are given. These values are plotted on a rectangular coordinate system, i.e., the xy-plane. The resulting set of points is known as a scatter diagram (Fig. 5.1). The scatter diagram exhibits the trend and it is possible to visualize a smooth curve approximating the data. Such a curve is known as an approximating curve.



METHODS:

1.Linear Regression: One of the simplest forms of curve fitting, linear regression assumes a linear relationship between the independent and dependent variables. The goal is to find the best-fitting line (or hyperplane in higher dimensions) that minimizes the sum of squared differences between observed and predicted values.($Y = AX + B \; or \; X = AY + B$).

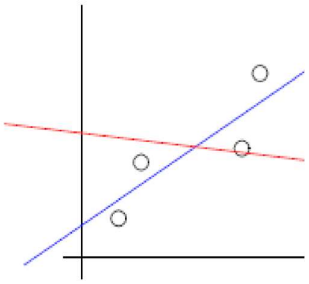
2.Polynomial Regression: Polynomial regression extends linear regression by allowing the model to include higher-degree polynomials. This flexibility enables a better fit for nonlinear relationships in the data.($Y = AX^2 + BX + C) or (\; X = AY^2 + BY + C$).

3.Exponential and Logarithmic: Exponential and logarithmic curve fitting is suitable for datasets exhibiting exponential growth or decay. These models are often used in fields like biology, physics, and finance.($Y = e^{ax}$)

Linear Regression:

Given the general form of a straight line

$$f(x) = ax + b$$

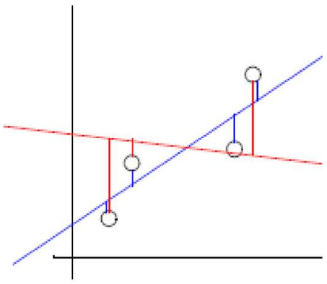


How can we pick the coefficients that best fits the line to the data?

First question: What makes a particular straight line a 'good' fit?

Why does the blue line appear to us to fit the trend better?

• Consider the distance between the data and points on the line



• Add up the length of all the red and blue verticle lines

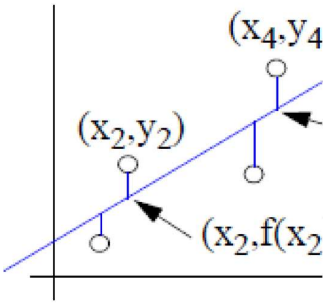
• This is an expression of the 'error' between data and fitted line

• The one line that provides a minimum error is then the 'best' straight line

Quantifying errors in a curve fit

(1) positive or negative error have the same value (data point is above or below the line)

(2) Weight greater errors more heavily



we can do both of these things by squaring the distance denote data values as (x, y) ======>>

denote points on the fitted line as (x, f(x))

sum the error at the four data points

$$err = \sum_{i=1}^{n} d_i^{\,2} = (y_1 - f(x_1))^2 + (y_2 - f(x_2))^2 + \ldots\ldots(y_n - f(x_n))^2$$

$$= (y_1 - (ax_1 + b))^2 + (y_2 - (ax_2 + b))^2 + \ldots\ldots + (y_n - (ax_n + b))^2$$

$$= \sum_{i=1}^{n} (y_i - (ax_i + b))^2$$

Error is minimum if first ordered partial derivatives=0

$$\frac{\partial(err)}{\partial a} = \sum_{i=1}^{n} -2x_i\left(y_i - (ax_i + b)\right) = 0 \qquad \frac{\partial(err)}{\partial b} = \sum_{i=1}^{n} -2\left(y_i - (ax_i + b)\right) = 0$$

$$\therefore \sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i^{\,2} - b\sum_{i=1}^{n} x_i = 0 \qquad \therefore \sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} 1 = 0$$

$$\therefore \sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i^{\,2} + b\sum_{i=1}^{n} x_i \qquad \text{and} \qquad \therefore \sum_{i=1}^{n} y_i = a\sum_{i=1}^{n} x_i + n\,b$$

Solve the equations

$$\sum_{i=1}^{n} y_i = a\sum_{i=1}^{n} x_i + nb \qquad (1)$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i \qquad (2)$$

Example: 1 Fit a straight line to the following data:

x	1	2	3	4	6	8
y	2.4	3	3.6	4	5	6

Solution

Let the straight line to be fitted to the data be

$$y = a + bx$$

$$\sum y = na + b\sum x \qquad (1)$$

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

$$n = 6$$

x	y	x^2	xy
1	2.4	1	2.4
2	3	4	6
3	3.6	9	10.8
4	4	16	16
6	5	36	30
8	6	64	48
$\sum x = 24$	$\sum y = 24$	$\sum x^2 = 130$	$\sum xy = 113.2$

Substituting these values inn Eqs (1) and (2)

$$24 = 6a + 24b \qquad (3)$$

$$113.2 = 24a + 130b \qquad (4)$$

Solving Eqs (3) and (4), we get

$$a = 1.9764$$

$$b = 0.5059$$

Hence, the required equation of straight line is $\boldsymbol{y = 1.9764 + 0.5059x}$

Example: 2 Fit a straight line to the following data. Also, estimate the value of y at $x = 2.5$.

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

(Winter 2022-23)

Example: 3 Fit a straight line using least square method.

x	0	0.5	1	1.5	2	2.5
y	0	1.5	3	4.5	6	7.5

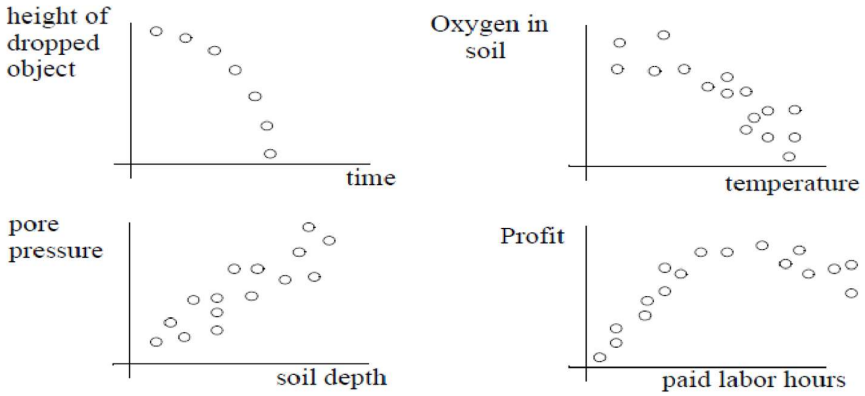
(Winter 2023-24)

Example: 4 Fit a straight line to the following data and hence find y when $x = 70$

x	71	68	73	69	67	65	66	67
y	69	72	70	70	68	67	68	64

(Summer 2023-24)

Polynomial Regression: We started the linear curve fit by choosing a generic form of the straight line $f(x) = ax + b$
This is just one kind of function. There are an infinite number of generic forms we could choose from for almost any shape we want. Let's start with a simple extension to the linear regression concept recall the examples of sampled data.



Error - Least squares approach

$$err = \sum_{i=1}^{n} d_i^{\,2} = (y_1 - f(x_1))^2 + (y_2 - f(x_2))^2 + \ldots\ldots(y_n - f(x_n))^2$$
$$= \left(y_1 - \left(a + bx_1 + cx_1^{\,2}\right)\right)^2 + \left(y_2 - \left(a + bx_2 + +cx_2^{\,2}\right)\right)^2 + \ldots\ldots + \left(y_n - \left(a + bx_n + cx_n\right)\right)^2$$
$$= \sum_{i=1}^{n} \left(y_i - \left(a + bx_i + cx_i^{\,2}\right)\right)^2$$

To minimize the error, derivatives with respect to $a, b \text{ and } c$ equal to 0.

$$\frac{\partial(err)}{\partial a} = \sum_{i=1}^{n} -2\left(y_i - \left(a + bx_i + cx_i^{\,2}\right)\right) = 0$$
$$\frac{\partial(err)}{\partial b} = \sum_{i=1}^{n} -2x_i\left(y_i - \left(a + bx_i + cx_i^{\,2}\right)\right) = 0$$
$$\frac{\partial(err)}{\partial b} = \sum_{i=1}^{n} -2x_i^{\,2}\left(y_i - \left(a + bx_i + cx_i^{\,2}\right)\right) = 0$$

Simplify these equations, we get

$$\sum_{i=1}^{n} y_i = a\,n + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^{\,2}$$
$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^{\,2} + c\sum_{i=1}^{n} x_i^{\,3}$$
$$\sum_{i=1}^{n} x_i^{\,2} y_i = a\sum_{i=1}^{n} x_i^{\,2} + b\sum_{i=1}^{n} x_i^{\,3} + c\sum_{i=1}^{n} x_i^{\,4}$$

Example: 1

Fit a least squares quadratic curve to the following data:

x	1	2	3	4
y	1.7	1.8	2.3	3.2

Estimate $y(2.4)$.

Solution:

Let the equation of the least squares quadratic curve (parabola) be $y = a + bx + cx^2$.

The normal equations are

$$\sum y = na + b\sum x + c\sum x^2 \qquad (1)$$

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

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

Here, $n = 4$

x	y	x^2	x^3	x^4	xy	x^2y
1	1.7	1	1	1	1.7	1.7
2	1.8	4	8	16	3.6	7.2
3	2.3	9	27	81	6.9	20.7
4	3.2	16	64	256	12.8	51.2
$\sum x = 10$	$\sum y = 9$	$\sum x^2 = 30$	$\sum x^3 = 100$	$\sum x^4 = 354$	$\sum xy = 25$	$\sum x^2y = 80.8$

Substitute these values in equations (1), (2) and (3),

$$9 = 4a + 10b + 30c$$
$$25 = 10a + 30b + 100c$$
$$80.8 = 30a + 100b + 354c$$

Solving the above equations, we get

$$a = 2, b = -0.5, c = 0.2$$

Hence, the required equation of quadratic curve is

$$y = 2 - 0.5x + 0.2x^2$$

$$y(2.4) = 2 - (0.5)(2.4) + (0.2)(2.4)^2 = 1.952$$

Example: 2

Fit a second-degree polynomial using least square method to the following data:

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

Example: 3

Fit a second order polynomial $y = a + bx + cx^2$ to following data, using least square method. **(Summer 2022-23)**

x	0	5	10	15	20
y	7	11	16	20	26

Curve fitting - Other nonlinear fits (exponential)
Q: Will a polynomial of any order necessarily fit any set of data?
A: Nope, lots of phenomena don't follow a polynomial form. They may be, for example, exponential

(1) General exponential equation $f(x) = C\,e^{Ax}$
Now, take log on both side, we get

$$\ln y = \ln C + Ax$$

$$Y = b + aX; \qquad \text{where } Y = \ln y,\ X = x, \ln C = b \text{ and } a = \ln A$$

Which is equation of line, the original data in xy- plane mapped into XY-plane. This is called *linearization.* The data (x, y) transformed as $(x, \ln y)$.

To find the value of a and b we will use the equations

$$\sum_{i=1}^{n} Y_i = a\sum_{i=1}^{n} X_i + n\,b \qquad (1)$$

$$\sum_{i=1}^{n} X_i Y_i = a\sum_{i=1}^{n} X_i^{\,2} + b\sum_{i=1}^{n} X_i \qquad (2)$$

After getting values of a and b, $A = \text{antilog}\, a,\ C = \text{antilog}\, b$.

Example: An experiment gave the following values:

X	1	5	7	9	12
Y	10	15	12	15	21

Fit an exponential curve $y = Ce^{Ax}$
Solution:

$X_I = x_i$	y_i	$Y_i = \ln y_i$	$X_i^{\,2}$	$X_i Y_i$
1	10	2.302585	1	2.302585
5	15	2.70805	25	13.54025
7	12	2.484906	49	17.39435
9	15	2.70805	81	24.37245
12	21	3.044522	144	36.53427
$\sum_{i=1}^{5} X_i$ =34		$\sum_{i=1}^{5} Y_i$ =13.24811	$\sum_{i=1}^{5} X_i^{\,2}$ =300	$\sum_{i=1}^{5} X_i Y_I$ =94.1439

$$13.24811 = 34A + 5B$$
$$94.1439 = 300A + 34B$$

A=2.00479, B=2.248664
a=antilog2.00479=7.424536, b=antilog (2.248664) =9.475068

Hence, best fit curve is $y = 9.475068e^{2.248664x}$

(2) $y = bx^a$

Taking $\log_{10}$ on both the side

$$\log_{10} y = \log_{10} b + a \log_{10} x$$
$$Y = B + AX; \qquad \text{where } Y = \log_{10} y,\ X = \log_{10} x \text{ and } a = A, B = \log_{10} b$$

$$\sum_{i=1}^{n} Y_i = nB + A\sum_{i=1}^{n} X_i \qquad (1)$$

$$\sum_{i=1}^{n} X_i Y_i = B\sum_{i=1}^{n} X_i + A\sum_{i=1}^{n} X_i^2 \qquad (2)$$

Example: An experiment gave the following values:

v (ft/min)	350	400	500	600
t (min)	61	26	7	2.6

It is known that v and t are connected by the relation $v = bt^a$, find the best possible values of a and b.

V	t	$Y=logv$	$X=logt$	X^2	XY
350	61	2.544068	1.78533	3.18740262	4.542001
400	26	2.60206	1.414973	2.002149575	3.681846
500	7	2.69897	0.845098	0.714190697	2.280894
600	2.6	2.778151	0.414973	0.17220288	1.152859
		$\sum_{i=1}^{4} Y_i$ =10.62325	$\sum_{i=1}^{4} X_i$ =4.460375	$\sum_{i=1}^{4} X_i^2$ =6.075945772	$\sum_{i=1}^{4} X_i^3$ =11.6576

Substitute in given equation,

$$\sum_{i=1}^{n} Y_i = nB + A\sum_{i=1}^{n} X_i \qquad (1)$$

$$\sum_{i=1}^{n} X_i Y_i = B\sum_{i=1}^{n} X_i + A\sum_{i=1}^{n} X_i^2 \qquad (2)$$

$$10.62325 = 4B + 4.460375A$$
$$11.6575 = 4.460375B + 6.075945772A$$

On solving these equations B=2.845 A=a= - 0.17.
$b = anti\log(2.845) = 699.842$

3)The following values of T and l follow the law $T= al^n$. Test if this is so and find the best values of a and n.

T	1.0	1.5	2.0	2.5
L	25	56.2	100	1.56