Introduction to Statistical Methods

Team ISM

BITS
Pilani

innovate achieve lead

BITS Pilani

Webinar-4

Correlation Analysis

(Correlation, Regression and Maximum likelihood estimation)

(24th August 2023)

Contact session	List of topic	References
Webinar-4	**Correlation, Regression Analysis and Maximum likelihood estimation.**	

CORRELATION Analysis

1.1 Definition of Correlation

1.2Types of Correlation

1.3 Methods of studying Correlation

Definition of Correlation

- Correlation is a statistical tool which is used to know the relationship between two or more than two variables.
- In bivariate frequency distribution f(x,y) if a change in one variable x creates or effects any change in another variable y, then the two variables are set to be correlated.

Types of Correlation

- Positive or Negative Correlation
- Simple, Partial and Multiple correlation
- Linear and non linear correlation

Positive or Negative Correlation

- If a change in the variable X creates any change in the variable Y in the same direction then X and Y are said to be Positive correlated.
- If one variable increases the second variable will also increases correspondingly.

Examples: 1. Income and expenditure

2. Distance travelled by train and train fare

3.Age and sickness

4. Advertising and sales

If a change in one variable creates change in other variable, but in opposite direction then two variables are said to be negatively correlated. i.e if one variable increases, then the other decreases or vice versa.

Examples: 1.Price and demand

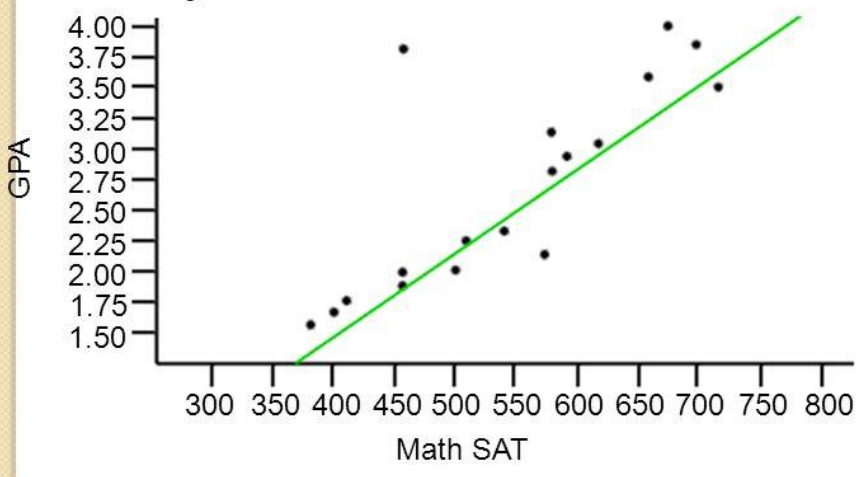
2. Demand for sweaters and the day temperature.

Cont…

Scatter Plots and Types of Correlation

x = SAT score

y = GPA

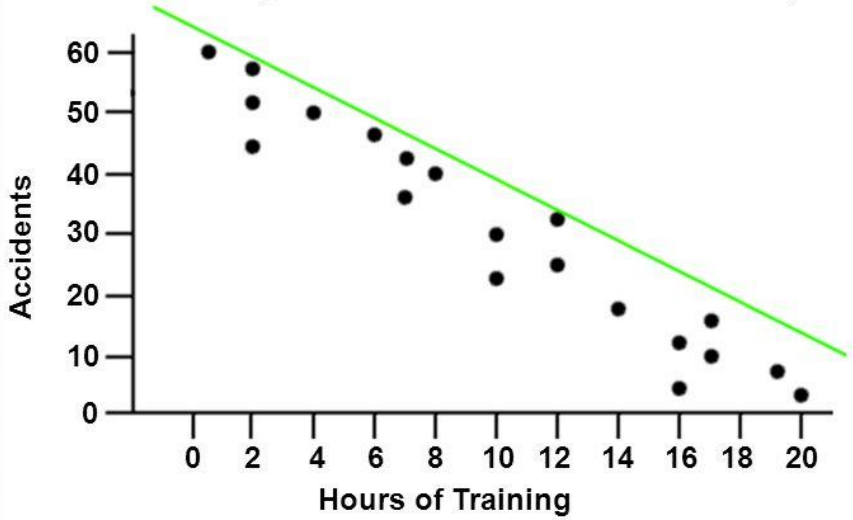


Positive Correlation–as x increases, y increases

Cont…

Scatter Plots and Types of Correlation

x = hours of training (horizontal axis)
y = number of accidents (vertical axis)



Negative Correlation–as *x* increases, *y* decreases

Simple, partial and multiple correlation

- The distinction between simple, partial and multiple correlation is based on the number of variables studied.
- When only two variables are studied, it is a simple correlation

  Example: Intelligent quotient and marks secured.

  - When 3 or more variables are studied then it is a multiple correlation.

  Example: yield of rice per acre and both the amount of rainfall and the amount of fertilizers used.

  - In partial correlation we recognize more than two variables but consider only two variables are influencing each other and ignore other variables by keeping them as constant.

  Example: Mileage of car and car type can only be studied by ignoring road conditions, driving skills etc..

Linear and Non-linear(curvi linear correlation)

- If the change in one variable causes the change in other variable in constant ratio then it is called linear correlation.
- If such variables are plotted on a graph paper, plotted points will fall on a straight line.

example: Amount of raw material required and number of finished goods.

When the amount of change in one variable is not in a constant ratio to the amount of change in other variable then it is called non-linear correlation.

Example: Amount of rainfall and yield of crop.

Methods of studying Correlation

There are various methods to know whether the two variables are correlated or not. They are

- Scatter diagram
- Karl Pearson's coefficient of correlation

Scatter Diagram

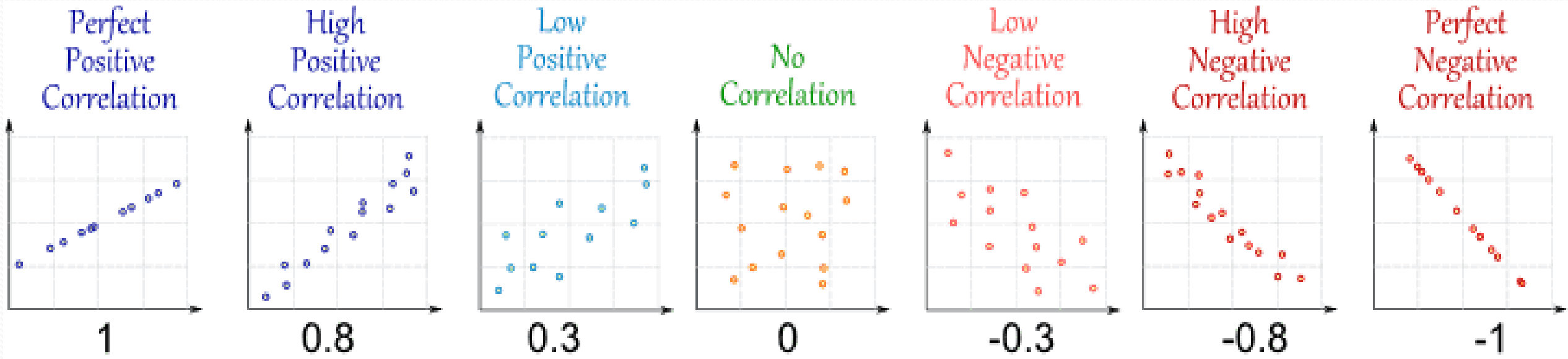
To know the correlation between any two variables , scatter diagrams are used.

Procedure: given n pair of values (x1,y1) (x2,y2)…(xn,yn) of two variables X and Y.

- Take the independent variable on x axis and dependent variable on y axis. Then plot the n points on the graph.
- The diagram of dots thus obtained is called scatter diagram.
  - If the points reveal any upward or downward trend the variables are said to be correlated , otherwise un-correlated.
  - If the points are very closed to each other, a good amount of correlation exists.
  - Upward trend indicates +ve correlation and downward trend indicates –ve correlation.
  - Limitation: By this method we cannot establish the exact degree of correlation between the variables.

Scattered diagram- Some specific cases

$$-1 \leq Correlation(x, y) \leq +1$$



Karl Pearson's coefficient of correlation

The extent relationship between the variables calculated with the help of statistical techniques known as correlation coefficient. It was developed by Karl Pearson. It is denoted by r(x,y) or r_{xy}

It always varies between +1 or -1

When r= +1 then it denotes perfect +ve correlation

r= 0 denotes no correlation

r= -1 perfect –ve correlation

Correlation coefficent between two random variables X and Y is as follows:

$$r_{xy} = \frac{\operatorname{cov}(x, y)}{\sigma_x \sigma_y}$$

$$= \frac{E(X,Y) - E(X)E(Y)}{\sqrt{E(X^2) - [E(X)]^2} . \sqrt{E(\mathrm{Y}^2) - [E(\mathrm{Y})]^2}}$$

$$= \frac{\frac{1}{N}\sum (X_i - \bar{X})(\mathrm{Y}_i - \bar{Y})}{\sqrt{\frac{1}{N}\sum (X_i - \bar{X})^2} \sqrt{\frac{1}{N}\sum (\mathrm{Y}_i - \bar{Y})^2}}$$

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

Properties of the coefficient of correlation

1. The coefficient of correlation lies between -1 and +1.
2. Correlation is independent of change of origin and scale.
3. The coefficient of correlation is the geometric mean of two regression coefficients

$$r_{xy} = \sqrt{b_{xy} b_{yx}}$$

1. The degree of relationship between two variables is symmetric

$$r_{xy} = r_{yx}$$

Calculate karl pearson's coefficient of correlation for the following data

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

Sol: Correlation coefficient

$$r_{xy} = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}}$$

X	Y	X^2	Y^2	XY
9	15	81	225	135
8	16	64	256	128
7	14	49	196	98
6	13	36	169	78
5	11	25	121	55
4	12	16	144	48
3	10	9	100	30
2	8	4	64	16
1	9	1	81	9
$\sum X = 45$	$\sum Y = 108$	$\sum X^2 = 285$	$\sum Y^2 = 1356$	$\sum XY = 597$

Therefore, N=9, $\sum XY = 597, \sum X = 45, \sum X^2 = 285, \sum Y = 108, \sum Y^2 = 1356$

$$r_{xy} = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}}$$

$$= \frac{9*597 - 45*108}{\sqrt{9*285 - (45)^2}\sqrt{9*1356 - (108)^2}}$$

$$= \frac{513}{\sqrt{540}\sqrt{540}} = \frac{513}{540}$$

$$= +0.95$$

There is a high degree of +ve correlation between variables X and Y.
i.e if x increases y also increases

Calculate karl pearson's coefficient of correlation for the following data

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

X	Y	$u = x - \bar{x}$	$v = y - \bar{y}$	u^2	v^2	uv
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
$\sum x = 544$	$\sum y = 552$	$\sum u = 0$	$\sum v = 0$	$\sum u^2 = 36$	$\sum v^2 = 44$	$\sum uv = 24$

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

$$r_{xy} = \frac{n\sum uv - (\sum u)(\sum v)}{\sqrt{n\sum u^2 - (\sum u)^2}\sqrt{n\sum v^2 - (\sum v)^2}}$$

$$= \frac{8(24) - 0}{\sqrt{8(36) - 0}\sqrt{8(44) - 0}}$$

$$= 0.603$$

sales	15	18	25	27	30	35
Advertising expenditure	50	65	82	95	110	120

Find the correlation coefficient for the above data

Regression Analysis

The term regression literally means stepping back towards the average

Definition: Regression analysis is a mathematical measure of the average relationship between 2 or more variables in terms of original units of the data.

- In regression analysis there are 2 variables.
- The variable whose values are influenced or can be predicted is called dependent variable and the variable which influences or used for prediction is called independent variable.
- In regression analysis the dependent variable is known as regressed or explained variable, while the independent variable is known as regressor or predictor or explanator.

Lines of Regression

- If we plot the points of a bivariate distribution in the scatter diagram the points will cluster round a curve called curve of regression.
- If the curve is straight line it is called line of regression or linear regression otherwise the regression is said to be non linear regression.
- Line of regression of y on x is the line which gives the best estimate for the value of y for any specified value of x.
- Line of regression of x on y is the line which gives the best estimate for the value of x for any specified value of y.

Types of Regression

- Simple Regression: the regression analysis confined to the study of only two variables at a time is called as simple regression
- Multiple Regression: The regression analysis for studying more than two variables at a time is termed as Multiple Regression.
- Utility of Regression Test: Regression lines are useful in prediction of values of one variable for a specified value of other variable.
- Examples:

  1.When price and demand are related, we can estimate the future demand for a specified price.

  2. When crop yield depend on amount of rain fall then regression test can predict crop yield for a particular amount of rain fall.

Cont…

- Let x and y are two variables for regression analysis then there will be two regression lines.
- Regression line of x on y = $x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$

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

- Regression line of y on x= $y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$

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

- The intersection of two lines x on y and y on x gives the mean values of x series and y series. i.e., and $\bar{x}$ $\bar{y}$

Properties of Regression coefficients

- Correlation coefficient is the geometric mean of the regression coefficients. $r = \sqrt{b_{xy}.b_{yx}}$

- If one regression coefficient is > unity then the other must be < unity.
- The arithmetic mean of the two regression coefficient is > the correlation coefficient.

Problems

- The folowing table gives the normal weight of kids during the first 8 years of life;

Age	0	1	2	3	5	6	7	8
Weight	5	7	8	10	15	17	20	22

i. Obtain two lines of regression
ii. Estimate the weight of kid at the age of 4 years

solution : we know that regression line y on x is $y-\bar{y}=b_{yx}(x-\bar{x})$

Where $\bar{x},\bar{y}$ are means

$b_{yx}=\dfrac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$ is regression coefficient.

Line of regression x on y is $x-\bar{x}=b_{xy}(y-\bar{y})$

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

N=no of observations.

$$\bar{x}=\frac{\sum x}{N}=\frac{32}{8}=4$$

$$\bar{y}=\frac{\sum y}{N}=\frac{104}{8}=13$$

Age(x)	Weight(y)	$(x-\bar{x})=x-4$	$(y-\bar{y})=y-13$	$(x-\bar{x})^2=(x-4)^2$	$(y-\bar{y})^2=(y-13)^2$	$(x-\bar{x})(y-\bar{y})=(x-4)(y-13)$
0	5	-4	-8	16	64	32
1	7	-3	-6	9	36	18
2	8	-2	-5	4	25	10
3	10	-1	-3	1	9	3
5	15	1	2	1	4	2
6	17	2	4	4	16	8
7	20	3	7	9	49	21
8	22	4	9	16	81	36
Total=32	104	0	0	60	284	130

From the table $\sum x = 32, \sum y = 104, \sum(x-\bar{x}) = 0, \sum(y-\bar{y}) = 0, \sum(x-\bar{x})^2 = 60$

$\sum(y-\bar{y})^2 = 284, \sum(x-\bar{x})(y-\bar{y}) = 130$

$$b_{yx} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{130}{60} = 2.167$$

$$b_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(y_i - \bar{y})^2} = \frac{130}{284} = 0.458$$

Line of regression y on x

$$y - \bar{y} = b_{yx}(x - \bar{x})$$
$$y - 13 = 2.167(x - 4)$$
$$y - 13 = 2.167x - 8.668$$
$$y = 2.167x + 5.668$$

Line of regression x on y

$$x - \bar{x} = b_{xy}(y - \bar{y})$$
$$x - 4 = 0.458(y - 13)$$
$$x - 4 = 0.458y - 5.954$$
$$x = 0.458y - 1.954$$

Weight of a kid at the age of 4 is estimating y when x=4
We use y on x regression line.

$$y = 2.167x + 5.668$$

$$y = 2.167(4) + 5.668$$

$$y = 14.336$$

The weight of a kid at the age of 4is 14.336.

Problem: The following are the regression equations.
Find the regression line x on y and y on x and also find

i. $\bar{x}$ And $\bar{y}$
ii. The regression coefficients
iii. Correlation Coefficients

$$8x - 10y + 66 = 0$$

$$40x - 18y = 214$$

Solution:

$$8x - 10y + 66 = 0 \quad \text{.....(1)}$$

$$40x - 18y = 214 \quad \text{.....(2)}$$

i. By solving two regression lines we get $\bar{x}$ And $\bar{y}$

$\bar{x}$ = 13 And $\bar{y}$ =17

ii. Rewrite the equations of regression to get the form x on y and y on x
From eq-(1)

$$10y = 8x + 66 \Rightarrow y = \frac{8}{10}x + \frac{66}{10}$$

$$y = a + bx \Rightarrow y = \frac{66}{10} + \frac{8}{10}x$$

$$b_{yx} = \frac{8}{10} = \frac{4}{5}$$

Equation (2) $40x - 18y = 214$

$$\therefore x = \frac{18}{40}y + \frac{214}{40}$$

$$x = a + by \Rightarrow x = \frac{214}{40} + \frac{18}{40}y$$

$$\therefore b_{xy} = \frac{18}{40} = \frac{9}{20}$$

Regression coefficient x on y $= \frac{9}{20} = b_{xy}$

Regression coefficient y on x $= \frac{4}{5} = b_{yx}$

iii)Correlation coefficient $r = \pm\sqrt{b_{xy}.b_{yx}} = \pm\sqrt{\frac{9}{20}.\frac{4}{5}} = \pm\sqrt{\frac{9}{25}} = \pm 0.6$

Since both regression coefficients are positive ,r must be positive there fore r=o.6

Least square Method

Algebraically method-:

1. Least Square Method-:

The regression equation of X on Y is :

$$X = a+bY$$

Where,

X=Dependent variable

Y=Independent variable

The regression equation of Y on X is:

$$Y = a+bX$$

Where,

Y=Dependent variable

X=Independent variable

And the values of a and b in the above equations are found by the method of least of Squares-reference . The values of a and b are found with the help of normal equations given below:

(I)

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

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

(II)

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

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

Cont…

Example1-: From the following data obtain the two regression equations using the method of Least Squares.

X	3	2	7	4	8
Y	6	1	8	5	9

Solution-:

X	Y	XY	X^2	Y^2
3	6	18	9	36
2	1	2	4	1
7	8	56	49	64
4	5	20	16	25
8	9	72	64	81
$\sum X = 24$	$\sum Y = 29$	$\sum XY = 168$	$\sum X^2 = 142$	$\sum Y^2 = 207$

Cont…

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

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

Substitution the values from the table we get

29=5a+24b……(i)
168=24a+142b
84=12a+71b……..(ii)

Multiplying equation (i) by 12 and (ii) by 5

348=60a+288b……(iii)
420=60a+355b……(iv)

By solving equation(iii)and (iv) we get

a=0.66 and b=1.07

Cont…

By putting the value of a and b in the **Regression equation Y on X** we get

Y=0.66+1.07X

Now to find the regression equation of X on Y ,
The two normal equation are

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

Substituting the values in the equations we get

24=5a+29b……(i)
168=29a+207b……..(ii)

Multiplying equation (i)by 29 and in (ii) by 5 we get

a=0.49 and b=0.74

Substituting the values of a and b in the **Regression equation X and Y**

X=0.49+0.74Y

- Fit a straight line to the following data

x	0	1	2	3	4
y	1	1.8	3.3	4.5	6.5

- Show that the line of fit of the following data is given by y=0.7x+11.28

x	0	5	10	15	20	25
y	12	15	17	22	24	30

Non- linear least square approximation:

Parabola: let the equation of parabola to be fit be $y = a + bx + cx^2 \ldots\ldots\ldots\ldots(1)$

The normal equations are

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

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

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

Solving these equations for a,b,c and substituting in (1) we get required parabola of best fit.

Problem

1)Fit the second degree parabola to the following data

x	0	1	2	3	4
y	1	5	10	22	38

sol: let the equation of the parabola be

$$y = a + bx + cx^2 \ldots\ldots\ldots\ldots (1)$$

The normal equations are

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

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

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

x	y	xy	x^2	x^2y	x^3	x^4
0	1	0	0	0	0	0
1	5	5	1	5	1	1
2	10	20	4	40	8	16
3	22	66	9	198	27	81
4	38	152	16	608	64	256
$\sum x = 10$	$\sum y = 76$	$\sum xy = 152$	$\sum x^2 = 30$	$\sum x^2 y = 851$	$\sum x^3 = 100$	$\sum x^4 = 354$

Sub these values in the normal equations $76 = 5a + 10b + 30c$

$$243 = 10a + 30b + 100c$$

$$851 = 30a + 100b + 354c$$

Solving above a=1.42, b=0.26, c=2.221

Sub a, b, c in (1)

$y = 1.42 + 0.26x + 2.21x^2$ Is the required parabola of the best fit.

Power curve: The power curve is $y = ax^b \ldots\ldots(1)$

Taking log on both sides

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

$$Y = A + bX \ldots\ldots\ldots\ldots(2)$$

$$where, Y = \log_{10} y, A = \log_{10} a, X = \log_{10} x$$

(2) Is a linear equation in X and Y
The normal equations are

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

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

Solving for A and b and substitute in (2).

Problem: Fit a curve $y = ax^b$ to the following data

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

Sol: $y = ax^b$ .....(1)

Taking log on both sides

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

$$Y = A + bX \text{.....(2)}$$

$$where, Y = \log_{10} y, A = \log_{10} a, X = \log_{10} x$$

This is a linear eq in X and Y
The normal eq s are

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

$$\sum XY = A \sum X + b \sum X^2 \qquad \text{.....(3)}$$

x	X=log x	y	Y=log y	XY	X^2
1	0	2.98	0.4742	0	0
2	0.3010	4.26	0.6294	0.1894	0.0906
3	0.4771	5.21	0.7168	0.3420	0.2276
4	0.6021	6.10	0.7853	0.4728	0.3625
5	0.6990	6.80	0.8325	0.5819	0.4886
6	0.7782	7.50	0.8751	0.6810	0.6056
TOTAL	$\sum X = 2.8574$		$\sum Y = 4.3133$	$\sum XY = 2.2671$	$\sum X^2 = 1.7749$

Substituting the above values in (3)

$4.3133 = 6A + 2.8574\ b$

$2.2671 = 2.8574\ A + 1.7749\ b$

Solving the equations for A and b

$A = 0.4739$ And $b = 0.5143$

$a = (10)^A = (10)^{0.4739} = 2.978$

Substituting a and b in (1)

$$y = 2.978x^{0.5143}$$

Exponential curve: $1) y = ae^{bx}, 2) y = ab^{x}$

1)The exponential curve $y = ae^{bx} \ldots\ldots\ldots (1)$

Taking log on both sides

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

$$Y = A + Bx \ldots\ldots\ldots\ldots (2)$$

$$where, Y = \log_{10} y, A = \log_{10} a, \mathrm{B} = b \log_{10} e$$

Equation (2) is linear equation in x and Y
So the normal equations are given by

$$\sum Y = nA + B \sum x$$

$$\sum xY = A \sum x + B \sum x^{2}$$

Problem: Fit acurve of the form $y = ae^{bx}$ to the following data

x	0	5	8	12	20
y	3.0	1.5	1.0	0.55	0.18

Sol: Let $y = ae^{bx} \ldots\ldots\ldots(1)$

Taking log on both sides

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

$$Y = A + Bx \ldots\ldots\ldots\ldots(2)$$

$$where, Y = \log_{10} y, A = \log_{10} a, \mathrm{B} = b \log_{10} e$$

Equation (2) is linear equation in x and Y
So the normal equations are given by

$$\sum Y = nA + B\sum x$$

$$\sum xY = A\sum x + B\sum x^2 \ldots\ldots\ldots\ldots(3)$$

x	y	$Y = \log_{10} y$	x^2	xY
0	3.0	0.4771	0	0
5	1.5	0.1761	25	0.8805
8	1.0	0	64	0
12	0.55	-0.2596	144	-3.1152
20	0.18	-0.7447	400	-14.894
$\sum x = 45$		$\sum Y = -0.3511$	$\sum x^2 = 633$	$\sum xY = -17.1287$

Sub these values in (3) we get

$$5A + 45B = -0.3511$$

$$45A + 633B = -17.1287$$

Solving above for A and B
A=0.4815 and B=-0.0613

$$A = \log_{10} a = 0.4815$$

$$\therefore a = (10)^{A} = (10)^{0.4815} = 3.0304$$

$$B = b \log_{10} e = -0.0613$$

$$\therefore b = \frac{-0.0613}{0.4343} = -0.1411$$

Sub a and b values in (1)

$$y = 3.0304e^{-0.1411x}$$

Maximum Likelihood Estimation (MLE)

- Suppose we have a random sample $x_1, x_2, .., x_n$ whose assumed probability distribution depends on some unknown parameter θ.
- Ex:

1) For Binomial unknown parameters are n, p.

2) For Poisson unknown parameter is λ.

Our goal is to find good estimator of θ(population parameter) using sample and which can be done with the help of MLE.

Maximum Likelihood function

- Let $x_1, x_2, .., x_n$ be i.i.d. random variables drawn from some probability distribution that depends on some unknown parameter θ.
- The goal of MLE to maximize likelihood function

$$L(\theta) = f(x_1, x_2, \ldots, x_n \mid \theta)$$

$$= f(x_1 \mid \theta) f(x_2 \mid \theta) \ldots f(x_n \mid \theta)$$

$$L(\theta) = \prod_{i=1}^{n} f(x_i / \theta)$$

- The maximum likelihood estimate (MLE) of θ is that value of θ that maximizes likelihood(θ).

It is defined as

$$L(\theta) = \prod_{i=1}^{n} f(x_i / \theta)$$

$$\log L(\theta) = \sum_{i=1}^{n} \log f(x_i / \theta)$$

For maximization, we have

$$\frac{dL}{d\theta} = 0 \; ; \quad \frac{d^2L}{d\theta^2} < 0$$

Example: An unfair coin is flipped 100 times, and 61 heads are observed. The coin either has probability $\frac{1}{3}, \frac{1}{2}$, or $\frac{2}{3}$ of flipping a head each time it is flipped. Which of the three is the MLE?

Solution: Here the distribution is the binomial distribution with $n = 100$.

$$P\left(H = 61 \mid p = \frac{1}{3}\right) = \binom{100}{61}\left(\frac{1}{3}\right)^{61}\left(\frac{2}{3}\right)^{39} \approx 9.6 \times 10^{-9}$$

$$P\left(H = 61 \mid p = \frac{1}{2}\right) = \binom{100}{61}\left(\frac{1}{2}\right)^{61}\left(\frac{1}{2}\right)^{39} = 0.007$$

$$P\left(H = 61 \mid p = \frac{2}{3}\right) = \binom{100}{61}\left(\frac{2}{3}\right)^{61}\left(\frac{1}{3}\right)^{39} = 0.040$$

Since $P\left(H = 61 \mid p = \frac{2}{3}\right)$ is maximum and hence MLE is $p = \frac{2}{3}$

p.m.f.

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$0 \le p \le 1$$

$$x = 0, 1, 2, \ldots, n;$$

Example: An unfair coin is flipped 100 times, and 61 heads are observed. What is the MLE when nothing is previously known about the coin?

Solution: Since the distribution follow is Binomial distribution, with parameter p.

Here $n = 100$. The likelihood function (MLE) is

$$P(H = 61|p) = \binom{100}{61} p^{61}(1-p)^{39}$$

For maximization

$$\frac{d}{dp} P(H = 61|p) = 0$$

$$\Rightarrow \binom{100}{61} [61p^{60}(1-p)^{39} - 39p^{61}(1-p)^{38}] = 0$$

$$\Rightarrow p^{60}(1-p)^{38}(61 - 100p) = 0$$

$$\Rightarrow p = 0, \frac{61}{100}, 1$$

Thus, the likelihoods are

$$P(H = 61|p = 0) = 0$$

$$P\left(H = 61 \middle| p = \frac{61}{100}\right) = \binom{100}{61}\left(\frac{61}{100}\right)^{61}\left(\frac{39}{100}\right)^{3}$$

$$P(H = 61|p = 1) = 0$$

Since $P\left(H = 61 \middle| p = \frac{61}{100}\right)$ is maximum and hence $p = \frac{61}{100}$ is the MLE.

Maximum Likelihood for a Binomial distribution

- Suppose we wish to find the maximum likelihood estimate (MLE) of θ for a Binomial distribution,

$$p_k(k,\theta) = nC_k \theta^k (1-\theta)^{n-k}$$

$$\log p_k(k,\theta) = \log(nC_k) + k\log(\theta) + (n-k)\log((1-\theta)$$

$$\frac{\partial \log p_k(k,\theta)}{\partial \theta} = 0 \Rightarrow 0 + \frac{k}{\theta} - \frac{n-k}{1-\theta} = 0$$

$$k - k\theta = n\theta - k\theta \Rightarrow \theta = \frac{k}{n}$$

Let $X_1, X_2, \ldots, X_n \in R$ be a random sample from a Poisson distribution

The p.d.f. of a Poisson Distribution is :

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$; where x =0,1,2,...

The likelihood function is:

$$L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = e^{-\lambda n} \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i}$$

The log-likelihood is:

$$lnL(\lambda) = -\lambda n + \sum_{i=1}^{n} xi.\, ln(\lambda) - ln(\prod_{i=1}^{n} xi)$$

Setting its derivative with respect to λ to zero, we have:

$$\frac{d}{d\lambda} lnL(\lambda) = -n + \sum_{i=1}^{n} xi. \frac{1}{\lambda} = 0$$

giving,

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{X}$$

which is the maximum likelihood estimate

Thank you