UNIT-V

APPLIED STATISTICS

- Curve fitting by the method of least squares- fitting of straight lines, second degree parabolas and more general curves
- Correlation and regression – Rank correlation

Dr.R.MURUGESAN/FME CIVIL-NRCM

Curve Fitting by the method of least squares:

Let (x_i, y_i); i=1, 2 , …. , m denote the set of data points. 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 be the error of approximation at $x=x_i$ then $e_i=y_i- f(x_i)$.

Let $S=[y_1 - f(x_1)]^2+[y_2 - f(x_2)]^2+[y_3 - f(x_3)]^2+\dots\dots+[y_m - f(x_m)]^2$

$=e_1{}^2+e_2{}^2+\dots\dots+e_m{}^2$

We minimize S by partially differentiating on both sides with respect to constants in the method of least squares.

Fitting of Straight line:

Let y= a+bx be the straight line to be fitted for the given data.

Then $S=[y_1 - (a + bx_1)]^2+[y_2 - (a + bx_2)]^2+[y_3 - (a + bx_3)]^2+\dots\dots$
$+[y_m - (a + bx_m)]^2$

$S=\sum_{i=1}^{m}[y_i - (a + bx_i)]^2$………. (1)

Partially differentiating (1) w.r.t 'a' and equating $\frac{\partial S}{\partial a}$ to 0, we get

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

$\Longrightarrow \sum_{i=1}^{m}[y_i - (a + bx_i)]=0$

$\Longrightarrow \sum_{i=1}^{m} y_i=\sum_{i=1}^{m} a+\sum_{i=1}^{m} bx_i$

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

$\Longrightarrow \sum_{i=1}^{m} y_i=ma+b \sum_{i=1}^{m} x_i$

(1) $\Rightarrow$ $\boxed{\therefore \sum_{i=1}^{m} y_i = ma + b\sum_{i=1}^{m} x_i}$

Partially differentiating (1) w.r.t 'b' and equating $\frac{\partial S}{\partial b}$ to 0, we get

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

$$\Rightarrow \sum_{i=1}^{m} x_i\ [y_i - (a + bx_i)] = 0$$

$$\Rightarrow \sum_{i=1}^{m} x_i\, y_i = \sum_{i=1}^{m} a x_i + \sum_{i=1}^{m} b x_i^2$$

$$\therefore \sum_{i=1}^{m} x_i\, y_i = a\sum_{i=1}^{m}(x_i) + b\sum_{i=1}^{m} x_i^2$$

$$\boxed{(2) \Rightarrow \therefore \sum_{i=1}^{m} x_i\, y_i = a\sum_{i=1}^{m}(x_i) + b\sum_{i=1}^{m} x_i^2}$$

The equations (1) and (2) are called normal equations.

Solving equations (1) and (2) , we get values of a and b which when substituted in the equation gives straight line of best fit.

Problem 1: By the method of least squares, find the straight line that best fits the following data:

x	1	2	3	4	5
y	14	27	40	55	68

Solution: Let y=a+bx…… (1) be the equation of straight line to be fitted to given set of points.

		2	
1	14	1	14
2	27	4	54
3	40	9	120

4	55	16	220
5	68	25	340
15	204	55	748

The normal equations of (1) are na +b$\sum x$=$\sum y$

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

$\Longrightarrow$5a+15b=204

15a+55b=748

Solving , we get a=0 and b=13.6

Substituting the values of a and b in (1), we get y=13.6 x which is the required straight line equation of best fit.

Problem 2: By the method of least squares, find the straight line that best fits the following data:

x	6	7	7	8	8	8	9	9	10
y	5	5	4	5	4	3	4	3	3

Solution: Let y=a+bx…… (1) be the equation of straight line to be fitted to given set of points.

		2	
6	5	36	30
7	5	49	35
7	4	49	28
8	5	64	40
8	4	64	32
8	3	64	24
9	4	81	36
9	3	81	27
10	3	100	30
72	36	588	282

The normal equations of (1) are na +b$\sum x$=$\sum y$

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

$\Rightarrow$9a+72b=36

72a+588b=282

Solving , we get a=-0.5 and b=8

Substituting the values of a and b in (1), we get y=-0.5+8x which is the required straight line equation of best fit.

<u>Fitting of Second degree polynomial or parabola:</u>

Let y= a+bx+cx^2 be the straight line to be fitted for the given data.

Then S=$[y_1 - (a + bx_1 + cx_1^2)]^2$+$[y_2 - (a + bx_2 + cx_2^2)]^2$+$[y_3 - (a + bx_3 + cx_3^2)]^2$+.......... +$[y_m - (a + bx_m + cx_m^2)]^2$

S=$\sum_{i=1}^{m}[y_i - (a + bx_i + cx_i^2)]^2$......... (1)

Partially differentiating (1) w.r.t 'a' and equating $\frac{\partial S}{\partial a}$ to 0, we get

$\frac{\partial S}{\partial a}$=$2\sum_{i=1}^{m}[y_i - (a + bx_i + cx_i^2)]\ (-1)$=0

$\Rightarrow \sum_{i=1}^{m}[y_i - (a + bx_i + cx_i^2)]$ =0

$\Rightarrow \sum_{i=1}^{m} y_i$=$\sum_{i=1}^{m} a$+$\sum_{i=1}^{m} bx_i$+$\sum_{i=1}^{m} c\ {}_i^2$

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

$\Rightarrow \sum_{i=1}^{m} y_i$=ma+$b\ \sum_{i=1}^{m} x_i$+c$\sum_{i=1}^{m} x_i^2$

(1) $\Rightarrow$ $\boxed{\therefore \sum_{i=1}^{m} y_i = ma + b\sum_{i=1}^{m} x_i + c\sum_{i=1}^{m} x_i^2}$

Partially differentiating (1) w.r.t 'b' and equating $\frac{\partial S}{\partial b}$ to 0, we get

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

$\Longrightarrow \sum_{i=1}^{m} x_i\ [y_i - (a + bx_i + cx_i^2)] =0$

$\Longrightarrow \sum_{i=1}^{m} x_i\, y_i=\sum_{i=1}^{m} ax_i+\sum_{i=1}^{m} bx_i^2+ c\sum_{i=1}^{m} x_i^3$

$\therefore \sum_{i=1}^{m} x_i\ y_i=a \sum_{i=1}^{m}(x_i)+b \sum_{i=1}^{m} x_i^2+ c\sum_{i=1}^{m} x_i^3$

$$\boxed{(2) \Longrightarrow \therefore \sum_{i=1}^{m} x_i\ y_i=a \sum_{i=1}^{m}(x_i)+b \sum_{i=1}^{m} x_i^2 + c \sum_{i=1}^{m} x_i^3}$$

Partially differentiating (1) w.r.t 'c' and equating $\frac{\partial S}{\partial c}$ to 0, we get

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

$\Longrightarrow \sum_{i=1}^{m} x_i^2\ [y_i - (a + bx_i + cx_i^2)] =0$

$\Longrightarrow \sum_{i=1}^{m} x_i^2\, y_i=\sum_{i=1}^{m} ax_i^2+\sum_{i=1}^{m} bx_i^3+ c\sum_{i=1}^{m} x_i^4$

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

$$\boxed{(3) \Longrightarrow \therefore \sum_{i=1}^{m} x_i^2\ y_i=a \sum_{i=1}^{m} x_i^2+b \sum_{i=1}^{m} x_i^3 + c \sum_{i=1}^{m} x_i^4}$$

The equations (1) , (2)and (3) are called normal equations.

Solving equations (1) , (2) and (3) we get values of a, b and c which when substituted in the equation gives parabola of best fit.

Fitting of exponential curve:

Let y=ae^{bx} .... (1) be the exponential curve to be fitted to given set of points.

Taking logarithm on both sides of (1),

log y=log ( )

$\Rightarrow$logy =log a+bx (loge)

$\Rightarrow Y = A + bx \ldots\ldots(2)$ where Y=logy , A=log a and loge=1

Eq(2) represents straight line equation with normal equations

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

A$\sum x$+ b$\sum x^2$=$\sum xY$

Solving these equations for constants A and b, we get the values and then substitute a=   and b values in (1) , we get required exponential curve of best fit.

Problem 1: Fit the curve of the form y=ae^{bx} to the following data:

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

Solution: Let y=ae^{bx} .... (1) be the exponential curve to be fitted to given set of points.

The normal equations

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

A$\sum x$+ b$\sum x^2$=$\sum xY$

x		=	2	
0	20	2.995	0	0
1	30	3.401	1	3.401
2	52	3.951	4	7.902
3	77	4.343	9	13.029
4	135	4.905	16	19.62
5	211	5.351	25	26.755
6	320	5.768	36	34.608
7	550	6.309	49	44.163
8	1052	6.958	81	55.644

36		43.981	204	205.142

$\therefore \sum x = 36, \sum Y = 43.981$, $\sum x^2$=204,$\sum xY = 205.142$,n=9

Substituting these values in the above normal equations, we get

9A+36b=43.981

36A+204b=205.142

Solving, we get a=e^A=18.495 and b=0.486

$\therefore$The required exponential curve of best fit is y=(18.495) $e^{(0.486)x}$.

Problem 2: Fit the curve of the form y=ae^{bx} to the following data:

x	0	0.5	1	1.5	2	2.5
y	0.10	0.45	2.15	9.15	40.35	180.75

Solution: Let y=ae^{bx} .... (1) be the exponential curve to be fitted to given set of points.

The normal equations

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

A$\sum x$+ b$\sum x^2$=$\sum xY$

x		=	2	
0	0.1	-2.3026	0	0
0.5	0.45	-0.7985	0.25	-0.39925
1	2.15	0.7655	1	0.7655
1.5	9.15	2.2138	2.25	3.3207
2	40.35	3.6976	4	7.3952
2.5	180.75	5.1971	6.25	14.107
7.5		8.7729	13.75	25.1892

$\therefore \sum x = 7.5, \sum Y = 8.7729, \sum x^2$=13.75,$\sum xY = 25.1892$,n=6

Substituting these values in the above normal equations, we get

6A+7.5b=8.7729

7.5A+13.75b=25.1892

Solving, we get a= =0.058 and b=3.448

∴The required exponential curve of best fit is y=(0.058) $e^{(3.448)x}$.

<u>Fitting of Power curve:</u>

Let y=ab^x …. (1) be the power curve to be fitted to given set of points.

Taking logarithm on both sides of (1),

log y=log ( )

⟹logy =log a +x (logb)

$\Longrightarrow Y = A + Bx \ldots..(2)$ where Y=logy , A=log a and B=logb

Eq(2) represents straight line equation with normal equations

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

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

Solving these equations for constants A and b, we get the values and then substitute a= and b= values in (1) , we get required power curve of bestfit.

Problem 1: Fit the curve of the form y=a to the following data:

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

Solution: Let y=ab^x …. (1) be the power curve to be fitted to given set of points.

The normal equations

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

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

x		=	2	
2	8.3	0.9191	4	1.8382
3	15.4	1.1875	9	3.5625
4	33.1	1.5198	16	6.0792
5	65.2	1.8142	25	9.0710
6	127.4	2.1052	36	12.631
20		7.5455	90	33.1819

5A+20B=7.5455

20A+90B=33.1819

Solving A=0.31 and B=0.3

$\Rightarrow$a=$e^{0.31}$=1.3634 and b=$e^{0.3}$=1.3499

$\therefore$The required power curve of best fit is y=(1.3634)$(1.3499)^x$

Problem 2: Fit the curve of the form y=a to the following data:

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

Solution: Let y=ab^x …. (1) be the power curve to be fitted to given set of points.

The normal equations

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

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

x		=	2	
0	10	1	0	0
1	21	1.3222	1	1.3222
2	35	1.5441	4	3.0882
3	59	1.7708	9	5.3124
4	92	1.9638	16	7.8552
5	200	2.3010	25	11.5050
6	400	2.6021	36	15.6126
7	610	2.7853	49	19.4971
28		15.2893	140	64.1927

8A+28B=15.2893

28A+140B=64.1927

Solving A=1.02115 and B=0.2543

$\Rightarrow$a=$e^{1.02115}$ =2.7764 and b=$e^{0.2543}$=1.2896

$\therefore$The required power curve of best fit is y=(2.7764)$(1.2896)^x$

<u>CORRELATION:</u>

Definition: Correlation is the study of relationship between two or more variables.

Types of correlation:

Correlation is classified into the following types:

1. Positive and Negative Correlation: This depends on the change of variables. If two variables tend to move in the same direction i.e. increase in the value of one variable is associated with increase in another variable or a decrease in the value of one variable is associated with decrease in another variable. Examples: Height and weight, rainfall and production of crops etc. This type of correlation is called Positive Correlation.

If increase or decrease in the value of one variable is associated with decrease or increase in another variable then this type of correlation is called Negative Correlation. Examples: Price and demand etc.

1. Simple and Multiple Correlation: If relationship between two variables is defined then such type of Correlation is called Simple Correlation. If relationship between many variables is defined then such type of Correlation is called Multiple Correlation.

2. Partial and Total Correlation: If the relationship exists between some variables from many variables then such type of correlation is called partial correlation and If the relationship exists between all variables then such type of correlation is called total correlation.

3. Linear and non-linear Correlation: If the ratio of change between two variables is uniform, then linear Correlation exists otherwise non-linear correlation between variables.

   Methods of studying correlation:

   1. Graphic methods and (2) Mathematical methods

1. <u>Graphic methods:</u>

(a)Scatter diagram and (b) simple graph

(a) Scatter diagram: The scatter diagram is a chart obtained by plotting two variables to find out whether there is any relationship between them.

(b)Simple graph: The values of the two variables are plotted on a graph paper. We get two curves, one for X variable and other for variable Y. These two curves reveal the direction and closeness of the two variables and also whether they are related or not. If both the curves move in the same direction i.e either upward or downward then correlation is said to be positive otherwise if they move in opposite direction then correlation is said to be negative.

2. <u>Mathematical methods:</u>

(a)Karl Pearson's coefficient of correlation

(b)Spearman's Rank coefficient of correlation

(a)Karl Pearson's coefficient of correlation:

This method is also called Pearson or Product-Moment correlation coefficient denoted by 'r'. The formula for calculating r is r=$\frac{\sum XY}{\sqrt{(X^2\ )(Y^2)}}$

<u>Properties of correlation coefficient:</u>

1. The correlation coefficient lies between -1 and +1(i.e -1≤r≤1)

    If r=1 then correlation is perfect and positive.

    If r=-1 then correlation is perfect and negative.

    If r=0 then there is no relationship between the variables.

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

3. If X and Y are random variables and a, b, c, d are any numbers such that a≠0, c≠0 then r(aX+b, cY+d)=$\frac{ac\ r(X,Y)}{|ac|}$

4. Two independent variables are uncorrelated .i.e r(X,Y)=0.

Problem 1: Calculate coefficient of correlation from the following data:

X	12	9	8	10	11	13	7
Y	14	8	6	9	11	12	3

S lution:

X	Y	2	2	
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
70	63	728	651	676

$\therefore \sum X = 70, \sum Y = 63, \sum X^2$=728, $\sum Y^2 = 651, \sum XY = 676$

coefficient of correlation is given by r=$\frac{\sum XY}{\sqrt{(X^2\ )(Y^2)}}$

$\Rightarrow$r=$\frac{676}{\sqrt{(728)(651)}}$=$\frac{676}{\sqrt{(473928)}}$=0.982

Problem 2: Calculate coefficient of correlation from the following data:

X	57	59	62	63	64	65	55	58	57
Y	113	117	126	126	130	129	111	116	112

Solution:

X	Y	x=X-$\bar{X}$	2	y=Y-$\bar{Y}$	2	xy
57	113	-3	9	-7	49	21
59	117	-1	1	-3	9	3
62	126	2	4	6	36	12
63	126	3	9	6	36	18
64	130	4	16	10	100	40
65	129	5	25	9	81	45
55	111	-5	25	-9	81	45

58 57	116 112	-2 -3	4 9	-4 -8	16 64	8 24
540	1080		102		472	216

$\bar{X}=\frac{\sum X}{n}=\frac{540}{9}=60$, $\bar{Y}=\frac{\sum Y}{n}=\frac{1080}{9}=120$

$\sum x^2=102, \sum y^2 = 472, \sum xy = 216$

coefficient of correlation is given by $r=\frac{\sum XY}{\sqrt{(X^2\ )(Y^2)}}$

$\Rightarrow r=\frac{216}{\sqrt{(102)(472)}}=\frac{216}{\sqrt{(48144)}}=0.9844$

$NOTE$:When actual mean is not a whole number, but a fraction(decimals) then correlation coefficient is given by the formula

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

Problem 1: Calculate Karl Pearson's coefficient of correlation from the following data:

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

Solution:

X	Y	x=X-$\bar{X}$	^{2}	y=Y-$\bar{Y}$	^{2}	xy
28	23	-0.75	0.5625	-8.3	68.89	6.225
41	34	5.5	30.25	2.7	7.29	14.85
40	33	4.5	20.25	1.7	2.89	7.65
38	34	2.5	6.25	2.7	7.29	6.75
35	30	-0.5	0.25	-1.3	1.69	0.65
33	26	-2.5	6.25	-5.3	28.09	13.25

40	28	4.5	20.25	-3.3	10.89	14.85
32	31	-3.5	12.25	-0.3	0.09	1.05
36	36	0.5	0.25	4.7	22.09	2.35
33	38	-2.5	6.25	6.7	44.89	16.75
355	313	7.75	102.8125	0		

$\bar{x}=\frac{\sum X}{n}=\frac{355}{10}=35.5$, $\bar{y}=\frac{\sum Y}{n}=\frac{313}{10}=31.3$

$\sum x^2$=102, $\sum y^2 = 472, \sum xy = 216$

coefficient of correlation is given by r=$\frac{N\sum XY-\sum X\sum Y}{\sqrt{(N\sum X^2-(\sum X)^2}\sqrt{N\sum Y^2-(\sum Y)^2}}$

$\Rightarrow$r=$\frac{216}{\sqrt{(102)(472)}}$=$\frac{216}{\sqrt{(48144)}}$=0.9844

Rank Correlation Coefficient:

It is based on the ranks given to the observations. It is applicable only to the individual observations. The formula for Spearman's rank correlation is given by

r= 1 - $\frac{6\sum D^2}{N(N^2-1)}$ where r=rank coefficient correlation , ^{2}=Sum of the squares of the differences of two ranks, N=number of observations.

Problem 1: Following are the rank obtained by 10 students in two subjects, Statistics and Mathematics. To what extent the knowledge of the students in two subjects is related?

Statistics	1	2	3	4	5	6	7	8	9	10
Mathematics	2	4	1	5	3	9	7	10	6	8

Solution: Let X denote the ranks in Statistics and Y denote the ranks in Mathematics.

X	Y	D=X-Y	^{2}

1	2	-1	1
2	4	-2	4
3	1	2	4
4	5	-1	1
5	3	2	4
6	9	-3	9
7	7	0	0
8	10	-2	4
9	6	3	9
10	8	2	4
			40

Therefore r= 1 $-\frac{6\sum D^2}{N(N^2-1)}$

$= 1 - \frac{6(40)}{10((10)^2-1))}$

$=1 - \frac{240}{10(99)} = 0.76$

Problem 2: A random sample of 5 college students is selected and their grades in Mathematics and Statistics are found to be

	1	2	3	4	5
Mathematics	85	60	73	40	90
Statistics	93	75	65	50	80

Calculate Spearman's rank correlation coefficient.

Sol: Let X denote grades in Mathematics and Y denote grades in Statistics.

Let x denote ranks in Mathematics and y denote ranks in Statistics.

X	Y	x	y	D=x-y	^{2}
85	93	2	1	1	1
60	75	4	3	1	1
73	65	3	4	-1	1
40	50	5	5	0	0
90	80	1	2	1	1
					4

Therefore r= $1 - \frac{6\sum D^2}{N(N^2-1)}$

$= 1 - \frac{6(4)}{5(25-1))}$

$=1 - \frac{24}{5(24)} =0.8$

Problem 3: Ten competitors in a musical test were ranked by the three judges A,B,C in the following order.

Ranks by A	1	6	5	10	3	2	4	9	7	8
Ranks by B	3	5	8	4	7	10	2	1	6	9
Ranks by C	6	4	9	8	1	2	3	10	5	7

Using rank correlation coefficient, discuss which pair of judges has the nearest approach to common likings in music.

Sol: Let X denote Ranks by A , Y denote Ranks by B and Z denote Ranks by C.

X	Y	Z	D_1=X-Y	D_2=Y-Z	D_3=X-Z	$^2_1$	$^2_2$	$^2_3$

1	3	6	-2	-3	-5	4	9	25
6	5	4	1	1	2	1	1	4
5	8	9	-3	-1	-4	9	1	16
10	4	8	6	-4	2	36	16	4
3	7	1	-4	6	2	16	36	4
2	10	2	-8	8	0	64	64	0
4	2	3	2	-1	1	4	1	1
9	1	10	8	-9	-1	64	81	1
7	6	5	1	1	2	1	1	4
8	9	7	-1	2	1	1	4	1
						200	214	60

Therefore $r(X,Y) = 1 - \frac{6\sum D_1^2}{N(N^2-1)}$

$= 1 - \frac{6(200)}{10(100-1)}$

$= 1 - \frac{1200}{990} = -0.2121$

$r(Y,Z) = 1 - \frac{6\sum D_2^2}{N(N^2-1)}$

$= 1 - \frac{6(214)}{10(100-1)}$

$= 1 - \frac{1284}{990} = -0.2969$

$r(X,Z) = 1 - \frac{6\sum D_3^2}{N(N^2-1)}$

$= 1 - \frac{6(60)}{10(100-1)}$

$= 1 - \frac{360}{990} = 0.6364$

Since r(X,Z) is maximum, we conclude that the pair of judges A and C has the nearest approach to common likings in music.

<u>Rank Correlation Coefficient for equal or repeated ranks:</u>

The formula for rank correlation coefficient for equal or repeated ranks is given by r=1- $\frac{6\{D^2+\frac{(m^3-m)}{12}+\frac{(m^3-m)}{12}+\cdots}{N^3-N}$ where m denotes the number of items whose ranks are common.

Problem 1: From the following data calculate the rank correlation coefficient for the following data:

X	48	33	40	9	16	16	65	24	16	57
Y	13	13	24	6	15	4	20	9	6	19

Solution:

X	Y	x	y	D=x-y	2
48	13	8	5.5	2.5	6.25
33	13	6	5.5	0.5	0.25
40	24	7	10	-3	9
9	6	1	2.5	-1.5	2.25
16	15	3	7	4	16
16	4	3	1	2	4
65	20	10	9	1	1
24	9	5	4	1	1
16	6	3	2.5	5	0.25
57	19	9	8	1	1
					41

r=1- $\frac{6\{41+\frac{(3^3-3)}{12}+\frac{(2^3-2)}{12}+\frac{(2^3-2)}{12}+\cdots}{(10)^3-10}$

=0.733

Problem 2: A sample of 12 fathers and their elder sons gave the following data about their elder sons. Calculate the coefficient of rank correlation.

Fathers	65	63	67	64	68	62	70	66	68	67	69	71
Sons	68	66	68	65	69	66	68	65	71	67	68	70

Solution:Let X denote Fathers and Y denote sons.Let x denote ranks of X and y denote ranks of Y

X	Y	x	y	D=x-y	2
65	68	9	5.5	3.5	12.25
63	66	11	9.5	1.5	2.25
67	68	6.5	5.5	1	1
64	65	10	11.5	-1.5	2.25
68	69	4.5	3	1.5	2.25
62	66	12	9.5	2.5	6.25
70	68	2	5.5	-3.5	12.25
66	65	8	11.5	3.5	12.25
68	71	4.5	1	-3.5	12.25
67	67	6.5	8	-1.5	2.25
69	68	3	5.5	-2.5	6.25
71	70	1	2	-1	1
					72.5

$$r=1-\frac{6\{72.5+\frac{(2^3-2)}{12}+\frac{(2^3-2)}{12}+\frac{(4^3-4)}{12}+\frac{(2^3-2)}{12}+\frac{(2^3-2)}{12}+\cdots..}{(12)^3-12}$$

$=0.722$

<u>Regression</u>

<u>Definition:</u>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 representing the two variables is called line of regression.

<u>Methods of studying Regression:</u>

1.Graphic method 2.Algebraic method

1.Graphic method: In this method, two variables are plotted on X-axis and Y-axis.These points form a scatter diagram.A regression line is drawn between these points by free hand.

2.Algebraic method: In this method, a regression line is a straight line fitted to the data by the method of least squares.There are two lines of regression X on Y and Y on X.

The regression equation of X on Y is X=a+bY with normal equations na+b∑Y=∑X and a∑Y+b∑Y^2=∑XY
The regression equation of Y on X is Y=a+bX with normal equations na+b∑X=∑Y and a∑X+b∑X^2=∑XY

Problem 1:Find the regression equation of X on Y and Y on X for the following data:

X	10	12	13	16	17	20	25
Y	10	22	24	27	29	33	37

Solution:

X	Y	^{2}	^{2}	XY
10	10	100	100	100
12	22	144	484	264
13	24	169	576	312
16	27	256	729	432
17	29	289	841	493
20	33	400	1089	660
25	37	625	1369	925
113	182	1938	5188	3186

X on Y:

7a+182b=113 and 182a+1938b=3186

Implies a=18.45 ,b=-0.089

Therefore the regression equation of X on Y is X=a+bY

Implies X=18.45-0.089 Y

Y On X:

7a+113b=182 and 113a+1938b=3186

Implies a=0.82 and b=1.56

Therefore the regression equation of Y on X is Y=a+bX

Implies Y=0.82+1.56

<u>Regression lines with Actual mean</u>

Regression line of X on Y is given by $X-\bar{X}=b_{xy}(Y-\bar{Y})$ where

$b_{xy}=\frac{\sigma_x}{\sigma_y}=\frac{\sum XY}{\sum X^2}$

Regression line of Y on X is given by $Y-\bar{Y}=b_{xy}(X-\bar{X})$ where

$b_{yx}=\frac{\sigma_y}{\sigma_x}=\frac{\sum XY}{\sum Y^2}$

PROBLEM 1: A panel of two judges P and Q graded seven dramatic performances by independently awarding marks as follows:

Performance	1	2	3	4	5	6	7
Marks by P	46	42	44	40	43	41	45
Marks by Q	40	38	36	35	39	37	41

The eigth performance, which judge Q would not attend, was awarded 37 marks by judge P. If Judge Q had also been present , how many marks would be expected to have been awarded by him to the eigth performance.

Solution: Let X denote marks by P and Y denote marks by Q.

Performance	X	$x=X-\bar{X}$	Y	$y=Y-\bar{Y}$	2	xy
1	46	3	40	2	9	6
2	42	-1	38	0	1	0
3	44	1	36	-2	1	2
4	40	-3	35	-3	9	9
5	43	0	39	1	0	0
6	41	-2	37	-1	4	2
7	45	2	41	3	4	6
	301		266		28	21

$\bar{X}=\frac{\sum X}{n}=\frac{301}{7}=43$, $\bar{Y}=\frac{\sum Y}{n}=\frac{266}{7}=38$

Regression line of Y on X is given by $Y-\overline{Y}=b_{xy}(X-\overline{X})$ where

$b_{yx}=\frac{r\sigma_y}{\sigma}=\frac{\sum XY}{\sum Y^2}$

Therefore $b_{yx}=\frac{\sum XY}{\sum X^2}=\frac{21}{28}=0.75$

Therefore Regression line of Y on X is given by $Y-\overline{Y}=b_{xy}(X-\overline{X})$

Implies Y-38=0.75(X-43)

Implies Y=0.75X-32.25+38

Implies Y=0.75X+5.75

If X=37 then Y=0.75(37)+5.75=33.5

Problem 2: Find the most likely production corresponding to a rainfall 40 from the following data:

	Rainfall	Production
Average	30	500 kgs
Standard Deviation	5	100 kgs
Coefficient of Correlation	0.8	

Solution: Let X denote Rainfall and Y denote Production.

Given $\overline{X}=30$, $\overline{Y}=500$, $\sigma_x=5$, $\sigma_y=100$, r=0.8

Regression line of Y on X is given by $Y-\overline{Y}=b_{xy}(X-\overline{X})$ where

$=\frac{(0.8)(100)}{5}=16$

Therefore Y-500=16(X-30)

Implies Y=16X-480+500

Implies Y=16X+20

For rainfall of 40,i.e. X=40 ,Y=16(40)+20=660

Thus for a rainfall of 40 the production is 660 kgs.

Problem 3: Comment the following:

Regression coefficient of Y on X is 0.7 and X on Y is 3.2.

Solution: since r^2=$b_{xy}b_{yx}$=(0.7)(3.2)=2.24

Implies r=$\sqrt{2.24}$=1.5(approximately)

Hence the data is inconsistent.

Angle between two regression lines:

If θ is the angle between two regression lines X on Y and Y on X then

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

Note:

1.If r=0 then tan θ=∞⟹ θ=$\frac{\pi}{2}$ .i.e, there is no relationship between the two variables or the two variables are independent.

2.If r=±1 then tan θ=0 ⟹ θ= 0 or π. Hence the two regression lines are parallel.

Problem 1: 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 θ.

Solution: Given =2 ,r=0.25

If θ is the angle between two regression lines X on Y and Y on X then

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

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

=3.75(2/5)=1.5

Problem 2: If $\sigma_{x=}\sigma_y$=σ and the angle between the regression lines is $Tan^{-1}(\frac{4}{3})$, find r.

Solution: tan θ= $\frac{(1-r^2)}{r}\frac{\sigma_x\sigma_y}{(\sigma_x^2+\sigma_y^2}$

Given $\sigma_{x=}\sigma_y$=σ ⟹ tan θ= $\frac{(1-r^2)}{r}\frac{\sigma\,\sigma}{(\sigma^2+\sigma^2}$

⟹ tan θ=$\frac{(1-r2)}{r}\frac{\sigma^2}{2\sigma^2}$

⟹ tan θ=$\frac{(1-r2)}{2r}$

⟹ θ=$Tan^{-1}[\frac{(1-r^2)}{2r}]$……..(1) **Dr.R.MURUGESAN/FME**

Also given θ=$Tan^{-1}(\frac{4}{3})$

Therefore $\frac{(1-r^2)}{2r}=\frac{4}{3}$

$\Rightarrow 3r^2+8r-3=0$

$\Rightarrow r=\frac{1}{3}$ or -3

Since -1≤r≤1, we take $r=\frac{1}{3}$

UNIT-IV

TESTING OF HYPOTHESIS

- Test of significance: Null Hypothesis; Alternative Hypothesis-Type Error –Type II Error
- Large sample test for single proportion
- Difference of proportions
- Single mean
- Difference of means
- Test for single mean; Difference of means for small samples
- Test for ratio of variances for small samples