ADVANCED ENGINEERING MATHEMATICS

IT (III SEM)

Course Objectives:

CO1 To prepare many optimization tools for applying them into different research areas as per the requirement.

CO2 To prepare important strategies of linear programming for applying them into solving the problems of transportation and assignment.

CO3 To establish different theorems and properties of random variables for understanding expectation, moments, moment generating function etc.

CO4 To analyze the various discrete and continuous distributions with their appropriate applications.

CORRELATION AND REGRESSION (CO4)

By Dr. Anil Maheshwari

Assistant Professor, Mathematics

Engineering College, Ajmer

INTRODUCTION

Many times we required to analyze that how a change in one variable affects the change in other variable. Such analysis comes under the field of correlation and regression.

CORRELATION

The two variables are said to be correlated if change in one variable affects the change in other variable e.g. price and demand of a commodity are correlated, income and expenditure are correlated etc. Here, we will discuss two types of coefficients of correlation i.e. Karl Pearson's correlation coefficient and rank correlation coefficient.

REGRESSION

Through regression, we can fit a curve over a given data. Here, we will discuss fitting of a straight line over a given data, which is called linear regression. Under this section, we will discuss two types of lines of regression i.e. the line of regression of Y on X (best estimate of Y for any given X) and the line of regression of X on Y (best estimate of X for any given Y).

We will also discuss the fitting of curves for a given data through method of least squares in this section.

KARL PEARSON'S CORRELATION COEFFICIOENT

KARL PEARSON'S CORRELATION COEFFICIENT

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}} \quad \text{OR} \quad r = \frac{\frac{1}{n}\sum xy - \bar{x}\bar{y}}{\sqrt{\frac{1}{n}\sum x^2 - \bar{x}^2}\sqrt{\frac{1}{n}\sum y^2 - \bar{y}^2}}$$

Karl Pearson's correlation coefficient measures the degree of linear relationship between two variables. It lies between -1 and 1 i.e. $-1 \le r \le 1$

Remarks : (1) If $r = 0$, then variables are said to be uncorrelated.

(2) If $r = 1$, then variables are said to be perfectly and positively correlated.

(3) If $r = -1$, then variables are said to be perfectly and negatively correlated.

(4) If $0 < r < 1$, then variables are said to be positively correlated.

(5) If $-1 < r < 0$, then variables are said to be negatively correlated.

Q.1. *Calculate the coefficient of correlation between the heights of fathers (x) and their sons (y) using the following data :*

x (inches) :	65	66	67	67	68	69	70	72
y (inches) :	67	68	65	68	72	72	69	71

Sol.

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

Here, $\bar{x} = \dfrac{\sum x}{n} = \dfrac{544}{8} = 68$ and $\bar{y} = \dfrac{\sum y}{n} = \dfrac{552}{8} = 69$

Now, Karl Pearson's coefficient of correlation

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

$$= \frac{24}{\sqrt{36}\sqrt{44}}$$

$$= \frac{24}{39.7994975}$$

$$= 0.60302269$$

Q.2. *Find the correlation coefficient between x and y, when it is given that :*

$n = 15, \quad \Sigma x = 50, \quad \Sigma y = -30, \quad \Sigma x^2 = 290, \quad \Sigma y^2 = 300, \quad \Sigma xy = -115$

Sol. We are given that :

$n = 15, \Sigma x = 50, \Sigma y = -30, \quad \Sigma x^2 = 290, \quad \Sigma y^2 = 300, \Sigma xy = -115$

Now, Coefficient of correlation

$$r = \frac{\frac{1}{n}\sum xy - \bar{x}\bar{y}}{\sqrt{\frac{1}{n}\sum x^2 - \bar{x}^2}\sqrt{\frac{1}{n}\sum y^2 - \bar{y}^2}} = \frac{-\frac{115}{15} - \left(\frac{50}{15}\right)\left(\frac{-30}{15}\right)}{\sqrt{\frac{290}{15} - \left(\frac{50}{15}\right)^2}\sqrt{\frac{300}{15} - \left(\frac{-30}{15}\right)^2}}$$

$$= \frac{-\frac{115}{15} + \frac{100}{15}}{\sqrt{\frac{290}{15} - \frac{100}{9}}\sqrt{\frac{300}{15} - 4}} = \frac{-1}{\sqrt{8.2222}\sqrt{16}} = -\frac{1}{114698}$$

$= -0.0872$

RANK CORRELATION COEFFICIOENT

SPEARMAN'S RANK CORRELATION COEFFICIENT

$$r = 1 - \frac{6\Sigma D^2}{n\left(n^2 - 1\right)}$$

Rank correlation is useful, when the data is ranked according to the particular characteristics instead of taking numeric measurements on them e.g. honesty, morality etc. In such cases, data turns out to be qualitative (non - numeric) in nature. It lies between -1 and 1 i.e. $-1 \leq r \leq 1$

Q.1. *The ranking of 10 students in two subjects Mathematics and Statistics are as follows :*

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

What is the coefficient of rank correlation ?

Sol.

Ranks of Mathematics R_x	*Ranks of Statistics* R_y	$D = R_x - R_y$	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
			$\Sigma D^2 = 214$

Now, rank correlation coefficient is given by

$$r = 1 - \frac{6\Sigma D^2}{n(n^2 - 1)}$$

$$= 1 - \frac{6 \times 214}{10(100 - 1)} = 1 - \frac{1284}{990}$$

$$= 1 - 1.2969697$$

$$= -0.2969697$$

Q.2. *Compute rank correlation coefficient from the following data of marks obtained by eight students in the papers of Physics and Mathematics :*

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

Calculate coefficient of rank correlation .

Sol.

Marks in Physics (x)	*Marks in Mathematics* (y)	R_x	R_y	$D = R_x - R_y$	D^2
15	**50**	7	**3**	**4**	**16**
20	**30**	**5.5**	**5**	**0.5**	**0.25**
27	**55**	**4**	**2**	**2**	**4**
13	**30**	**8**	**5**	**3**	**9**
45	**25**	**3**	7	**−4**	**16**
60	**10**	**2**	**8**	**−6**	**36**
20	**30**	**5.5**	**5**	**0.5**	**0.25**
75	70	**1**	**1**	**0**	**0**
					$\Sigma D^2 = 81.5$

Two students have secured equal marks 20 in Physics, so the rank awarded to them is the arithmetic mean of the ranks that they would have got viz. 5 and 6. Hence the rank awarded to them is $\frac{1}{2}(5+6) = 5.5$

Similar process is for awarding ranks to students in Mathematics.

For marks in Physics $m_1 = 2$ and in Maths $m_2 = 3$, where m is the number of times a particular mark has been repeated.

Rank correlation coefficient is given as

$$r = 1 - \frac{6\left[\Sigma D^2 + \frac{(m_1^3 - m_1)}{12} + \frac{(m_2^3 - m_2)}{12}\right]}{n(n^2-1)}$$

$$= 1 - \frac{6\left[815 + \frac{(2^3 - 2)}{12} + \frac{(3^3 - 3)}{12}\right]}{8(64-1)}$$

$$= 0$$

Q.3. *Ten competitors in a beauty contest are ranked by three judges in the following order :*

I Judge	:	1	6	5	10	3	2	4	9	7	8
II Judge	:	3	5	8	4	7	10	2	1	6	9
III Judge	:	6	4	9	8	1	2	3	10	5	7

Which two judges have the nearest approach to common test in beauty?

Sol. In order to find the required output, we compare the rank correlation between the judgements of :

(*a*) I and II judge

(*b*) II and III judge

(*c*) I and III judge

Rank by I Judge (R_1)	Rank by II Judge (R_2)	Rank by III Judge (R_3)	$D_1 = R_1 - R_2$	$D_2 = R_2 - R_3$	$D_3 = R_3 - R_1$	D_1^2	D_2^2	D_3^2
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
						$\sum D_1^2 = 200$	$\sum D_2^2 = 214$	$\sum D_3^2 = 60$

Rank correlation coefficient between the judgements of I and II judges

$$r = 1 - \frac{6\Sigma D_1^2}{n(n^2-1)}$$

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

$$= 1 - \frac{120}{99}$$

$$= 1 - 1.21212121$$

$$= -0.21212121$$

Rank correlation coefficient between the judgements of II and III judges

$$r = 1 - \frac{6\Sigma D_2^2}{n(n^2-1)} = 1 - \frac{6 \times 214}{10(100-1)}$$

$$= 1 - 1.2969697 = -0.2969697$$

Rank correlation coefficient between the judgements of I and III judges

$$r = 1 - \frac{6\Sigma D_3^2}{n(n^2-1)} = 1 - \frac{6 \times 60}{10(100-1)}$$

$$= 1 - 0.36363636 = 0.63636364$$

Since coefficient of correlation is maximum in the judgements of the first and third judges, so we conclude that they have the nearest approach to common test in beauty.

REGRESSION

LINES OF REGRESSION

line of regression of Y on X

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

line of regression of X on Y

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

Remarks : (I) Both the lines of regression pass through the point $(\bar{x}, \bar{y})$. Hence $(\bar{x}, \bar{y})$ is the point of intersection of two lines of regression.

(II) We have $b_{xy} = \frac{r\sigma_x}{\sigma_y}$ and $b_{yx} = \frac{r\sigma_y}{\sigma_x}$ as regression coefficients.

Now, $$b_{xy} \cdot b_{yx} = \frac{r\sigma_x}{\sigma_y} \times \frac{r\sigma_y}{\sigma_x} = r^2$$

or $$r = \pm\sqrt{b_{xy} \cdot b_{yx}}$$

Here, r is positive, if b_{xy} and b_{yx} both are positive
and r is negative, if b_{xy} and b_{yx} both are negative.

(III) As we have $-1 \le r \le 1$

or $r^2 \le 1$

or $b_{xy} \cdot b_{yx} \le 1$

Q.1. ***From the following data, obtain the two regression equations :***

x :	6	2	10	4	8
y :	9	11	5	8	7

Sol.

x	y	$x-\bar{x}$	$(x-\bar{x})^2$	$y-\bar{y}$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
6	9	0	0	1	1	0
2	11	−4	16	3	9	−12
10	5	4	16	−3	9	−12
4	8	−2	4	0	0	0
8	7	2	4	−1	1	−2
$\Sigma x = 30$	$\Sigma y = 40$	$\Sigma(x-\bar{x}) = 0$	$\Sigma(x-\bar{x})^2 = 40$	$\Sigma(y-\bar{y}) = 0$	$\Sigma(y-\bar{y})^2 = 20$	$\Sigma(x-\bar{x})(y-\bar{y}) = -26$

Here, $\bar{x} = \frac{\Sigma x}{n} = \frac{30}{5} = 6$ ***and*** $\bar{y} = \frac{\Sigma y}{n} = \frac{40}{5} = 8$

Now, **regression** equation of y on x is $(y - \bar{y}) = \dfrac{r\sigma_y}{\sigma_x}(x - \bar{x})$,

where $\dfrac{r\sigma_y}{\sigma_x} = \dfrac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2} = \dfrac{-26}{40} = -0.65$

So, equation becomes $y - 8 = -0.65\,(x - 6)$

or $y = 11.9 - 0.65x$

and **regression** equation of x on y is $(x - \bar{x}) = \dfrac{r\sigma_x}{\sigma_y}(y - \bar{y})$,

where $\dfrac{r\sigma_x}{\sigma_y} = \dfrac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(y - \bar{y})^2} = \dfrac{-26}{20} = -1.3$

So, equation becomes $x - 6 = -1.3\,(y - 8)$

or $x = -1.3y + 16.4$

Q.2. *Two random variables have the regression lines with equations* $3x + 2y - 26 = 0$ *and* $6x + y - 31 = 0$. *Find the mean values and the coefficient of correlation between* x *and* y.

Sol. The lines of regression are :

$$3x + 2y - 26 = 0 \qquad \ldots(1)$$

and $6x + y - 31 = 0$ $\qquad \ldots(2)$

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

$$3\bar{x} + 2\bar{y} = 26 \quad \text{and} \quad 6\bar{x} + \bar{y} = 31$$

On solving these two equations, we get $\bar{x} = 4$, $\bar{y} = 7$

Now, equations (1) and (2) can be rewritten as

$$y = \frac{-3}{2}x + 13 \qquad \ldots(3)$$

and $$x = \frac{-1}{6}y + \frac{31}{6} \qquad \ldots(4)$$

From equations (3) and (4), it is obvious to say that

$$b_{yx} = \frac{-3}{2} \text{ and } b_{xy} = \frac{-1}{6}$$

So, $$b_{yx} \cdot b_{xy} = \left(\frac{-3}{2}\right) \cdot \left(\frac{-1}{6}\right) = \frac{1}{4}$$

and $$r = \pm \sqrt{b_{yx} \cdot b_{xy}} = \pm \sqrt{\frac{1}{4}} = \pm 0.5$$

Since both the coefficients b_{yx} and b_{xy} are negative,

$$r = -0.5$$

Q.3. *Show that θ, an acute angle between the two lines of regression, is given by*

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

Interpret the case, when $r = 0, \pm 1$

Sol. The regression line of y on x is $(y-\bar{y}) = \dfrac{r\sigma_y}{\sigma_x}(x-\bar{x})$ ...(1)

Slope of line (1) is $m_1 = \dfrac{r\sigma_y}{\sigma_x}$

The regression line of x on y is $(x-\bar{x}) = \dfrac{r\sigma_x}{\sigma_y}(y-\bar{y})$ ...(2)

Slope of line (2) is $m_2 = \dfrac{\sigma_y}{r\sigma_x}$

Let θ be an acute angle between the lines (1) and (2),

$$\text{then} \quad \tan\theta = \pm \frac{(m_1 - m_2)}{(1 + m_1 m_2)}$$

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

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

Taking positive sign,

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

$\Rightarrow$ θ will be an obtuse angle.

Taking negative sign,

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

$\Rightarrow \theta$ will be an acute angle.

If $r = 0$, then $\tan \theta = \infty \Rightarrow \theta = \pi/2$ *i.e.* both the regression lines are perpendicular to each other.

If $r = \pm 1$, then $\tan \theta = 0 \Rightarrow \theta = 0$ or π *i.e.* both the regression lines coincide each other and there is perfect correlation between the variables involved.

CURVE FITTING THROUGH METHOD OF LEAST SQUARES

If we are given with n points as $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ and we have to fit a curve of given degree over these n points *i.e.* we have to obtain the equation of curve of given degree which passes through these n points, then it is not possible to obtain the equation of the curve which passes through all these points. But, we can find the equation of the curve which passes through a large numer of points given *i.e.* large number of the given points satisfy the equation of the curve approximately.

For fitting a straight line of the form $y = a + bx$,

normal equations of method of least squares are given as:

$$\sum y = a\sum(1) + b\sum x \quad \text{i.e.} \quad \sum y = an + b\sum x$$

$$\text{and} \quad \sum xy = a\sum x + b\sum x^2$$

For fitting a straight line of the form $x = a + by$,

normal equations of method of least squares are given as:

$$\sum x = a\sum(1) + b\sum y \quad \text{i.e.} \quad \sum x = an + b\sum y$$

$$\text{and} \quad \sum xy = a\sum y + b\sum y^2$$

For fitting a parabola of the form $y = a + bx + cx^2$,

normal equations of method of least squares are given as:

$$\sum y = a\sum(1) + b\sum x + c\sum x^2 \quad \text{i.e.} \quad \sum y = an + b\sum x + c\sum x^2,$$

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

$$\text{and} \quad \sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4$$

Q.1. *Fit a straight line to the following data treating y as the dependent variable :*

x :	1	2	3	4	5
y :	5	7	9	10	11

Sol. Let the straight line to be fitted be

$$y = a + bx$$

By the method of least squares, the two normal equations are :

$$\Sigma y = na + b\Sigma x$$

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

Here, $n = 5$

x	y	xy	x^2
1	5	5	1
2	7	14	4
3	9	27	9
4	10	40	16
5	11	55	25
$\sum x = 15$	$\sum y = 42$	$\sum xy = 141$	$\sum x^2 = 55$

Substituting these values in the normal equations, we get

$$42 = 5a + 15b$$

$$141 = 15a + 55b$$

Solving the above two equations, we get

$$a = 3.9, b = 1.5$$

Hence, the equation of the line of best fit is

$$y = 3.9 + 1.5\,x$$

Q.2. *Fit a parabola of second degree to the following data :*

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

Sol. Let the parabola to be fitted be

$$y = a + bx + cx^2$$

By the method of least squares, the three normal equations are :

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

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

$$\Sigma x^2 y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4$$

Here, $n = 5$

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

Substituting these values in the normal equations, we get

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

Solving the above three equations, we get

$$a = 1.42,\ b = -1.07 \text{ and } c = 0.55$$

Hence, the equation of the parabola of best fit is

$$y = 1.42 - 1.07x + 0.55x^2$$

Exercise

1. Obtain the coefficient of correlation for x and y from the following data:

x:	45	55	56	58	60	65	68	70	75	80	85
y:	56	50	48	60	62	64	65	70	74	82	90

2. Obtain the correlation coefficient between x and y, when it is given that:

$n = 40,\ \sum x = 120,\ \sum x^2 = 600,\ \sum y = 90,\ \sum y^2 = 250,\ \sum xy = 356$

3. The marks secured by 9 students in Mathematics and Statistics are given below:

Marks in Mathematics	:	10	15	12	17	13	16	24	14	22
Marks in Statistics	:	30	42	45	46	33	34	40	35	39

Calculate the rank correlation coefficient.

4. Calculate rank correlation coefficient for the following data:

x:	81	78	73	73	69	68	62	58
y:	10	12	18	18	18	22	20	24

5. Find line of regression of y on x for the following data:

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

6. For certain x and y series which are correlated, the two lines of regression are $5x-6y+90=0$ and $15x-8y-130=0$. Find the means of the two series and the coefficient of correlation.

7. Fit a straight line to the following data:

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

8. Fit a second degree parabola to the following data:

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

Answers

1. 0.92

2. 0.81

3. 0.4

4. - 0.9

5. $y = 0.95x + 8.20$

6. $\overline{x} = 30,\ \overline{y} = 40,\ r = 0.667$

7. $y = 0.72 + 1.33x$

8. $y = 1.42 + 0.26x + 2.21x^2$

THANKS