Semester - IV

US04CSTA21

(STATISTICAL TECHNIQUES)

Unit – II

Regression

Introduction:

By using correlation, we can study the relationship between two related variables; also the amount (degree or extent or strength) of correlation can be measured but it fails in giving any idea about the rate at which one variable changes for given change in the other variable and hence one cannot estimate the value of one variable for the given value of the other variable. For estimating the value of one variable for the given value of other variable, we must find out some functional relationship between the variables. Regression is statistical technique with the help of which the functional relationship between two variables can be established and which helps us in estimating the unknown value of one variable for a known value of other variable (Or the method of obtaining mathematical relationship between two related variables is known as regression). When this relationship is given by equation of straight line, regression is said to be linear.

<u>Linear Regression:</u>

For given pairs of observations on X and Y, first draw a scatter diagram. Then draw the line in this diagram which passes through maximum number of points in such a way that almost equal number of points are left on either side of the line. But this method is purely subjective. Different person may draw different lines on the scatter diagram for the same data. Hence we cannot use this method. So, to obtain a best line, we apply the Principle of least squares.

Line of regression is the line which gives the best estimate of one variable for any given value of the other variable.

***Regression line of Y on X*:**

Let $Y = a + bX$ be the equation of a line of regression of $Y\ on\ X$ where $a\ and\ b$ are unknown constants. Let $(Xi, Yi), i = 1, 2 \dots n$ be the given n pairs of observations. So the problem is to estimate the unknowns $a\ and\ b$ which gives the best line fitted for the given data. This can be done by minimizing the error sum of squares parallel to $y - axis$. Thus, according to Principle of least squares, we have to determine $a\ and\ b$ in such a way that error sum of squares is minimum

$i.e.$ $E = \sum_{i=1}^{n} ei^2 = \sum_{i=1}^{n}(Yi - \hat{Y})^2 = \sum_{i=1}^{n}(Yi - (a + bXi))^2 = \sum_{i=1}^{n}(Yi - a - bXi)^2$ **is minimum. From the principle of maxima and minima, differentiate E partially, w.r.to $a\ and\ b$ and equate to zero i.e.**

$$\frac{\partial E}{\partial a} = 0 \Rightarrow 2\sum_{i=1}^{n}(Yi - a - bXi)(-1) = 0$$

$$\Rightarrow (-2)\ \sum_{i=1}^{n}(Yi - a - bXi)(-1) = 0$$

$$\Rightarrow \sum_{i=1}^{n} Yi = na + b\sum_{i=1}^{n} Xi \text{ ------------ } \textbf{(I)}$$

And

$$\frac{\partial E}{\partial b} = 0 \Rightarrow 2\sum_{i=1}^{n}(Yi - a - bXi)(-Xi) = 0$$

$$\Rightarrow -2\sum_{i=1}^{n} Xi(Yi - a - bXi) = 0$$

$$\Rightarrow \sum_{i=1}^{n} XiYi = a\sum_{i=1}^{n} Xi + b\sum_{i=1}^{n} Xi^{2}$$ **------------ (II)**

And

$$\frac{\partial^2 E}{\partial a^2} = 2\sum_{i=1}^{n} 1 = 2n > 0$$

$$\frac{\partial^2 E}{\partial b^2} = 2\sum_{i=1}^{n} Xi^2 > 0$$

Multiplying (I) by ($\sum_{i=1}^{n} Xi$) and (II) by n and solving them, we get

$$\left(\sum_{i=1}^{n} Xi\right)\left(\sum_{i=1}^{n} Yi\right) = na\left(\sum_{i=1}^{n} Xi\right) + b\left(\sum_{i=1}^{n} Xi\right)^2$$

$$n\sum_{i=1}^{n} XiYi = na\left(\sum_{i=1}^{n} Xi\right) + nb\sum_{i=1}^{n} Xi^2$$

$$\therefore \left(\sum_{i=1}^{n} Xi\right)\left(\sum_{i=1}^{n} Yi\right) - n\sum_{i=1}^{n} XiYi = b\left(\left(\sum_{i=1}^{n} Xi\right)^2 - n\sum_{i=1}^{n} Xi^2\right)$$

$$\therefore b = \frac{\left(\sum_{i=1}^{n} Xi\right)\left(\sum_{i=1}^{n} Yi\right) - n\sum_{i=1}^{n} XiYi}{n\sum_{i=1}^{n} Xi^2 - \left(\sum_{i=1}^{n} Xi\right)^2}$$

$$= \frac{n\sum_{i=1}^{n} XiYi - \left(\sum_{i=1}^{n} Xi\right)\left(\sum_{i=1}^{n} Yi\right)}{n\sum_{i=1}^{n} Xi^2 - \left(\sum_{i=1}^{n} Xi\right)^2}$$

Dividing numerator and denominator by n^2, we get

$$\hat{b} = \frac{\frac{1}{n}\sum_{i=1}^{n} XiYi - \left(\overline{X}\right)\left(\overline{Y}\right)}{\frac{1}{n}\sum_{i=1}^{n} Xi^2 - \left(\overline{X}\right)^2}$$

$$= \frac{Cov(X,Y)}{S_X{}^2}$$

$= b_{YX}$ **------------- (III)**

$=$ **Regression coefficient of Y on X**

Also from (I)

$$\sum_{i=1}^{n} Yi = na + b\sum_{i=1}^{n} Xi$$

$\therefore \overline{Y} = a + \hat{b}\overline{X}$

$\therefore \hat{a} = \overline{Y} - \hat{b}\overline{X} = \overline{Y} - b_{YX}\overline{X}$ --------- **(IV)**

Substituting (III) and (IV) in the equation Y=a+bX, we have

$$Y = \overline{Y} - b_{YX} * \overline{X} + b_{YX} * X$$

$$= \overline{Y} + b_{YX}(X - \overline{X})$$

$$\boxed{\therefore Y - \overline{Y} = b_{YX}(X - \overline{X})}$$

The above equation is known as line of regression of $Y\ on\ X$ (or regression equation of$Y\ on\ X$). The regression coefficient of Y on X (b$_{YX}$) gives the rate at which Y changes when X changes by (one) unit amount. Regression line of $Y\ on\ X$ is the line which gives the best estimate of Y for any given value of X.

***<u>Regression line of X on Y</u>*:**

Let $X = a + bY$ be the equation of a line of regression of $X\ on\ Y$ where $a\ and\ b$ are unknown constants. Let $(Xi, Yi), i = 1, 2 \dots n$ be the given n pairs of observations. So the problem is to estimate the unknowns $a\ and\ b$ which gives the best line fitted for the given data. This can be done by minimizing the error sum of squares parallel to $x - axis$.

The regression line of X on Y is

$$\boxed{X - \overline{X} = b_{XY}(Y - \overline{Y})}$$

Where $b_{XY} =$**regression coefficient of X on Y**

$= \dfrac{Cov(X,Y)}{S_Y{}^2}$ **= rate at which X changes when Y changes by (one) unit amount.**

<u>Properties of Regression coefficients</u>:

1. $b_{XY} = r\dfrac{S_X}{S_Y}$ **and** $b_{YX} = r\dfrac{S_Y}{S_X}$ **,** S_X **and** S_Y **are the standard deviations of X and Y respectively, r is the correlation coefficient.**

Proof:

RHS = $r\dfrac{S_X}{S_Y}$ **=** $\dfrac{Cov(X,Y)}{S_X S_Y}\dfrac{S_X}{S_Y}$ **=** $\dfrac{Cov(X,Y)}{S_Y{}^2} = b_{XY}$ **=LHS**

2. The correlation coefficient is the geometric mean of two regression coefficients.

i.e. $r = \pm\sqrt{b_{XY} \times b_{YX}}$ **or** $r^2 = b_{XY} \times b_{YX}$

Proof:

Consider $b_{XY} \times b_{YX} = r\dfrac{S_X}{S_Y} \times r\dfrac{S_Y}{S_X}$ **=** r^2

$$\therefore r = \pm\sqrt{b_{XY} \times b_{YX}}$$

3. The correlation coefficient (r) and the two regression coefficients (b$_{XY}$ and b$_{YX}$) are of same sign. i.e. r, b$_{XY}$ and b$_{YX}$ are of same sign.

Proof:

We know that

$r^2 = b_{XY} \times b_{YX}$ **and** $-1 \le r \le 1$

$\Rightarrow 0 \le r^2 \le 1$

$$\Rightarrow 0 \le b_{XY} \times b_{YX} \le 1$$

This is possible if both b_{XY} and b_{YX} are of same sign. Hence the correlation coefficient and two regression coefficients are of same sign.

4. The arithmetic mean of the regression coefficients is greater than the correlation coefficient.

$$\boldsymbol{i.e.}\ \frac{b_{XY} + b_{YX}}{2} > r$$

Proof: We know that, for any two real distinct positive numbers a and b

A.M > G.M

$$\Rightarrow \frac{a+b}{2} > \sqrt{ab}$$

Taking $a = b_{XY}$ **and** $b = b_{YX}$ **, we get**

$$\frac{b_{XY} + b_{YX}}{2} > \sqrt{b_{XY} \times b_{YX}}$$

$$\Rightarrow \frac{b_{XY} + b_{YX}}{2} > r$$

Hence the result.

5. The regression coefficients are independent of change of origin but depend on scale.

Proof:

Let $(\boldsymbol{Xi}, \boldsymbol{Yi}), \boldsymbol{i = 1, 2 \ldots n}$ **be the given n pairs of observations.**

Then $b_{XY} = r\dfrac{S_X}{S_Y}$ **and** $b_{YX} = r\dfrac{S_Y}{S_X}$

= $\dfrac{Cov(X,Y)}{{S_Y}^2}$ $= \dfrac{Cov(X,Y)}{{S_X}^2}$

Let us transform Xi to Ui and Yi to Vi by the means of change of origin and scale

i.e. $Ui = \dfrac{Xi - A}{c}$ **and** $Vi = \dfrac{Yi - B}{h}$**,** $\boldsymbol{i = 1, 2, \ldots, n}$**.**

Where A and B are new origins, C **and** h **are new scales.**

Thus, we have

$Xi = A + cUi$ **and** $Yi = B + hVi$**, for** $\boldsymbol{i = 1, 2 \ldots n}$**.**

$\therefore \overline{X} = A + c\overline{U}$ **and** $\overline{Y} = B + h\overline{V}$

$\therefore \left(Xi - \overline{X}\right) = c\left(Ui - \overline{U}\right)$ **and** $\left(Yi - \overline{Y}\right) = h\left(Vi - \overline{V}\right)$ **,for** $\boldsymbol{i = 1, 2 \ldots n}$**.**

Now, the regression coefficient of X on Y is

$$b_{XY} = r\frac{S_X}{S_Y}$$

= $\dfrac{Cov(X,Y)}{{S_Y}^2}$ **------ (I)**

Where Cov(X, Y) = $\dfrac{1}{\text{n}}\sum_{i=1}^{n}\left(Xi - \overline{X}\right)\left(Yi - \overline{Y}\right)$

= $\dfrac{1}{n}\sum_{i=1}^{n}\left(c\left(Ui - \overline{U}\right)h\left(Vi - \overline{V}\right)\right)$

$$= ch\frac{1}{n}\sum_{i=1}^{n}\left(Ui-\overline{U}\right)\left(Vi-\overline{V}\right)$$

$$= chb_{UV}$$ **---- (II)**

i.e. Covariance is an independent of change of origin but depend on scale.

Also

Sx² = $\frac{1}{\mathrm{n}}\sum_{i=1}^{n}\left(Xi-\overline{X}\right)^2$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(c\left(Ui-\overline{U}\right)\right)^2$$

$$= c^2\frac{1}{n}\sum_{i=1}^{n}\left(Ui-\overline{U}\right)^2$$

$$= c^2 S_U{}^2$$ **----- (III)**

Similarly,

Sy² = $\frac{1}{\mathrm{n}}\sum_{i=1}^{n}\left(Yi-\overline{Y}\right)^2$

$$= h^2 S_V{}^2$$ **------ (IV)**

$\therefore b_{XY} = \frac{chCov(U,V)}{h^2 S_V{}^2}$ **=** $\frac{c}{h}b_{UV}$ **------- (V)**

i.e regression coefficient is an independent of change of origin but depend on scale.

Similarly, $b_{YX} = \frac{chCov(U,V)}{c^2 S_U{}^2}$ **=** $\frac{h}{c}b_{VU}$ **---- (VI)**

6. Two regression lines X on Y and Y on X intersects at the point ($\overline{X},\overline{Y}$ **).**

7. If one of the regression coefficients is greater than 1 then the other must be less than 1.

Proof:

We know that

$$-1 \le r \le 1$$

$$\therefore 0 \le r^2 \le 1$$

Also $r = \pm\sqrt{b_{XY}\times b_{YX}}\ \therefore r^2 = b_{XY}\times b_{YX}$

$\therefore 0 \le b_{XY}\times b_{YX} \le 1$

Let $b_{XY} > 1, b_{YX} > 1$

i.e. both the regression coefficients are greater than 1 then $r^2 = b_{XY}\times b_{YX} > 1$ **which is impossible. Hence if one of the regression coefficients is greater than 1 then the other must be less than 1.**

8. Angle between the two regression lines:

If θ is an acute angle between two lines of regression then

$\tan\theta = \frac{1-r^2}{|r|}\cdot\frac{S_X S_Y}{S_X{}^2+S_Y{}^2}$ **, where S$_X$ and S$_Y$ are the standard deviations of $X and\ Y$ respectively, r is the correlation coefficient between X and Y.**

Proof:

The regression line of $X\ on\ Y$ is

$$X-\overline{X}=b_{XY}(Y-\overline{Y})$$

$$\Rightarrow \left(Y-\overline{Y}\right)=\frac{1}{b_{XY}}\left(X-\overline{X}\right)$$

$\therefore m_1=\frac{1}{b_{YX}}=\frac{1}{r\frac{S_X}{S_Y}}$ **=** $\frac{S_Y}{rS_X}$ **--------- (I)**

The regression line of $Y \text{ on } X$ is

$$Y-\overline{Y}=b_{YX}(X-\overline{X})$$

$\therefore m_2=b_{YX}=r\frac{S_Y}{S_X}$ **--------- (II)**

Now if θ is an acute angle between two lines of regression then

$\tan\theta=\left|\frac{m_1-m_2}{1+m_1m_2}\right|$ **=** $\left|\frac{\frac{S_Y}{rS_X}-r\frac{S_Y}{S_X}}{1+\frac{S_Y}{rS_X}\cdot\frac{rS_Y}{S_X}}\right|=\left|\frac{\frac{S_Y}{S_X}\left(\frac{1}{r}-1\right)}{1+\frac{S_Y{}^2}{S_X{}^2}}\right|=\frac{\left|1-r^2\right|}{|r|}\cdot\frac{\frac{S_Y}{S_X}}{\left(\frac{S_X{}^2+S_Y{}^2}{S_X{}^2}\right)}$

$$=\frac{\left|1-r^2\right|}{|r|}\cdot\frac{\frac{S_Y}{S_X}}{\left(\frac{S_X{}^2+S_Y{}^2}{S_X{}^2}\right)}\quad=\frac{\left|1-r^2\right|}{|r|}\cdot\frac{S_XS_Y}{\left(S_X{}^2+S_Y{}^2\right)}$$

$$\therefore \theta=\tan^{-1}\left(\frac{\left|1-r^2\right|}{|r|}\cdot\frac{S_XS_Y}{\left(S_X{}^2+S_Y{}^2\right)}\right)$$

In particular, if $r=\pm 1$ **then** $\theta=\tan^{-1}(0)$

$\Rightarrow \theta=0$ **Or** π

i.e the two lines are either coincide $(\theta=0)$ **or they are parallel** $(\theta=\pi)$**,but since both the lines of regression intersect at the point (** $\overline{X},\overline{Y}$ **), they cannot be parallel. Hence in case of perfect correlation, positive or negative, the two lines of regression coincide.**

If $r=0$ **then** $\theta=\tan^{-1}(\infty)=\frac{\pi}{2}$

i.e. if two variables are uncorrelated, the two lines of regression become perpendicular to each other.

Why two lines of regression?

There are always two lines of regression, one of y on X and the other of X on Y. the line of regression of Y on X is used to estimate or predict the value of Y for any given value of X. i.e when Y is a dependent variable and X is an independent variable, is obtained by minimizing error sum of squares parallel to y-axis. On the other hand, the line of regression of X on Y is used to estimate or predict the value of X for any given value of Y i.e when X is a dependent variable and Y is an independent variable, is obtained by minimizing error sum of squares parallel to x-axis. The two regression equations are not reversible or interchangeable because of the simple reason that the basis and assumptions for deriving these equations are quite different.

Problem set:

(1) **In an experiment the number of grams of a given salt which dissolved in 100 gm of water was observed at eight different temperatures.**

Temp.(^{0}c)	**0**	**10**	**20**	**30**	**40**	**50**	**60**	**70**
Weight of salt	**51.5**	**61.5**	**67.2**	**72.6**	**73.5**	**82.2**	**83.5**	**88.0**

Find the regression equation which could be used to predict the weight of salt given the temperature. Predict the weight of salt which would dissolve at temperatures (i) 35^{0}c (ii) 45^{0}c.

(2) **Following table gives age and vital capacity for each of 12 workers in the cadmium industry. Estimate the vital capacity of a worker whose age is 64.**

Age	**39**	**40**	**41**	**41**	**45**	**49**	**52**	**47**	**61**	**65**	**58**	**59**
Vital Capacity	**4.62**	**5.29**	**5.52**	**3.71**	**4.02**	**5.09**	**2.70**	**4.31**	**2.70**	**3.03**	**2.73**	**3.67**

(3) **A panel of judges A and B graded seven debators and independently awarded the following marks:**

Debator	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Marks by A	**40**	**34**	**28**	**30**	**44**	**38**	**31**
Marks by B	**32**	**39**	**26**	**30**	**38**	**34**	**28**

An eighth debator was awarded 36 marks by judge A while judge b was not present. If judge B were also present, how many marks would you expect him to award to the eigth debator assuming that the same degree of relationship exists in their judgments?

(4) **Regression equations of two variables X and Y are as follows:**

$$3X + 2Y - 26 = 0$$
$$6X + Y - 31 = 0$$

Find

(a)*the mean of* X *and* Y (b)*the regression coefficient of* X *on* Y
(c)*The correlation coeff. between* X *and* Y (d)*the most probable value of* Y *when* $X = 5$.
(e) σ_Y **if** $\sigma_X = 5$

(5) **For 50 students of a class the regression equation of marks in Statistics (X) on marks in Mathematics (Y) is $3Y - 5X + 180 = 0$. The mean mark in Mathematics is 44 and the variance of marks in Statistics is $9/16th$ of the variance of marks in Mathematics. Find the mean marks in Statistics and coefficient of correlation between marks in two subjects.**

(6) **The regression equation of profits (X) on sales (Y) of a certain firm is$3Y - 5X + 108 = 0$. The average sales of the firm were Rs. 44,000 and the variance of profits is $9/16th$ of the variance of sales. Find the average profit and the coefficient of correlation between sale and profit.**

(7) **An instructor wants to show the students that there is a linear correlation between the number of hours they spent watching TV(X) during a certain weekend and their scores (Y) on a test taken the following Monday. The number of television viewing hours and the test scores for 12 randomly selected students are shown.**

(a) Obtain the regression line of Y on X

(b) Predict the score of a student who spent 4 hours to watch TV.

X	**0**	**1**	**2**	**3**	**3**	**5**	**5**	**5**	**6**	**7**	**7**	**10**
Y	**96**	**85**	**82**	**74**	**95**	**68**	**76**	**84**	**58**	**65**	**75**	**50**

(8) **An experiment was conducted to determine the mass (Y grams) of a given amount of chemical that dissolved in glycerin at (X^{0}C). the results of the experiment are given below:**

Temperature(X^{0}C)	**0**	**10**	**20**	**30**	**40**	**50**
Mass(Y grams)	**51.3**	**51.4**	**51.9**	**52.0**	**52.6**	**52.8**

(a) Identify an independent and dependent variable.

(b) Compute r, the correlation coefficient and comment on it.

(c) Predict the mass of chemical that dissolved in glycerin at 25^0C.

(9) **Hardik's parents recorded his height at various ages up to 66 months.**

Age(months)	36	48	54	60	66
Height(inches)	35	38	41	43	45

(a) Compute r, the correlation coefficient and comment on it.

(b) Predict the height at the age of 6 years.

(10) **The success of a shopping centre can be represented as a function of the distance (in kms) from the centre of the population and the number of clients (in hundreds of people) who will visit. The data given in the table below:**

No. of Customer (X)	8	7	6	4	2	1
Distance (Y)	15	19	25	23	34	40

To receive 500 customers, at what distance from the centre of the population should the shopping centre be located?