REGRESSION

Regression is the estimation or prediction of unknown values of one variable from known values of another variable. After establishing the fact of correlation between two variables, it is natural curiosity to know the extent to which one variable varies in response to a given variation in the other variable i.e, one is interested to know the nature of relationship between the two variables. **Regression measures the nature and extent of correlation.**

The **line of regression of y on x** gives the best estimate of y for any given value of x.

Its equation is $\boldsymbol{y - \bar{y} = \frac{r\sigma_y}{\sigma_x}(x - \bar{x})}$, which passes through pt. $(\bar{x}, \bar{y})$ and its slope is $b_{yx} = \frac{r\sigma_y}{\sigma_x}$

The **line of regression of x on y** gives the best estimate of x for any given value of y.

Its equation is $\boldsymbol{x - \bar{x} = \frac{r\sigma_x}{\sigma_y}(y - \bar{y})}$, which passes through pt. $(\bar{x}, \bar{y})$ and its slope is reciprocal of $b_{xy} = \frac{r\sigma_x}{\sigma_y}$

Note: (i) $b_{xy} = \frac{r\sigma_x}{\sigma_y} = \frac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma y^2 - n\bar{y}^2} = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma y^2 - (\Sigma y)^2} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(y-\bar{y})^2}$ is called the regression co – efficient of x on y

(ii) $b_{yx} = \frac{r\sigma_y}{\sigma_x} = \frac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma x^2 - n\bar{x}^2} = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$ is called the regression co – efficient of y on x

Alternative Method:

Instead of calculating $\bar{x}, \bar{y}, \sigma_x, \sigma_y$ and r, we may use the following method (as discussed in curve fitting),

Find sum, Σx, Σy, Σxy, Σx^2 and Solve the **Normal equations** $\Sigma y = aN + b\Sigma x$ and $\Sigma xy = a\Sigma x + b\Sigma x^2$ simultaneously for a and b. we get the required equation $y = a + bx$.

PROPERTIES OF REGRESSION CO – EFFICIENTS:

1. Correlation co – efficient is the geometric mean between the regression coefficients. i.e. $r = \sqrt{b_{yx} \times b_{xy}}$

2. If one of the regression co – efficient is greater than unity, the other must be less than unity.

3. Arithmetic mean of regression co – efficient is greater than the correlation co – efficient. **i.e.** $\frac{b_{yx} + b_{xy}}{2} > r$

4. Regression co – efficient are independent of the origin but not of scale.

i.e. If $u = \frac{x-a}{h}$, $v = \frac{y-b}{k}$ where a, b, h and k are constant then $b_{yx} = \frac{k}{h} b_{vu}$ and $b_{xy} = \frac{h}{k} b_{uv}$

Thus, b_{yx} and b_{xy} are both independent of a and b but not of h and k.

5. The correlation co – efficient and the two regression co – efficient have same sign

i.e. b_{yx} , b_{xy} and r have same sign.

6. If θ is the acute angle between the two regression lines then $\tan\theta = \frac{1-r^2}{r} \cdot \frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}$

when r = 0, $\theta = \frac{\pi}{2}$ $\therefore$ The two lines of regression are perpendicular.

Hence the estimated value of Y is the same for all values of X and vice versa.

when $\boldsymbol{r = \pm 1}$, $\tan\theta = 0$ so that $\theta = 0 \ \ or \ \pi$

Hence the lines of regression coincide and there is perfect correlation between the two variates X and Y.

EXERCISE

1. The following table gives the age of car of a certain make and annual maintenance cost. Obtain the equation of the line of regression of cost on age.

Age of a car : 2 4 6 8

Maintenance : 1 2 2.5 3

2. Obtain the equation of the line of regression of y on x from the following data and estimate y for $x = 73$.

x : 70 72 74 76 78 80

y : 163 170 179 188 196 220

3. The heights in cm of fathers (x) and of the eldest sons (y) are given below.

x : 165 160 170 163 173 158 178 168 173 170 175 180

y : 173 168 173 165 175 168 173 165 180 170 173 178

Estimate the height of the eldest son if the height of the father is 172 cm. And the height of the father if the height of the eldest son is 173 cm. Also find the coefficient of correlation between the heights of fathers and sons.

4. Find **(i)**The lines of regression, **(ii)** Coefficient of correlation for the following data.

x : 65 66 67 67 68 69 70 72

y : 67 68 65 66 72 72 69 71

5. From the following data, find the regression equation of X and Y and the coefficient of correlation.

$\sum x = 60,\ \sum y = 40,\ \sum xy = 1500,\ \sum x^2 = 4160,\ \sum y^2 = 1720,\ N = 10$

Where x, y denoted the actual values.

6. Calculate the two regression coefficients and the coefficient of correlation from the following data.

$N = 10,\ \sum x = 350,\ \sum y = 310,\ \sum (x-35)^2 = 162, \sum (y-31)^2 = 222, \sum (x-35)(y-31) = 92$

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

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

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

8.

Given	x series	y series	
Mean	18	100	
S. D.	14	20	$r = 0.8.$

Find the most probable value of y when x =70 and most probable value of x when $y = 90$.

9. Given the following information about marks of 60 students.

	Mathematics	English	
Mean	80	50	
S. D.	15	10	Coefficient of correlation $r = 0.4$

Estimate the marks of the student in mathematics who scored 60 marks in English.

10. Given the following results of weights X and heights Y of 1000 men

$\overline{x} = 150$ lbs. $\sigma_x = 20$ lbs.

$\overline{y} = 68$ inches, $\sigma_y = 2.5$ inches, $r = 0.6$.

Where $\overline{x}$ *and* $\overline{y}$ are means of X and Y, σ_x and σ_y are standard deviations of X and Y and r is the correlation coefficient between X and Y. John weight 200lbs, Smith is 5 feet tall. Estimate the height of john and weight of Smith. From the value of height of John estimate his weight. Why is it different from 200?

11. Out of the two equations given below which can be a line of regression of x on y and why?

$x + 2y - 6 = 0$ and $2x + 3y - 8 = 0$.

12. It is given that the means of x and y are 5 and 10. If the line of regression of y on x is parallel to the line $20y = 9x + 40$, estimate the value of y for $x = 30$

13. In a partially destroyed laboratory record of analysis of correlation data the following results are legible.

Variance $= 9$, equations of the lines of regression $4x - 5y + 33 = 0, 20x - 9y - 107 = 0$.

Find **(i)**The mean values of x and y, **(ii)**The standard deviation of y, **(iii)** Coefficient of correlation between x and y.

14. Find the angle between the lines of regression using the following data.

$n = 10, \sum x = 270, \sum y = 630, \sigma_x = 4, \sigma_y = 5, r_{xy} = 0.6$.

15. If the tangent of the angle made by the line of regression of y on x is 0.6 and $\sigma_y = 2\sigma_x$, find the correlation coefficient between x and y .

16. If the tangent of the angle between the two lines of regression is 0.6 and $\sigma_y = 2\sigma_x$, find the coefficient of correlation between x and y.

17. If $\sigma_x = \sigma_y = \sigma$ and the angle between the lines of regression is $tan^{-1} 3$, find the coefficient of correlation.

18. For a certain bivariate data $b_{xy} = -3$, $r^2 = 0.25$, find the value of b_{yx}

ANSWERS

1. $x = 0.325y + 0.5$ **2.** $y = 5.31x - 212.57;\ y = 175.37$

3. **(i)** $y = 1.016x - 5.078$, **(ii)** $x = 0.477y + 90.88$,

(iii) $172.97,\ 170.69$, **(iv)** $r = 0.696$

4. **(i)** $y = 19.64 + 0.72x$, **(ii)** $x = 33.29 + 0.5y$; $r = 0.604$

5. $x = 0.58y + 3.68;\ r = 0.37$ **6.** $b_{yx} = 0.56,\ b_{xy} = 0.41,\ r = 0.479$

7. $r = 0.6, y = 0.1x - 82$ **8.** $y = 159.3,\ x = 12.4$

9. 86 **10.** Height of john $= 71.75$ inches, Weight of smith $= 111.6$ lbs

11. $2x + 3y - 8 = 0$ **12.** $21.25, y = 0.45x + 7.75$

13. **(i)** $\overline{x} = 13,\ \overline{y} = 17$ **(ii)** $\sigma_y = 4$ **(iii)** $r = 0.6$

14. 0.52 **15**. $r = 0.3$ **16**. $r = \frac{1}{2}$ **17.** -0.17

18. -0.08