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.

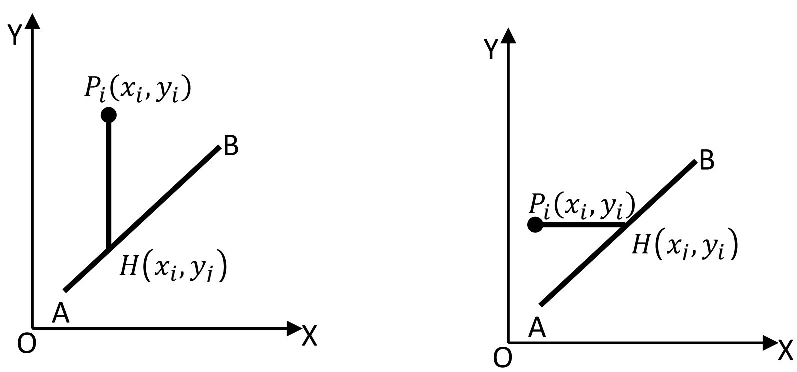
LINEAR REGRASSION:

If two variates x and y are correlated i.e, there exists an association or relationship between them, then the scatter diagram will be more or less concentrated round a curve, This curve is called the curve of regression and the relationship is said to be expressed by means of **curvilinear regression.** In the particular case, when the curve is a straight line, it is called a line of regression and the regression is said to be linear.

A line of regression is the straight line which gives the best fit in the least square sense to the given frequency.

If the line of regression is so chosen that the sum of squares of deviation parallel to the axis of y is minimised. It is called the **line of regression of y on x** and it gives the best estimate of y for any given value of x.

If the line of regression is so chosen that the sum of squares of deviation parallel to the axis of x is minimised. It is called the **line of regression of x on y** and it gives the best estimate of x for any given value of y.



Its equation is $\boldsymbol{y - \bar{y} = \frac{r\sigma_y}{\sigma_x}(x - \bar{x})}$, the line of regression of y on x

Similarly the equation of the line of regression of x on y is $\boldsymbol{x - \bar{x} = \frac{r\sigma_x}{\sigma_y}(y - \bar{y})}$

Alternative Method:

Instead of calculating $\bar{x}, \bar{y}, \sigma_x, \sigma_y$ and r, we may use the following method,

Find sum, Σx, Σy, Σxy, Σx^2 and

Solve the equation $\Sigma y = aN + b\Sigma x$ and $\Sigma xy = a\Sigma x + b\Sigma x^2$ simultaneously for a and b

we get the required equations $y = a + bx$. The above equations are called **Normal equations**

Note: (i) $\frac{r\sigma_y}{\sigma_x}$ is called the regression co – efficient of y on x and is denoted by $b_{yx.}$

(ii) $\frac{r\sigma_x}{\sigma_y}$ is called the regression co – efficient of x on y and is denoted by $b_{xy.}$

(iii) If r = 0, the two lines of regression becomes $y = \bar{y}$ and $x = \bar{x}$ which are two straight lines parallel to X and Y axes respectively and passing through their means $\bar{y}$ and $\bar{x}$. They are mutually perpendicular.

(iv) If $r = \pm 1$, the two lines of regression will coincide.

(v) $b_{xy} = \frac{r\sigma_x}{\sigma_y} = \frac{\frac{1}{n}\Sigma xy - \bar{x}\bar{y}}{\sigma_x \sigma_y} . \frac{\sigma_x}{\sigma_y} = \frac{\frac{1}{n}\Sigma xy - \bar{x}\bar{y}}{\sigma_y^2} = \frac{\frac{1}{n}\Sigma xy - \bar{x}\bar{y}}{\frac{1}{n}\Sigma y^2 - \bar{y}^2} = \frac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma y^2 - n\bar{y}^2}$ **OR** $b_{xy} = \frac{\sum(\boldsymbol{x}-\bar{\boldsymbol{x}})(\boldsymbol{y}-\bar{\boldsymbol{y}})}{\sum(\boldsymbol{y}-\bar{\boldsymbol{y}})^2}$

(vi) $b_{yx} = \frac{r\sigma_y}{\sigma_x} = \frac{\frac{1}{n}\Sigma xy - \bar{x}\bar{y}}{\frac{1}{n}\Sigma x^2 - \bar{x}^2} = \frac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma x^2 - n\bar{x}^2}$ **OR** $b_{yx} = \frac{\sum(\boldsymbol{x}-\bar{\boldsymbol{x}})(\boldsymbol{y}-\bar{\boldsymbol{y}})}{\sum(\boldsymbol{x}-\bar{\boldsymbol{x}})^2}$

PROPERTIES OF REGRESSION CO – EFFICIENTS:

Property I. Correlation co – efficient is the geometric mean between the regression coefficients.

proof: The co –efficient of regression are $\frac{r\sigma_y}{\sigma_x}$ and $\frac{r\sigma_x}{\sigma_y}$.

G.M. between them $= \sqrt{\frac{r\sigma_y}{\sigma_x} \times \frac{r\sigma_x}{\sigma_y}} = \sqrt{r^2} = r =$ co – efficient of correlation

$\therefore r = \sqrt{b_{yx} \times \ b_{xy}}$

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

proof: Let $b_{yx} > 1, then\ \frac{1}{b_{yx}} < 1$

Since $b_{yx.}b_{xy} = r^2 \leq 1 \quad (\because -1 \leq r \leq 1)$

$\therefore b_{xy} \leq \frac{1}{b_{yx}} < 1$ Similarly, if $b_{yx} > 1$, then $b_{xy} < 1$

Property III. Arithmetic mean of regression co – efficient is greater than the correlation co – efficient.

Proof: We have prove that $\frac{b_{yx}+b_{xy}}{2} > r$ **or** $\frac{\frac{r\sigma_y}{\sigma_x}+\frac{r\sigma_x}{\sigma_y}}{2} > r$

or $\sigma_y^2 + \sigma_x^2 > 2\sigma_x\sigma_y$ **or** $\left(\sigma_x - \sigma_y\right)^2 > 0$ which is true.

Property IV: Regression co – efficient are independent of the origin but not of scale.

Proof: Let $u = \frac{x-a}{h}$, $v = \frac{y-b}{k}$ where a, b, h and k are constant

$\therefore r_{uv} = r_{xy};\ \ \sigma_x = h\sigma_u;\ \ \sigma_y = k\sigma_v$

$b_{yx} = \frac{r\sigma_y}{\sigma_x} = r.\ \frac{k\sigma_v}{h\sigma_u} = \frac{k}{h}\left(\frac{r\sigma_v}{\sigma_u}\right) = \frac{k}{h}b_{vu}$

Similarly $b_{xy} = \frac{h}{k}b_{uv}$.

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

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

Proof: Regression co – efficient of y on $x = b_{yx} = r\frac{\sigma_y}{\sigma_x}$

Regression co – efficient of x on y $= b_{xy} = r\frac{\sigma_x}{\sigma_y}$

Since $\sigma_x and\ \sigma_y$ are both positive $\therefore b_{yx}$, b_{xy} and r have same sign.

ANGLE BETWEEN TWO LINES OF REGRESSION:

If θ is the acute angle between the two regression lines in the case of two variables X and Y, then $\tan\theta = \frac{1-r^2}{r}.\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}$ where r, $\sigma_x\sigma_y$ have their usual meaning. Explain the significance of the formula when $r\ = 0$ and $r\ = \pm 1$.

SOME SOLVES EXAMPLES:

1. A chemical engineer is investigating the effect of process operating temperature X on product yield Y. The study results in the following data.

X	100	110	120	130	140	150	160	170	180	190
Y	45	51	54	61	66	70	74	78	85	89

Find the equation of the least square line which will enable to predict yield on the basis of temperature. Find also the degree of relationship between the temperature and the yield. Also verify that the sum of the coefficients of regression is greater than $2r$

Solution:

Sr. no.	u			v			uv
	X	$X-150$	u^2	Y	$Y-70$	v^2	
1	100	-50	25	45	-25	625	1250
2	110	-40	16	51	-19	361	760
3	120	-30	9	54	-16	256	480
4	130	-20	4	61	-9	81	180
5	140	-10	1	66	-4	16	40
6	150	00	0	70	0	0	00
7	160	10	1	74	4	16	40
8	170	20	4	78	8	64	160
9	180	30	9	85	15	225	450
10	190	40	16	89	19	361	760
$N = 10$		-50	85		-27	2005	4120

Calculations of b_{xy}, b_{yx} etc

$\bar{X} = 145$

$\bar{Y} = 67.3$

$\bar{u} = -5\ ,\ \bar{v} = -2.7$

$b_{yx} = b_{vu} = \frac{\Sigma uv - n\bar{u}\bar{v}}{\Sigma u^2 - n\bar{u}^2} = \frac{4120-10(-5)(-2.7)}{85-10(-5)^2} = 0.483$

$b_{xy} = b_{uv} = \frac{\Sigma uv - n\bar{u}\bar{v}}{\Sigma v^2 - n\bar{v}^2} = \frac{4120-10(-5)(-2.7)}{2005-10(-2.7)^2} = 2.06$

The line of regression of Y on X is $Y - \bar{Y} = b_{yx}(X - \bar{X})$

$\therefore Y - 67.3 = 0.483(X - 145)$

$\therefore Y = 0.483X - 2.735$

The coefficient of correlation $r = \sqrt{b_{yx} \times b_{xy}} = \sqrt{0.483 \times 2.06} = 0.9975$

Now, $b_{xy} + b_{yx} = 2.060 + 0.483 = 2.543$

And $2r = 2 \times 0.9975 = 1.995$

Hence, we see that $b_{yx} + b_{xy} > 2r$

2. A panel of two judges A and B graded dramatic performances by independently awarding marks as follow

Performance No.	1	2	3	4	5	6	7
Marks by A	36	32	34	31	31	32	35
Marks by B	35	33	31	30	34	32	36

The eighth performance, however, which judge B could not attend, got 38 marks by judge A. If judge B had also been present, how many marks would he be expected to have awarded to the eighteen performance?

Solution: We have to find the marks that would have been awarded by the judge B. Therefore, let the marks given by the judge B, be denoted by Y and those given by A by X

For Calculation of coefficient of correlation

$\therefore \bar{X} = \frac{\sum X}{N} = \frac{231}{7} = 33, \bar{Y} = \frac{\sum Y}{N} = \frac{231}{7} = 33$

Now $r = \frac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}} = \frac{16}{\sqrt{24 \times 28}} = \frac{16}{\sqrt{672}} = \frac{16}{25.92} = 0.62$

Calculation of coefficient of correlation etc

Sr no	$X - \bar{X}$			$Y - \bar{Y}$			Product
	X	x	x^2	Y	y	y^2	xy
1	36	3	9	35	2	4	6
2	32	-1	1	33	0	0	0
3	34	1	1	31	-2	4	-2
4	31	-2	4	30	-3	9	6
5	31	-2	4	34	1	1	-1
6	32	-1	1	32	-1	1	1
7	35	2	4	36	3	9	3
$N = 7$	$\sum X = 231$		$\sum x^2 = 24$	$\sum Y = 231$		$\sum y^2 = 28$	$\sum xy = 13$

$\sigma_x = \sqrt{\frac{\sum x^2}{N}} = \sqrt{\frac{24}{7}} = 1.85$ $\quad\quad$ $\sigma_y = \sqrt{\frac{\sum y^2}{N}} = \sqrt{\frac{28}{7}} = \sqrt{4} = 2$

The equation of the line of regression of Y on X is $Y - \bar{Y} = r\frac{\sigma_y}{\sigma_x}(X - \bar{X})$

$\therefore Y - 33 = 0.62 \times \frac{2}{1.85}(X - 33) = 0.67(X - 33)$

To find the value of Y when $X = 38$, put $X = 38$ in the above equation

$\therefore Y - 33 = 0.67(38 - 33) = 0.67 \times 5 = 3.35$

$\therefore Y = 33 + 3.35 = 36.5 = 37$ approximately

Therefore, the judge B would have given 37 marks to the eighth performance

3. 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 and weight

Solution: The coefficients of regression are given by

$$b_{yx} = \frac{\sum xy - \frac{\sum x \cdot \sum y}{N}}{\sum x^2 - \frac{(\sum x)^2}{N}} = \frac{1022250 - \frac{15000 \times 6800}{100}}{2272500 - \frac{(15000)^2}{100}} = \frac{2250}{22500} = 0.1$$

$$b_{xy} = \frac{\sum xy - \frac{\sum x \cdot \sum y}{N}}{\sum y^2 - \frac{(\sum y)^2}{N}} = \frac{1022250 - \frac{15000 \times 6800}{100}}{463025 - \frac{(6800)^2}{100}} = \frac{2250}{625} = 3.6$$

$$\therefore r = \sqrt{b_{yx} \times b_{xy}} = \sqrt{0.1 \times 3.6} = 0.6$$

The equation of the lines of regression of y on x is $y - \bar{y} = b_{yx}(x - \bar{x})$

$$y - 68 = 0.1(x - 1500)$$

$$\therefore y = 0.1x - 82$$

4. From 10 observations on price X and supply Y of a commodity the following summary figures were obtained $\sum X = 130, \sum Y = 220, \sum X^2 = 2288, \sum XY = 3467$. Compute the equation of the line of regression of Y on X and interpret the result. Estimate the supply when price is 16 units

Solution: We obtain the values of a and b of the equation of the line of regression of Y on X i.e. of the equation $Y = a + bX$ from the normal equations

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

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

But $N = 10, \sum X = 130, \sum Y = 220, \sum X^2 = 2288, \sum XY = 3467$

$\therefore 220 = 10a + 130b \qquad \therefore 3467 = 130a + 2288b$

Multiply the first equation by 13 and subtract the result from the second equation

$\therefore 607 = 598b \qquad \therefore b = 1.002$

Substituting this value of b in any of the two equations we get, $a = 8.974$

Hence, the equation of the line of regression is, $Y = 8.974 + 1.002X$

When $X = 16$, putting this value, $Y = 8.974 + 1.002 \times 16 = 25.006$

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

$\bar{x} = 150$ lbs. $\qquad \sigma_x = 20$ lbs.

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

Where $\bar{x}$ *and* $\bar{y}$ 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 ?

Solution: With the given notation the line of regression of Y on X is $Y - \bar{Y} = r\frac{\sigma_y}{\sigma_x}(X - \bar{X})$

Substituting the given values, $Y - 68 = 0.6 \times \frac{2.5}{20}(X - 150) = \frac{15}{200}(X - 150)$

Put $X = 200$

$\therefore Y - 68 = \frac{15}{200}(200 - 150) = \frac{15}{4} = 3.75$

$\therefore Y = 68 + 3.75 = 71.75$ inches

Now the line of regression of X of Y is $X - \bar{X} = r\frac{\sigma_x}{\sigma_y}(Y - \bar{Y})$

Substituting the given values, $X - 150 = 0.6 \times \frac{20}{2.5}(Y - 68) = \frac{24}{5}(Y - 68)$

Put $Y = 5$ feet $= 60$ inches

$\therefore X - 150 = \frac{24}{5}(60 - 68) = -\frac{192}{5}$

$\therefore X = 150 - \frac{192}{5} = 111.6$ lbs

Hence, the height of John $= 71.75$ inches and weight of Smith $= 111.6$ lbs

To estimate the weight of John from his height 71.25 we have to use the equation of line of regression of X on Y (and not of Y on X)

i.e. $X - 150 = \frac{24}{5}(Y - 68)$

Putting $Y = 71.75$, we get $X - 150 = \frac{24}{5}(3.75) = 18$ $\quad\therefore X = 168$

The difference is due to the fact that for estimating Y we use one equation and for estimating X we use another equation.

6. 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 $x = 30$

Solution: The line of regression of y on x is $y - \bar{y} = b_{yx}(x - \bar{x})$

Its slope is b_{yx}. But this line is parallel to $20y = 9x + 40$ i.e. $y = \frac{9}{20}x + 2$ whose slope is $\frac{9}{20}$.

$\therefore b_{yx} = \frac{9}{20}$

But by data $\bar{x} = 5$ and $\bar{y} = 10$. Hence, the equation of the line of regression of y on x is

$y - 10 = \frac{9}{20}(x - 5)$ i.e. $y = \frac{9}{20}x + \frac{155}{20}$

When $x = 30, y = \frac{270}{20} + \frac{155}{20} = \frac{425}{20} = 21.25$

7. Given $6Y = 5X + 90, 15X = 8Y + 130, {\sigma_x}^2 = 16$. Find (i) $\bar{x}$ and $\bar{y}$, (ii) r and (iii) ${\sigma_y}^2$

Solution: **(i)** To find $\bar{x}$ and $\bar{y}$: We solve the given equations simultaneously. Multiply the first equation by 3

$\therefore -15X + 18Y = 270$ and add $15X - 8Y = 130$

$\therefore 10Y = 40$ $\quad\therefore \bar{Y} = 40$

Putting this value in any of the given equations $6 \times 40 = 5X + 90$ $\quad\therefore X = 30$ $\quad\therefore \bar{X} = 30$

(ii) To find r: Suppose the first equation represents, the line of regression of X on Y

Writing it as $X = \frac{6}{5}Y - 18$, we find $b_{xy} = \frac{6}{5}$

Suppose the second equation represents the line of regression of Y on X

Writing it as $Y = \frac{15}{8}X - \frac{130}{8}$, we find $b_{yx} = \frac{15}{8}$

$\therefore r = \sqrt{b_{xy} \times b_{yx}} = \sqrt{\frac{6}{5} \times \frac{15}{8}} = \sqrt{\frac{9}{4}} = \sqrt{2.25} = 1.5$

But the value of r can never be greater than 1 numerically. Hence, our supposition is wrong

Now treating the first equation as representing the line of regression of Y on X, we write it as,

$Y = \frac{5}{6}X + 15$ $\qquad \therefore b_{yx} = \frac{5}{6}$

Treating the second equation as representing the line of regression of X on Y, we write it as,

$X = \frac{8}{15}Y + \frac{130}{158}$ $\qquad \therefore b_{xy} = \frac{8}{15}$

$\therefore\ r = \sqrt{b_{xy} \times b_{yx}} = \sqrt{\frac{8}{15} \times \frac{5}{6}} = \sqrt{\frac{4}{9}} = \frac{2}{3} = 0.667$

(iii) For σ_y $\qquad b_{yx} = r\frac{\sigma_y}{\sigma_x}$ $\qquad \therefore \frac{5}{6} = \frac{2}{3} \times \frac{\sigma_y}{4}$ $\qquad \therefore \sigma_y = 5$

8. The regression lines of a sample are $x + 6y = 6$, and $3x + 2y = 10$. Find (i) Sample means $\bar{x}$ and $\bar{y}$, (ii) coefficient of correlation between x and y. Also estimate y when $x = 12$.

Also verify that the sum of the coefficients of regressions is greater than $2r$.

Solution: **(i)** Mean $\bar{x}$ and $\bar{y}$ are obtained by solving the two given equations.

$$\begin{array}{rl} 3x + 18y = 18 & \qquad \therefore y = 1/2 \\ 3x + 2y = 10 & \qquad \therefore x = 3 \\ \hline 16y = 8 & \end{array}$$

(ii) If the line $x + 6y = 6$ is the line of regression of y on x, then

$6y = -x + 6$ i.e. $y = -\frac{1}{6}x + 1$ $\qquad \therefore\ b_{yx} = -\frac{1}{6}$

If the line $3x + 2y = 10$ is the line of regression of x on y, then

$3x = -2y + 10$ i.e. $x = -\frac{2}{3}y + \frac{10}{3}$ $\qquad \therefore b_{xy} = -\frac{2}{3}$

$\therefore r = \sqrt{b_{yx} \times b_{xy}} = \sqrt{\left(-\frac{1}{6}\right) \times \left(-\frac{2}{3}\right)} = \sqrt{\frac{1}{9}} = \frac{1}{3}$

Since b_{yx} and b_{xy} are negative, r is negative

$\therefore r = -1/3$

Since $b_{yx} + b_{xy} = \frac{1}{6} + \frac{2}{3} = \frac{5}{6}$ (Numerically) and $2r = \frac{2}{3}$, we see that $b_{yx} + b_{xy} > 2r$

(iii) To estimate y when $x = 12$, we use the line of regression of y on x i.e. $y = -\frac{1}{6}x + 1$, when $x = 12, y = -2 + 1 = -1$

9. Find the angle between the lines of regression using the following data $n = 10, \sum x = 270, \sum y = 630$, $\sigma_x = 4,\ \sigma_y = 5,\ r_{x,y} = 0.6$

Solution: The angle between the lines of regression is given by $\tan\theta = \left(\frac{1-r^2}{r}\right)\left(\frac{\sigma_x \cdot \sigma_y}{{\sigma_x}^2+{\sigma_y}^2}\right)$

Putting the given values $\tan\theta = \left[\frac{1-(0.6)^2}{0.6}\right]\left(\frac{4\times 5}{16+25}\right) = 0.52$

10. 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

Solution: If the equation of the line of regression of y on x is $y - \bar{y} = b_{yx}(x - \bar{x})$ then we know that b_{yx} is the slope of the line of regression. We are thus, given $b_{yx} = 0.6$

But $b_{yx} = r\frac{\sigma_y}{\sigma_x}$ and $\sigma_y = 2\sigma_x$

Putting these values $0.6 = r \cdot \frac{2\sigma_x}{\sigma_x} = 2r$ $\quad\therefore r = \frac{0.6}{2} = 0.3$

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

Solution: We know that the tangent of the angle between two lines of regression is given by

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

But $\sigma_y = 2\sigma_x$ and $\tan\theta = 0.6$

$\therefore 0.6 = \left(\frac{1-r^2}{r}\right)\left(\frac{2{\sigma_x}^2}{5{\sigma_x}^2}\right)$ $\quad\therefore \frac{3}{5} = \left(\frac{1-r^2}{r}\right)\frac{2}{5}$

$\therefore 2r^2 + 3r - 2 = 0$ $\quad\therefore (2r-1)(r+2) = 0$

$\therefore r = -2$ or $r = 1/2$ $\quad\therefore r = \frac{1}{2}$ ($|r|$ cannot be > 1)

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

Solution: We have $\tan\theta = \frac{1-r^2}{r}\left(\frac{\sigma_x \cdot \sigma_y}{{\sigma_x}^2+{\sigma_y}^2}\right)$

$\therefore 3 = \frac{1-r^2}{r} \cdot \left(\frac{\sigma^2}{\sigma^2+\sigma^2}\right) = \frac{1-r^2}{2r}$

$\therefore \frac{1-r^2}{r} = 6$ $\quad\therefore r^2 + 6r - 1 = 0$

$\therefore r = \frac{-6\pm\sqrt{36+4}}{2} = -3 \pm \sqrt{10}$

Since r cannot be greater than 1, $r = -3 \pm \sqrt{10} = 0.17$

13. If the arithmetic mean of regression coefficients is p and their difference is $2q$, find the correlation coefficient.

Solution: Let the coefficients of regression be b_1 and b_2. Now by data $\frac{b_1+b_2}{2} = p$ and $b_1 - b_2 = 2q$

$\therefore b_1 + b_2 = 2p$ and $b_1 - b_2 = 2q$

$\therefore b_1 = p + q$ and $b_2 = p - q$ $\quad\therefore$ Coefficient of correlation $= r = \sqrt{b_1 b_2} = \sqrt{p^2 - q^2}$