Equations of Regression

The correlation coefficient tells us if there is some relation between the random variables X and Y. The regression equations express the relation mathematically. Here we obtain ***linear*** relation between the variables.

Regression line of Y on X (***Y is the dependent variable***)

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$

or

$$Y - \overline{Y} = r\frac{\sigma_y}{\sigma_x}\left(X - \overline{X}\right)$$

Here $r\frac{\sigma_y}{\sigma_x}$ which is the slope of the line is denoted by b_{yx} and is called the **regression coefficient** of y on x.

Regression line of X on Y (***X is the dependent variable***)

$$X - \bar{X} = b_{xy}(Y - \bar{Y})$$

$$X - \overline{X} = r\frac{\sigma_x}{\sigma_y}\left(Y - \overline{Y}\right)$$

Here $r\frac{\sigma_x}{\sigma_y}$ which is the slope of the line is denoted by b_{xy} and is called the **regression coefficient** of x on y.

Remarks:

1. The point $\left(\overline{X}, \overline{Y}\right)$ lies on both the lines of regression.

2. We have $b_{yx} * b_{xy} = r^2; \Rightarrow$ both b_{yx} and b_{xy} have the same sign. Also $b_{yx} = r\dfrac{\sigma_y}{\sigma_x} \Rightarrow b_{yx}$ and *r* have the same sign. (Since σ_x *and* σ_y are positive). That is b_{yx}, b_{xy} *and* r ***all*** have the same sign.
3. To ***estimate y***, use the ***regression line of y on x***. Similarly to ***estimate x***, use the ***regression line of x on y.***
4. Angle between regression lines: $\tan\theta = \left(\dfrac{1-r^2}{r}\right)\dfrac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}$

Examples:

1. Obtain the equations of two lines of regression for the following data. Also obtain the estimate of X for Y=70.

X	65	66	67	67	68	69	70	72
Y	67	68	65	68	72	72	69	71

Solution: We have,

$$n = 8;\quad \sum X_i = 544;\ \sum Y_i = 552;\ \sum X_iY_i = 37560$$

$$\sum X_i^2 = 37028;\ \sum Y_i^2 = 38132$$

$$\overline{X} = \frac{1}{n}\sum X_i = \frac{1}{8}(544) = 68$$

$$\overline{Y} = \frac{1}{n}\sum Y_i = \frac{1}{8}(552) = 69$$

$$\sigma_x = \sqrt{\left(\frac{1}{n}\sum X_i^2 - \left(\frac{1}{n}\sum X_i\right)^2\right)} = \sqrt{\left(\frac{1}{8}(37028) - (68)^2\right)} = \sqrt{4.5} = 2.1213$$

$$\sigma_y = \sqrt{\left(\frac{1}{n}\sum Y_i^2 - \left(\frac{1}{n}\sum Y_i\right)^2\right)} = \sqrt{\left(\frac{1}{8}(38132) - (69)^2\right)} = \sqrt{5.5} = 2.3452$$

$\Rightarrow r = 0.603$

The regression equation of Y on X is:

$$Y - \overline{Y} = r\frac{\sigma_y}{\sigma_x}\left(X - \overline{X}\right)$$

i.e $Y = 0.665X + 23.78$

Similarly the regression equation of X on Y is:

$$\mathrm{X} - \overline{X} = r\frac{\sigma_x}{\sigma_y}\left(Y - \overline{Y}\right) \text{ i.e } \mathrm{X} = 0.54Y + 30.74$$

$\therefore$ y = 70$\Rightarrow x = 68 + 0.603\dfrac{2.1213}{2.3452}(70 - 69)$ {using $\mathrm{X} - \overline{X} = r\dfrac{\sigma_x}{\sigma_y}\left(Y - \overline{Y}\right)$}

i.e. $x = 68.5454$

2. Consider the two regression lines: $3x + 2y = 26$ & $6x + y = 31$. (a) Find the mean values and the correlation coefficient between X and Y. (b) If the variance of Y is 4, find the S.D of X.

Solution: We know that the point $\left(\overline{X}, \overline{Y}\right)$ lies on both the lines of regression.

$$\left[\overline{X} = E(X) \; and \; \overline{Y} = E(Y)\right]$$

Solving the regression equations $3x + 2y = 26$ & $6x + y = 31$ we get

- $x = 4,\ y = 7 \Rightarrow \overline{X} = 4, \overline{Y} = 7$

Now let us assume that the regression line of x on y is $3x + 2y = 26$

$$3x + 2y = 26 \Rightarrow x = -\frac{2}{3}y + \frac{26}{3} \Rightarrow slope = b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{-2}{3} \quad .....(1)$$

Similarly, let us assume that the regression line of y on x is $6x+y=31$

Then $6x+y=31 \Rightarrow y=-6x+31 \Rightarrow slope = b_{yx} = r\frac{\sigma_y}{\sigma_x} = -6$ .....(2)

$$\therefore \text{r}^2 = b_{xy} * b_{yx} = 4, \text{ which is not possible, since} -1 \le r \le 1$$

Hence our assumption is wrong. Therefore the regression line of y on x is

$3x+2y=26 \Rightarrow y=\frac{-3}{2}x+13$ and the regression line of x on y is

$6x+y=31 \Rightarrow x=\frac{-1}{6}y+31$

We have,

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{-3}{2} \text{ and } b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{-1}{6} \Rightarrow r^2 = b_{yx} * b_{xy} = \frac{1}{4}$$

$$\Rightarrow r = \frac{-1}{2} \quad (\because b_{yx} \text{ and } b_{xy} \text{ are both negative})$$

- Hence the correlation coefficient $r=\frac{-1}{2}$

Now, $b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{-1}{6}$

$$\Rightarrow \frac{1}{2}\left(\frac{\sigma_x}{2}\right) = \frac{1}{6}$$

$$\Rightarrow \sigma_x = S.D \text{ of } X = \frac{2}{3}$$

Practice Problems

1. The regression lines are $x+6y=6$ & $3x+2y=10$. Find (i) $\overline{x}, \overline{y}$. (ii) r. Also estimate y when x=12.

[(i) $\overline{x}=3, \overline{y}=\frac{1}{2}$. (ii) $r=\frac{-1}{3}$ $(b_{yx}=\frac{-1}{6}; b_{xy}=\frac{-2}{3})$ y=-1, when x=12]

2. 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.

[Slope $b_{yx}=\frac{9}{20}; (y-10)=\frac{9}{20}(x-5); \textit{estimate for } y=21.25$]

3. For the following data,

X	1	2	3	4	5	6	7	8	9
Y	9	8	10	12	11	13	14	16	15

find the lines of regression. Show that for X = 6.2, the estimated value of Y= 13.14. Also estimate the value of X for Y = 13.14. Explain why this value of X differs from 6.2.

[This is because we use two different regression lines: To estimate the value of y, given x=6.2, we use the line of regression of y on x. But to estimate the value of x for y=13.14, we use the line of regression of x on y. Since the two lines are not the same, we get a different value of x.]

Prof. Nancy Sinollin