Regression

Definition: Method of estimating the value of one variable when that of the other is known and when the two variables are correlated is called regression

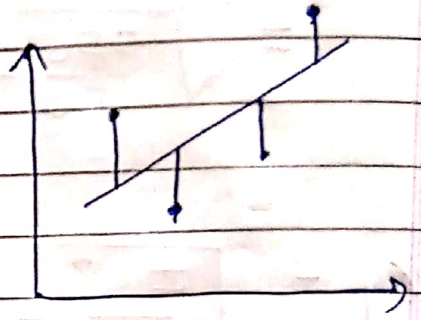
Lines of regression: It is line such that the sum of the distances of the points from the line is minimum

Types: i) Line of regression of y on x

It is given by

$y = a + bx$

'b' is called regression coefficient of y on x

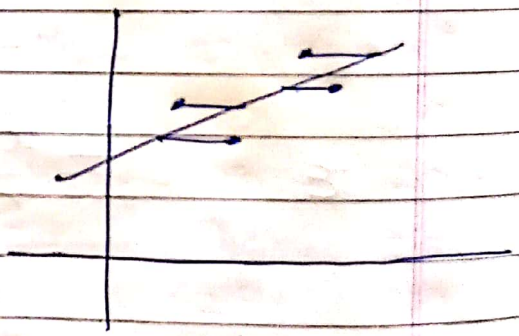


ii) Line of regression of x on y

It is given by

$x = a' + b'y$

b' is called regression coefficient of x on y



* The lines of regression of y on x is given by

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

The line of regression of x on y is given by

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

r is coefficient of correlation
σ_x standard deviation of x
σ_y Standard " of y

Definition

$b_{yx} = r \cdot \frac{\sigma_y}{\sigma_x}$ is called regression coeffficient of y on x &

$b_{xy} = r \cdot \frac{\sigma_x}{\sigma_y}$ is regression coeffficient of x on y

Remark:

$\therefore b_{xy} \cdot b_{yx} = r^2 \Rightarrow r = \sqrt{b_{xy} \cdot b_{yx}}$

Note: $b_{xy} \cdot b_{yx}$ is positive $\Rightarrow$ if one of them is negative other must be negative

Ex.① Find Karl pearson's coefficient of correlation and two lines of regression for the following. Estimate y when $x = 80$

x	62	64	65	69	70	71	72	74
y	126	125	139	145	165	152	180	208

Sol^n. We know that coefficient of correlation

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

$$\sigma_x = \sqrt{\frac{\sum (x-\bar{x})^2}{n}} \qquad \sigma_y = \sqrt{\frac{\sum (y-\bar{y})^2}{n}}$$

x	y	$x-\bar{x}$	$(x-\bar{x})^2$	$y-\bar{y}$	$(y-\bar{y})^2$	$(x-\bar{x})\cdot(y-\bar{y})$
62	126	-6	36	-29	841	174
64	125	-4	16	-30	900	120
65	139	-3	9	-16	256	48
69	145	1	1	-10	100	-10
70	165	2	4	10	100	20
71	152	3	9	-3	9	-9
72	180	4	16	25	625	100
74	208	6	36	53	2809	318
			127		5640	761

$$\bar{x} = \frac{\Sigma x}{n} = 68.37 \simeq 68 \qquad \bar{y} = 155$$

$$\sigma_x = \sqrt{\frac{127}{8}} = 3.984$$

$$\sigma_y = \sqrt{\frac{5640}{8}} = 26.55$$

$$\Rightarrow \varepsilon = \frac{761}{8 \times 26.55 \times 3.98}$$

$$\varepsilon = 0.86$$

The line of regression of y on x

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

$$y - 155 = \frac{0.86 \times 26.55}{3.984} (x - 68)$$

If $x = 80$ $y = 223.7$

The line of regression x on y

$$(x - 68) = 0.86 \times \frac{3.984}{26.55} (y - 155)$$

X. Below is given the respective heights x & y of a sample of 12 fathers and their eldest sons

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

using linear regression estimate son's height if father's height is 172 cm and estimate father's height if son's height is 173. also find coefficient of correlation between the heights of fathers & sons

x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
165	173	-4.4	1.25	19.36	1.65	-5.5
160	168	-9.4	-3.75	83.36	14.06	35.25
170	173	0.6	1.25	0.36	1.56	0.75
163	165	-6.4	-5.75	40.96	33.06	36.8
173	175	3.6	3.25	13.96	10.57	11.7
158	168	-11.4	-3.75	123.96	14.06	42.75
178	173	8.6	1.25	73.96	1.56	10.75
168	165	-1.4	-5.75	1.96	33.06	9.05
173	180	3.6	3.25	13.96	68.06	29.7
170	170	0.6	-1.75	0.36	3.05	-1.05
175	173	5.6	1.25	31.96	1.56	7
180	178	10.6	6.25	113.36	39.06	65.25
2033	2061			517.52	221.2	244.5

$\bar{x} = 169.4 \quad \bar{y} = 171.75$

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{n.\sigma_x \sigma_y} = 0.76$$

Line of regression of y on x is

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

10
18

$$(y - 171.75) = \frac{0.76 \times 4.25}{6.56} (x - 169.4)$$

$$= 0.49 (x - 169.4)$$

$$\Rightarrow y = 0.49x + 88.74$$

If, $x = 172 \Rightarrow y = 173$ estimated height of son.

IIy line of regression of x on y is

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

$$\Rightarrow (x - 169.4) = 1.12 (y - 171.75)$$

If $y = 173 \Rightarrow x = 170.8$ estimated height of father