Lecture 17

REGRESSION LINE OF X ON Y

The line which expresses the trend of two observed values is called a regression line. For example if the sample data is given then the value of $\boldsymbol{y}$ corresponding to the given value of $\boldsymbol{x}$ can be estimated by the method of least squares. Now because the value of $\boldsymbol{y}$ is estimated from given value of $\boldsymbol{x}$ therefore the resulting line is called regression line of $\boldsymbol{y}$ on $\boldsymbol{x}$ which means that $\boldsymbol{y}$ is dependent on $\boldsymbol{x}$. The general equation of $\boldsymbol{y}$ on $\boldsymbol{x}$ is

$$\boldsymbol{Y = a + bx}$$

where $\boldsymbol{b} = \frac{\boldsymbol{n\sum xy - (\sum y)(\sum x)}}{\boldsymbol{n\sum x^2 - (\sum x)^2}}$ and $\mathbf{a} = \overline{\boldsymbol{y}} - \boldsymbol{b}\overline{\boldsymbol{x}} = \frac{\boldsymbol{\sum y}}{\boldsymbol{n}} - \boldsymbol{b}\frac{\boldsymbol{\sum x}}{\boldsymbol{n}}$

Question

Problem: The following table shows the chart of price and demand for an item at different periods of time.

I. Forecast demand for the price of $ 25

- Solution:

Price (x)	*Demand* (y)	xy	x^2	y^2
18	45	810	324	2025
22	40	880	484	1600
25	34	850	625	1156
30	26	780	900	676
18	42	756	324	1764
15	44	660	225	1936
12	45	540	144	2025
140	276	5276	3026	11182

Regression Line of Y on X	
$Y = a + bx$	**And**
Where,	$a = \bar{y} - b\bar{x}$
$b = \dfrac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$	$a = \dfrac{276}{7} - (-1.08)\left(\dfrac{140}{7}\right)$
$b = \dfrac{7(5276) - (140)(276)}{7(3026) - (140)^2}$	$a = 39.43 + 21.60 = 61.03$
$b = \dfrac{36932 - 38640}{21182 - 19600}$	**Therefore,**
	$y = 61.03 - 1.08x$
$b = \dfrac{-1708}{1582}$	*By putting the value of* $x = 25$ *we get*
	$Y = 61.03 - 1.08(25)$
$b = -1.08$	$Y = 61.03 - 27 = 34.03$

Correlation

Correlation measures the degree of interdependence (association) between two variables. If two variables are so related that an increase or decrease of one is found in connection with increase or decrease of the other, then the two variables are said to be correlated. Here it is important to note that there might be a similar movement between two variables such as automobile sales and demand for shoes. But these two variables have no connection due to which the calculation for these two variables is wrong because it does not make any sense. Therefore care must be taken that the two variables have some connection before a calculation can make sense.

Correlation Coefficient

- **The correlation coefficient gives a mathematical value for measuring the strength of the linear relationship between two variables.**
- **r lies between -1 and +1**
- **+1 indicates perfect positive relation**
- **-1 indicates perfect negative relation**
- **0 shows no correlation**

Pearson product moment Correlation Coefficient

- The formula for calculating linear correlation coefficient is called product-moment formula presented by Karl Pearson. Therefore it is also called Pearsonian coefficient of correlation. The formula is given as:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Properties of Coefficient of Correlation

- Coefficient of correlation lies between -1 and +1,i.e. $-\mathbf{1} \leq \boldsymbol{r} \leq +\mathbf{1}$.
- Coefficients of correlation are independent of change of origin and scale.
- Coefficients of correlation possess the property of symmetry. i.e. $\boldsymbol{r}_{\boldsymbol{xy}} = \boldsymbol{r}_{\boldsymbol{yx}}$
- The coefficient of correlation is a geometric mean of two regression coefficient. $\mathbf{r} = \pm\sqrt{\boldsymbol{b}_{\boldsymbol{xy}}.\,\boldsymbol{b}_{\boldsymbol{yx}}}$

  $\mathbf{r} = +\sqrt{\boldsymbol{b}_{\boldsymbol{xy}}.\,\boldsymbol{b}_{\boldsymbol{yx}}}$, if $\boldsymbol{b}_{\boldsymbol{xy}} \boldsymbol{and}\ \boldsymbol{b}_{\boldsymbol{yx}}$ are positive.

  $\mathbf{r} = -\sqrt{\boldsymbol{b}_{\boldsymbol{xy}}.\,\boldsymbol{b}_{\boldsymbol{yx}}}$, if $\boldsymbol{b}_{\boldsymbol{xy}} \boldsymbol{and}\ \boldsymbol{b}_{\boldsymbol{yx}}$ are negative.

- Note: Correlation is the geometric mean of absolute values of two regression coefficients i.e.

$$r = \sqrt{bd}$$

$where,$

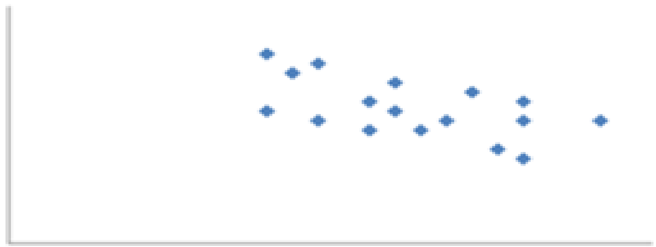
$b = byx \ regression \ coefficient \ of \ Y \ on \ X$

$d = bxy \ regression \ coefficient \ of \ X \ on \ Y$

Scatter Diagrams for different degrees of correlation

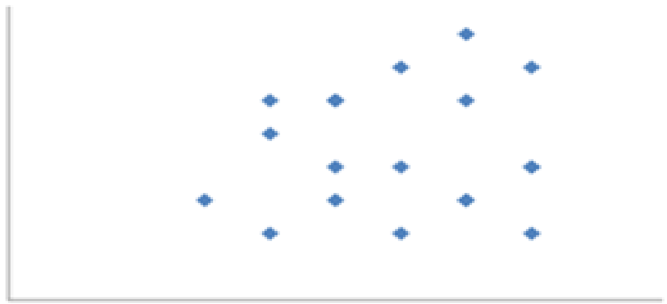
$Perfect \ positive \ correlation$

Perfect negative correlation



Negative correlation

Positive correlation



No correlation

Question

Problem: A researcher wants to know the relation between advertisement expenditure and total sales. For this purpose he took a sample data of 7 companies for one year. The data is given below in the table. Find the correlation coefficient and interpret your result.

Advertisement exp (Million dollars)	Annual Sales (Million dollars)
9	19
7	13
5	12
8	16
6	15
3	10
4	8

Solution

Solution:

x	y	xy	x^2	y^2
9	19	171	81	361
7	13	91	49	169
5	12	60	25	144
8	16	128	64	256
6	15	90	36	225
3	10	30	9	100
4	8	32	16	64
42	93	602	280	1319

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

$$r = \frac{7(602) - (42)(93)}{\sqrt{[7(280) - (42)^2][7(1319) - (93)^2]}}$$

$$r = \frac{4214 - 3906}{\sqrt{[1960 - 1764][9233 - 8649]}}$$

$$r = \frac{308}{\sqrt{[196][584]}}$$

$$r = \frac{308}{\sqrt{114464}}$$

$$r = \frac{308}{338.33}$$

$$r = 0.910$$

Example 14.10.

Calculate and analyze the correlation coefficient between the number of study hours and the number of sleeping hours of different students.

Number of study hours	2	4	6	8	10
Number of sleeping hours	10	9	8	7	6

Solution:

The necessary calculations are given below:

X	Y	$(X-\bar{X})$	$(Y-\bar{Y})$	$(X-\bar{X})(Y-\bar{Y})$	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$
2	10	–4	+2	–8	16	4
4	9	–2	+1	–2	4	1
6	8	0	0	0	0	0
8	7	+2	–1	–2	4	1
10	6	+4	–2	–8	16	4
$\Sigma X = 30$	$\Sigma Y = 40$	$\Sigma(X-\bar{X}) = 0$	$\Sigma(Y-\bar{Y}) = 0$	$\Sigma(X-\bar{X})(Y-\bar{Y}) = -20$	$\Sigma(X-\bar{X})^2 = 40$	$\Sigma(Y-\bar{Y})^2 = 10$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{30}{5} = 6 \quad \text{and} \quad \bar{Y} = \frac{\Sigma Y}{n} = \frac{40}{5} = 8$$

$$r_{xy} = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma(X-\bar{X})^2\,\Sigma(Y-\bar{Y})^2}} = \frac{-20}{\sqrt{(40)\,(10)}} = \frac{-20}{20} = -1$$

There is perfect negative correlation between the number of study hours and the number of sleeping hours.

Example 14.11

Example 14.13.

From the following data, compute the coefficient of correlation between X and Y:

	X series	Y series
Number of items	15	15
Arithmetic mean	25	18
Sum of square of deviations from arithmetic mean	136	138

Summation of products of deviations of X and Y series from their arithmetic means = 122.

Solution:

Here $n = 15, \bar{X} = 25, \ \bar{Y} = 18, \ \Sigma(X - \bar{X})^2 = 136, \Sigma(Y - \bar{Y})^2 = 138,$

$\Sigma(X - \bar{X})(Y - \bar{Y}) = 122$ and hence

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma(X - \bar{X})^2 \, \Sigma(Y - \bar{Y})^2}} = \frac{122}{\sqrt{(136)(138)}} = \frac{122}{137} = 0.89$$