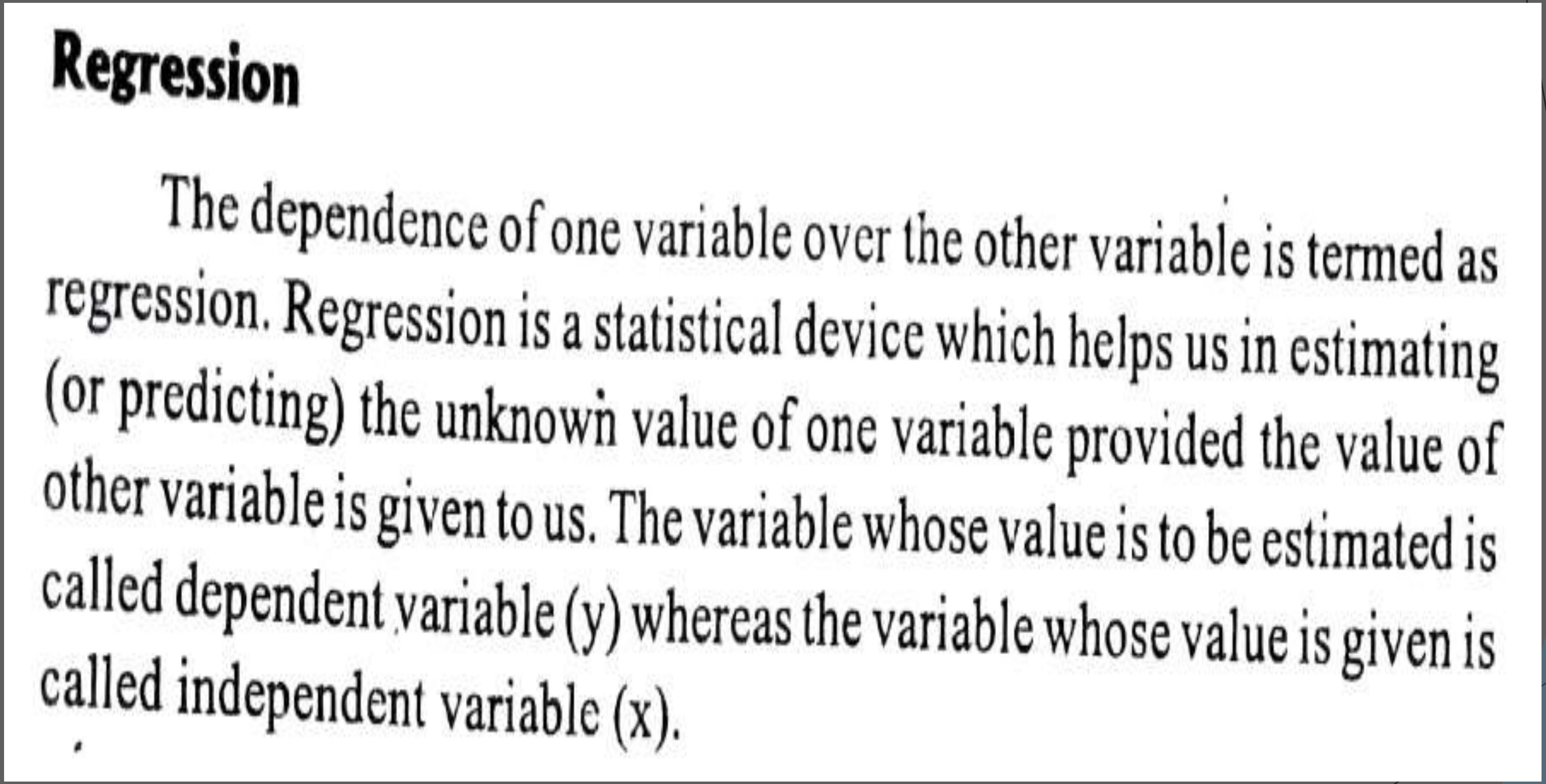
بسم الله الرحمن الرحيم

Probability & Statistics MS-301 Software Engineering

Lecture # 7

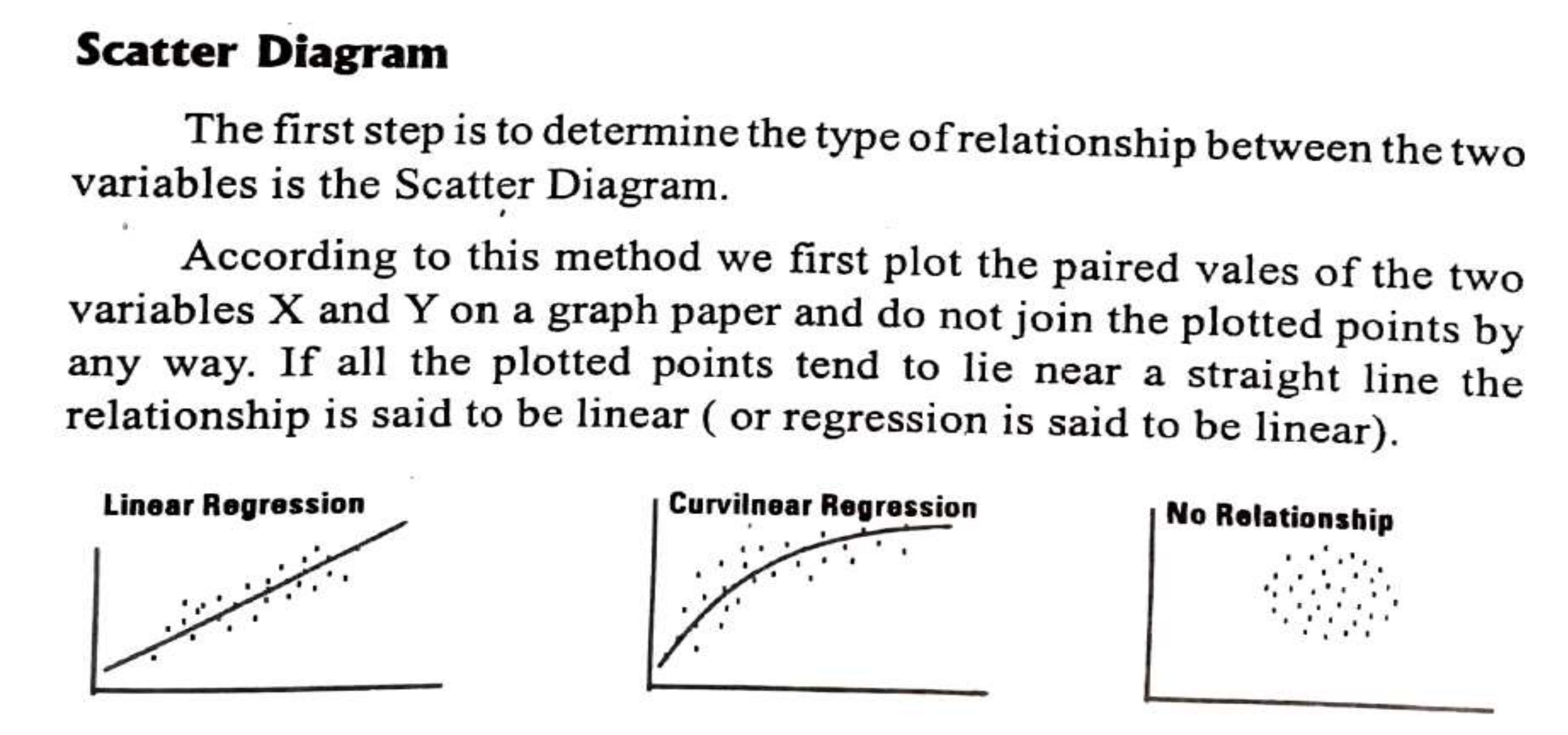
(Regression and Correlation)

Prepared by Dr. Asma Zaffar & Fouzia Iqbal



Regression

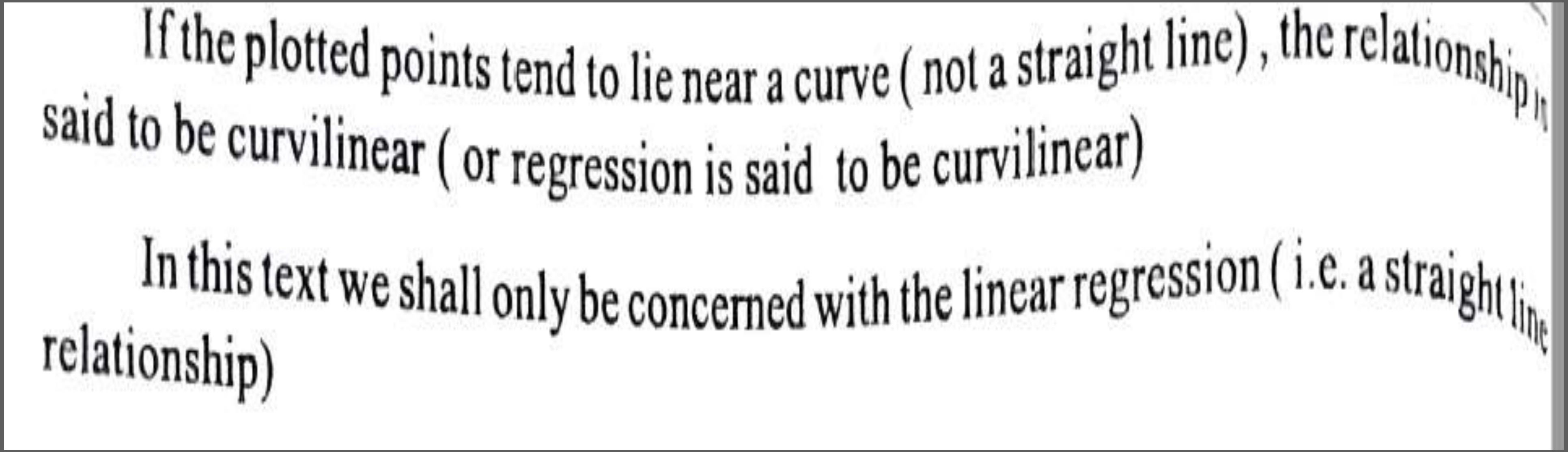
The dependence of one variable over the other variable is termed as regression. Regression is a statistical device which helps us in estimating (or predicting) the unknown value of one variable provided the value of other variable is given to us. The variable whose value is to be estimated is called dependent variable (y) whereas the variable whose value is given is called independent variable (x).



Scatter Diagram

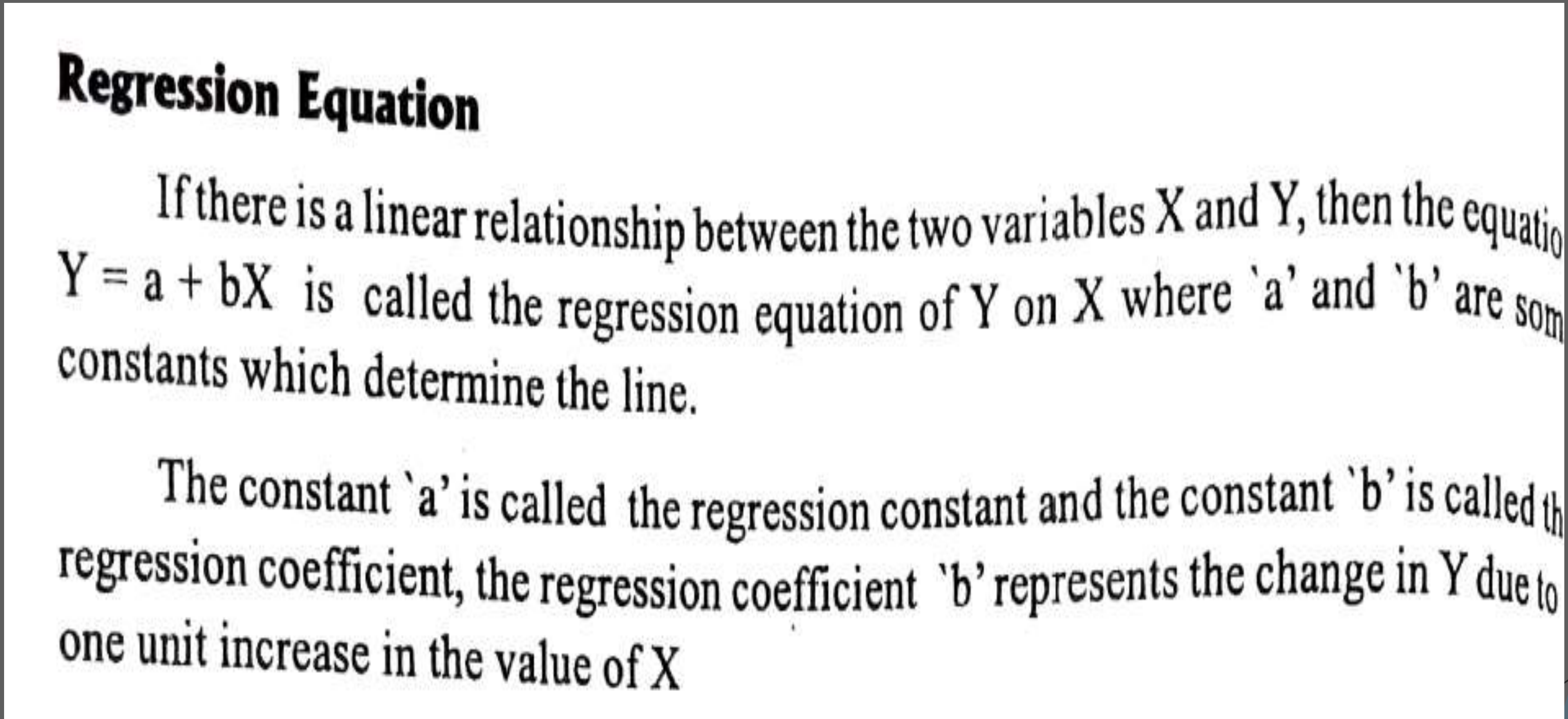
The first step is to determine the type of relationship between the two variables is the Scatter Diagram.

According to this method we first plot the paired vales of the two variables X and Y on a graph paper and do not join the plotted points by any way. If all the plotted points tend to lie near a straight line the relationship is said to be linear (or regression is said to be linear).



If the plotted points tend to lie near a curve (not a straight line) , the relationship is said to be curvilinear (or regression is said to be curvilinear)

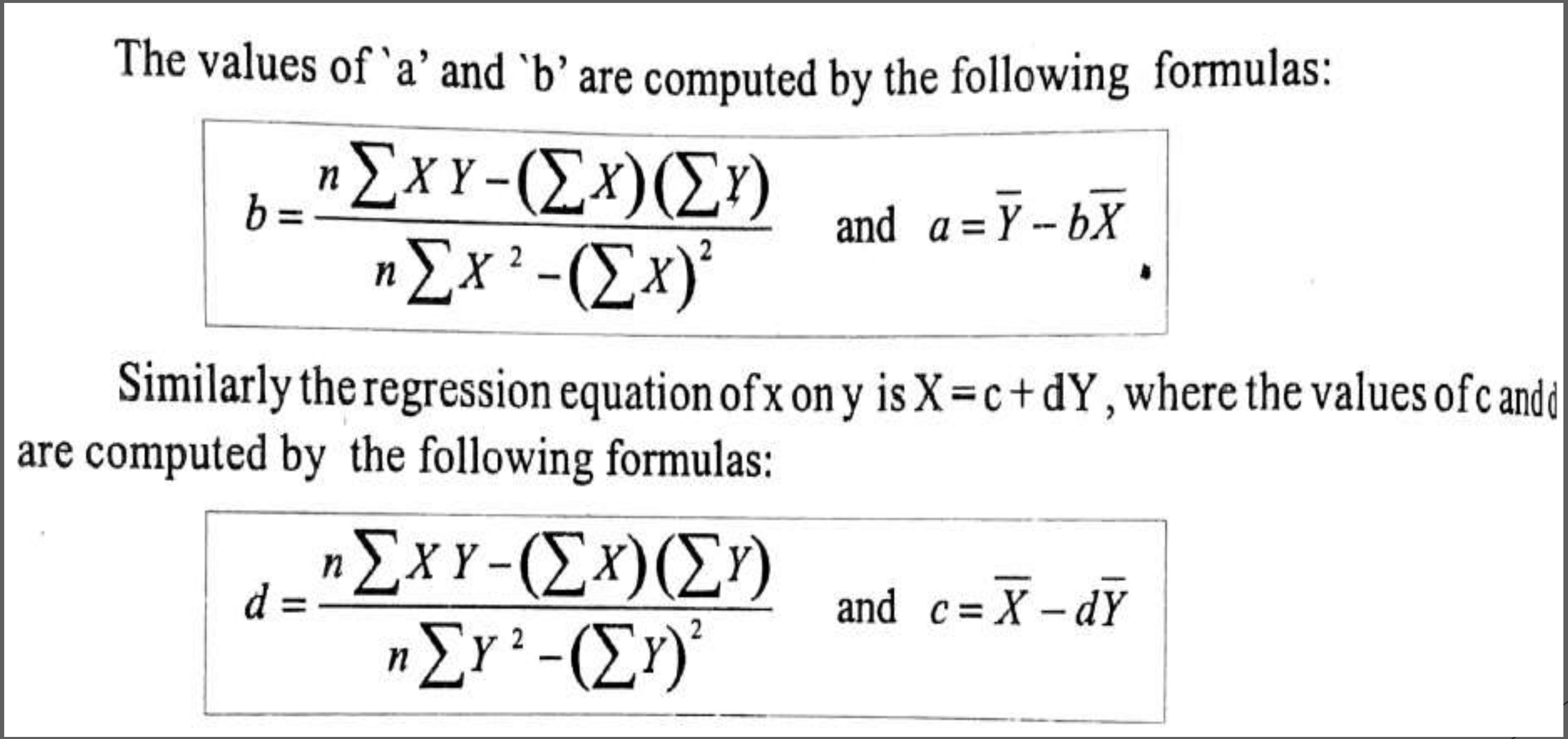
In this text we shall only be concerned with the linear regression (i.e. a straight line relationship)



Regression Equation

If there is a linear relationship between the two variables X and Y, then the equatio $Y = a + bX$ is called the regression equation of Y on X where `a’ and `b’ are som constants which determine the line.

The constant `a’ is called the regression constant and the constant `b’ is called th regression coefficient, the regression coefficient `b’ represents the change in Y due to one unit increase in the value of X

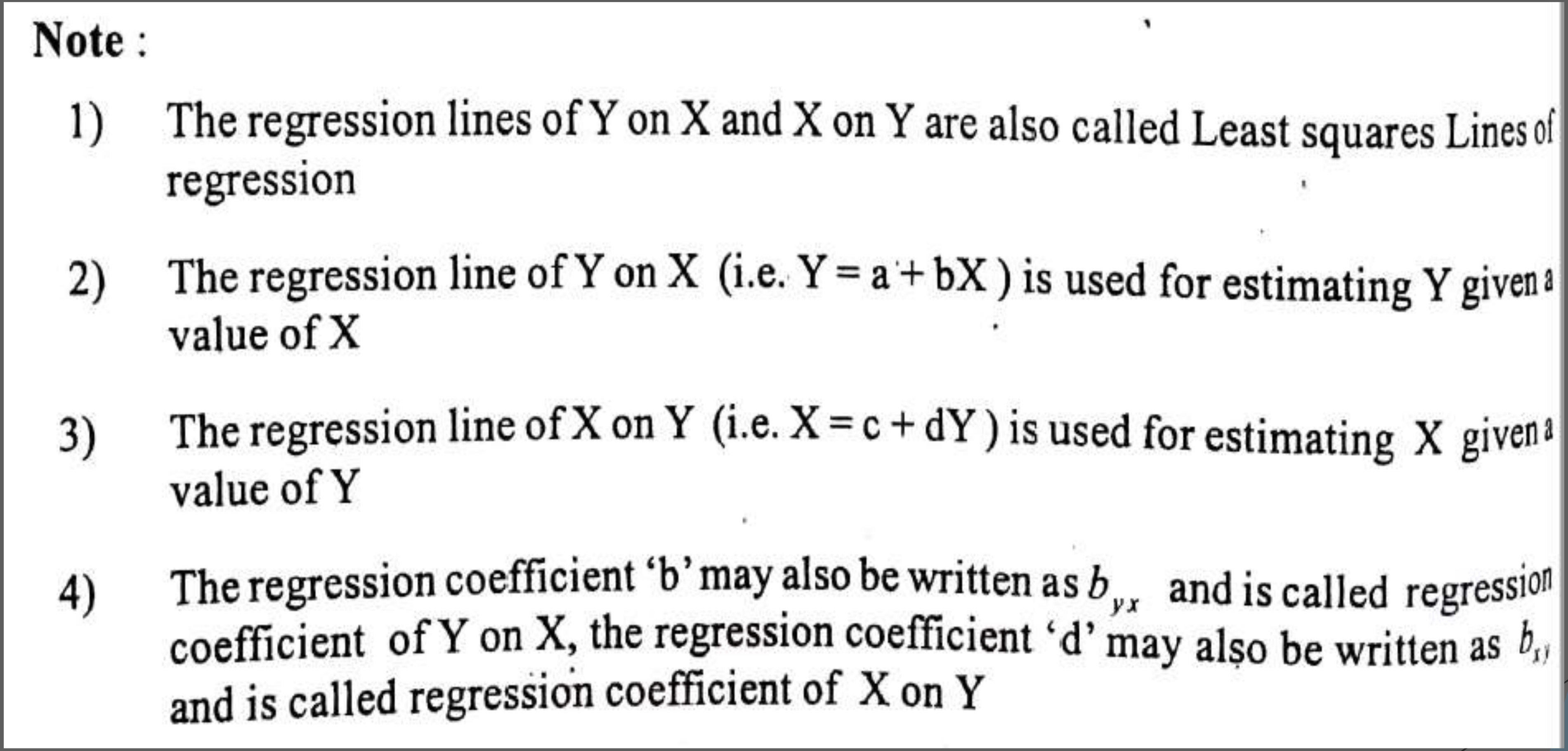


The values of `a' and `b' are computed by the following formulas:

$$\boxed{b = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2} \quad \text{and} \quad a = \bar{Y} - b\bar{X}}$$

Similarly the regression equation of x on y is $X = c + dY$, where the values of c and d are computed by the following formulas:

$$\boxed{d = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum Y^2 - (\sum Y)^2} \quad \text{and} \quad c = \bar{X} - d\bar{Y}}$$



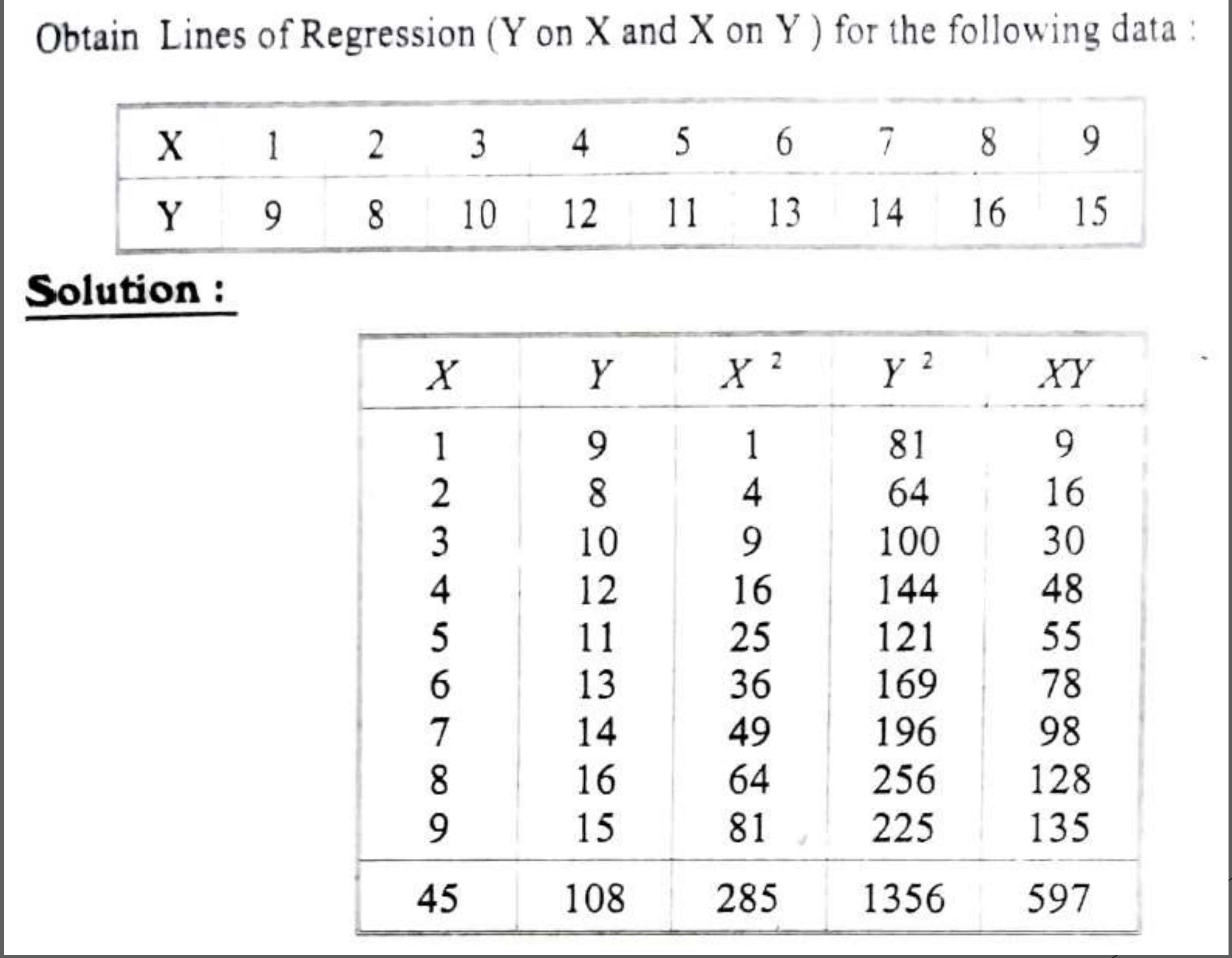
Note :

1) The regression lines of Y on X and X on Y are also called Least squares Lines of regression

2) The regression line of Y on X (i.e. $Y = a + bX$) is used for estimating Y given a value of X

3) The regression line of X on Y (i.e. $X = c + dY$) is used for estimating X given a value of Y

4) The regression coefficient 'b' may also be written as b_{yx} and is called regression coefficient of Y on X, the regression coefficient 'd' may also be written as b_{xy} and is called regression coefficient of X on Y

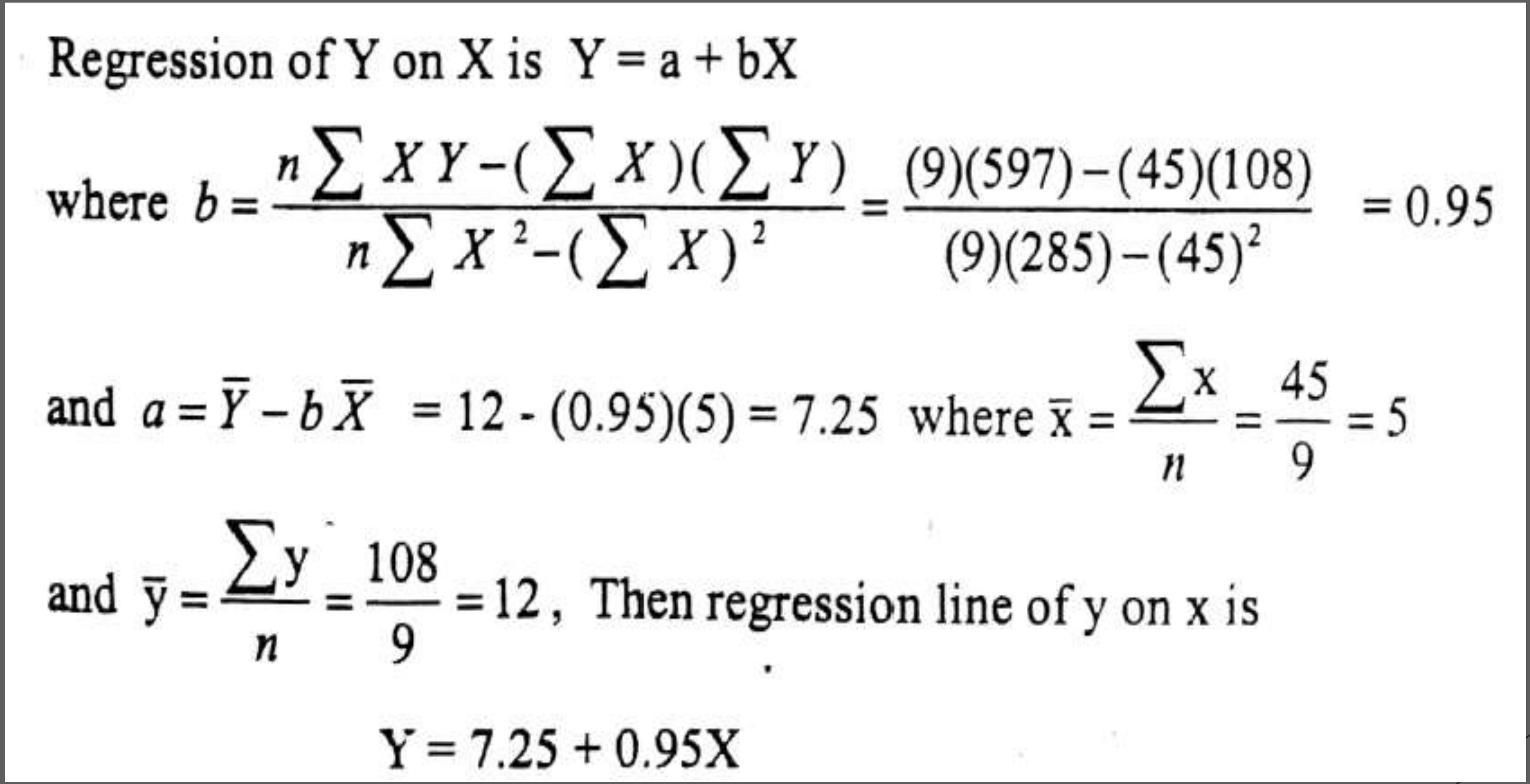


Obtain Lines of Regression (Y on X and X on Y) for the following data :

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

Solution :

X	Y	X^2	Y^2	XY
1	9	1	81	9
2	8	4	64	16
3	10	9	100	30
4	12	16	144	48
5	11	25	121	55
6	13	36	169	78
7	14	49	196	98
8	16	64	256	128
9	15	81	225	135
45	108	285	1356	597



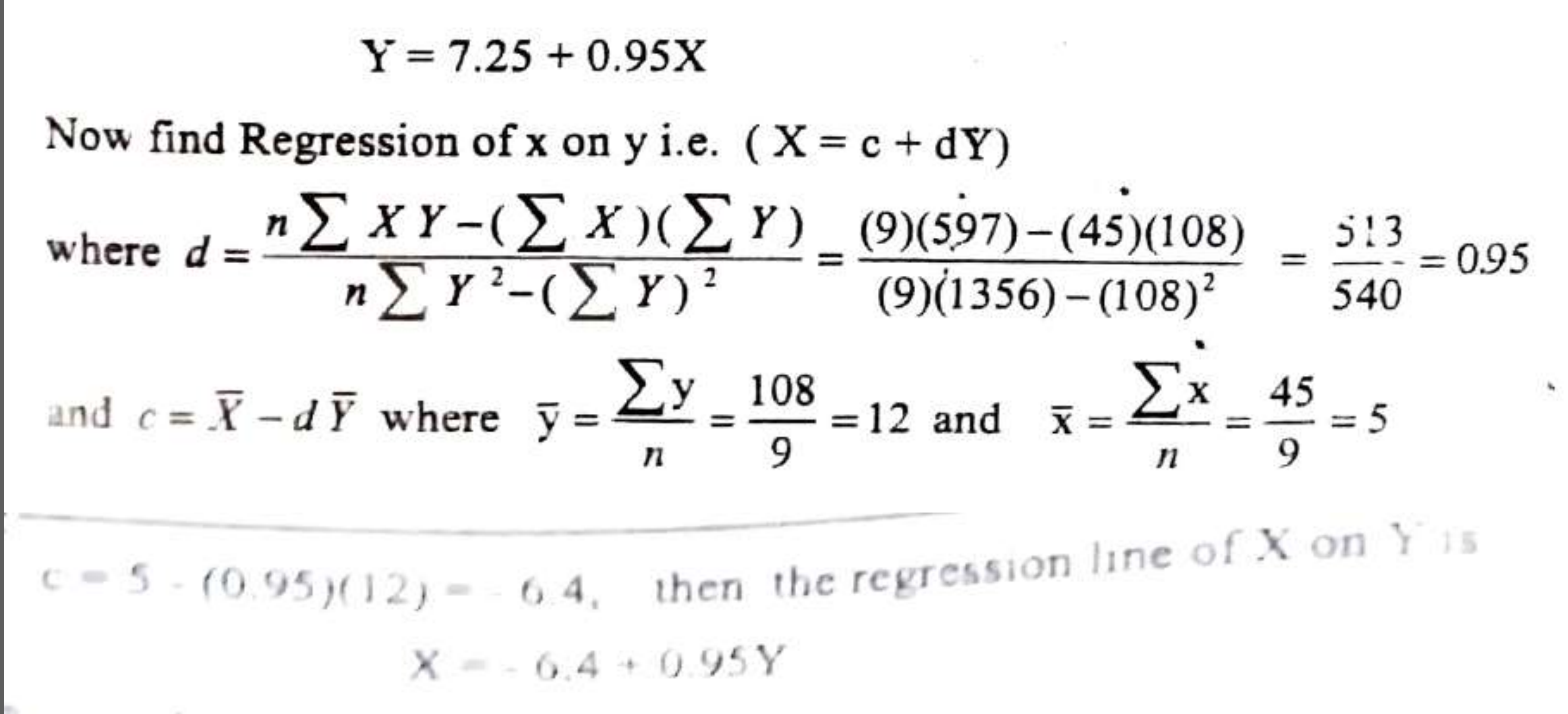
Regression of Y on X is $Y = a + bX$

$$\text{where } b = \frac{n\sum XY-(\sum X)(\sum Y)}{n\sum X^{2}-(\sum X)^{2}} = \frac{(9)(597)-(45)(108)}{(9)(285)-(45)^{2}} = 0.95$$

$$\text{and } a = \bar{Y} - b\bar{X} = 12 - (0.95)(5) = 7.25 \quad \text{where } \bar{x} = \frac{\sum x}{n} = \frac{45}{9} = 5$$

and $\bar{y} = \frac{\sum y}{n} = \frac{108}{9} = 12$, Then regression line of y on x is

$$Y = 7.25 + 0.95X$$



$$Y = 7.25 + 0.95X$$

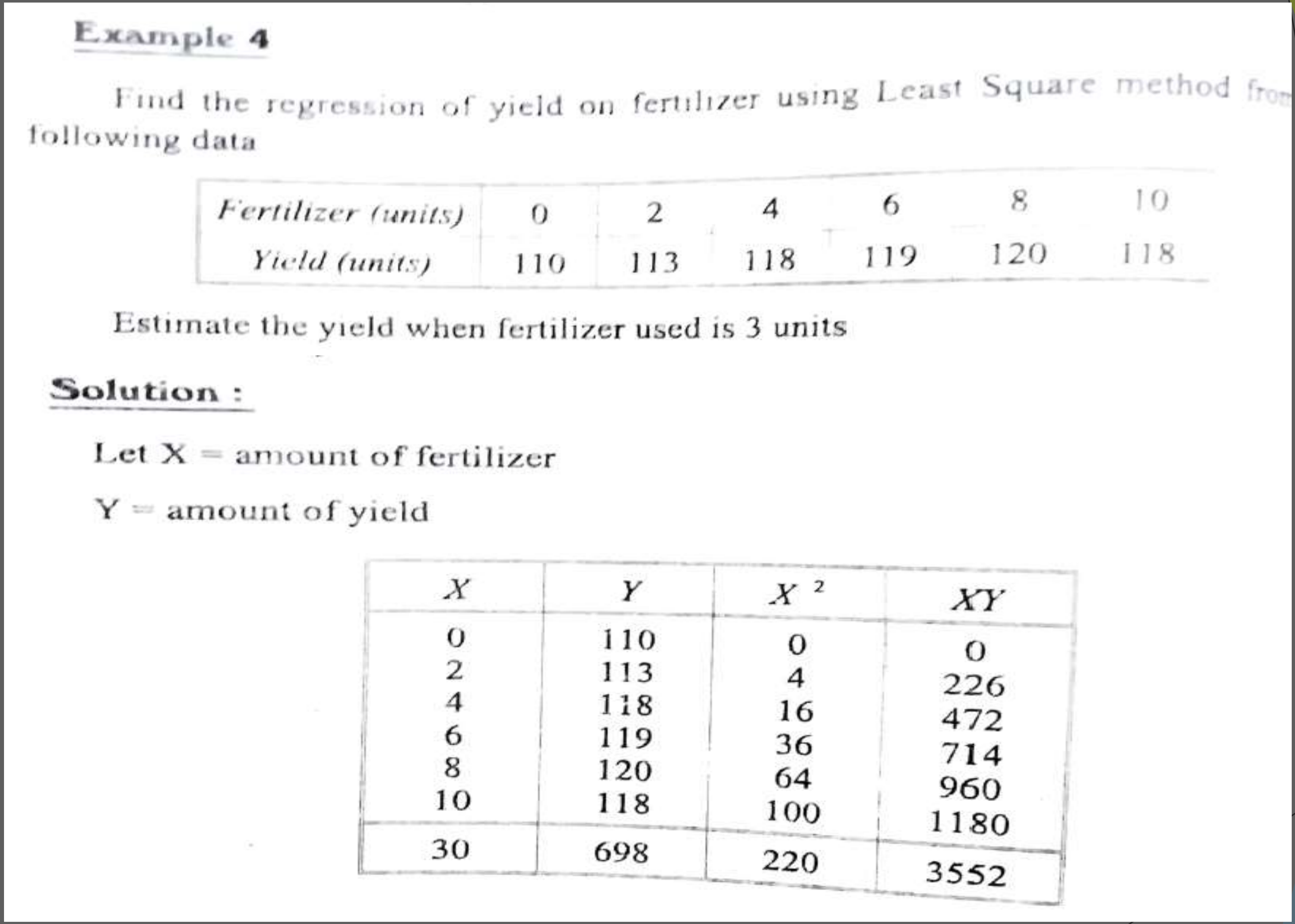
Now find Regression of x on y i.e. $(X = c + dY)$

where $d = \dfrac{n\sum XY-(\sum X)(\sum Y)}{n\sum Y^2-(\sum Y)^2} = \dfrac{(9)(597)-(45)(108)}{(9)(1356)-(108)^2} = \dfrac{513}{540} = 0.95$

and $c = \bar{X} - d\bar{Y}$ where $\bar{y} = \dfrac{\sum y}{n} = \dfrac{108}{9} = 12$ and $\bar{x} = \dfrac{\sum x}{n} = \dfrac{45}{9} = 5$

$c = 5 - (0.95)(12) = -6.4$, then the regression line of X on Y is

$$X = -6.4 + 0.95Y$$



Example 4

Find the regression of yield on fertilizer using Least Square method from following data

Fertilizer (units)	0	2	4	6	8	10
Yield (units)	110	113	118	119	120	118

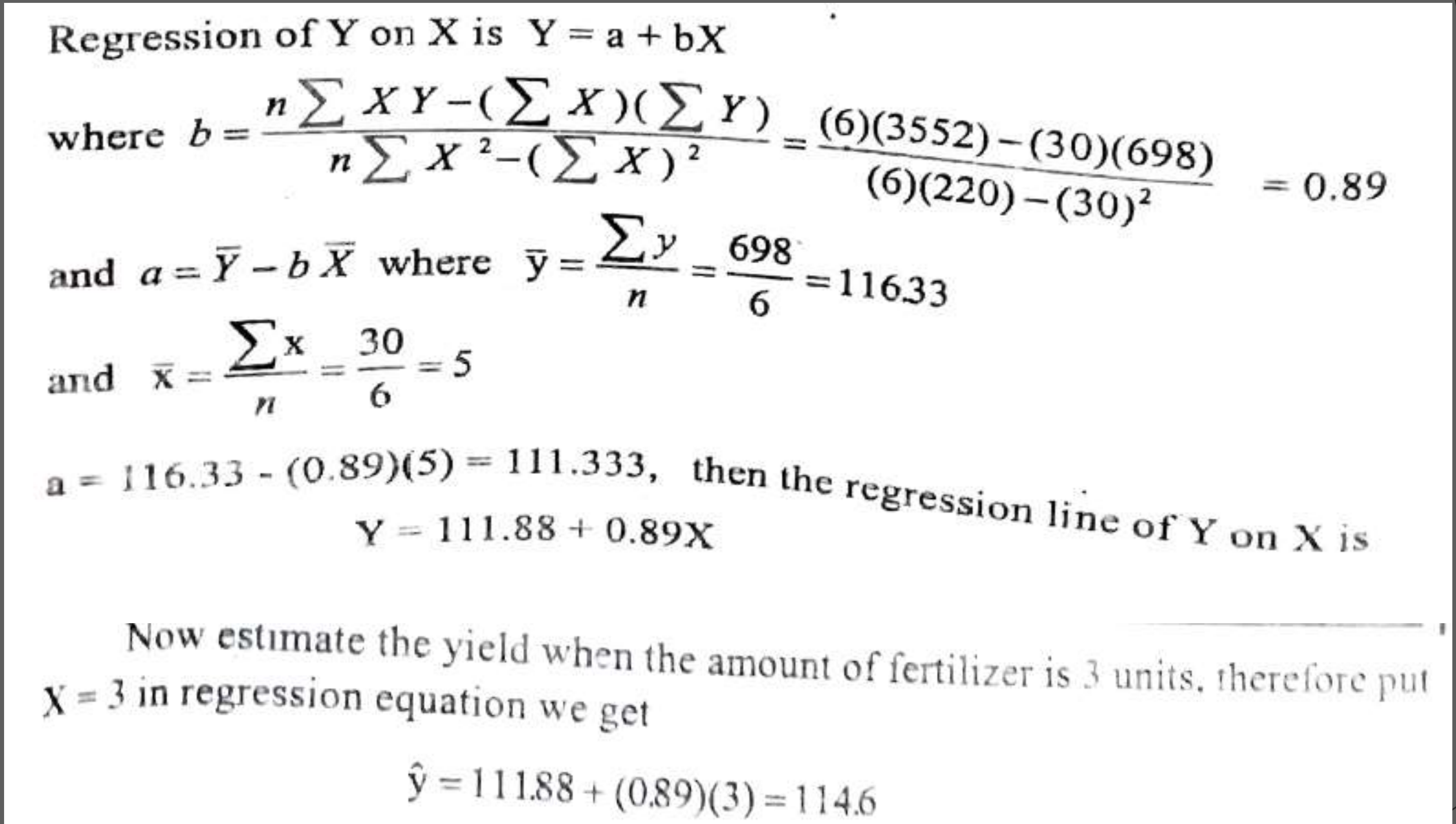
Estimate the yield when fertilizer used is 3 units

Solution :

Let X = amount of fertilizer

Y = amount of yield

X	Y	X^2	XY
0 2 4 6 8 10	110 113 118 119 120 118	0 4 16 36 64 100	0 226 472 714 960 1180
30	698	220	3552



Regression of Y on X is $Y = a + bX$

where $b = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2} = \frac{(6)(3552)-(30)(698)}{(6)(220)-(30)^2} = 0.89$

and $a = \bar{Y} - b\bar{X}$ where $\bar{y} = \frac{\sum y}{n} = \frac{698}{6} = 116.33$

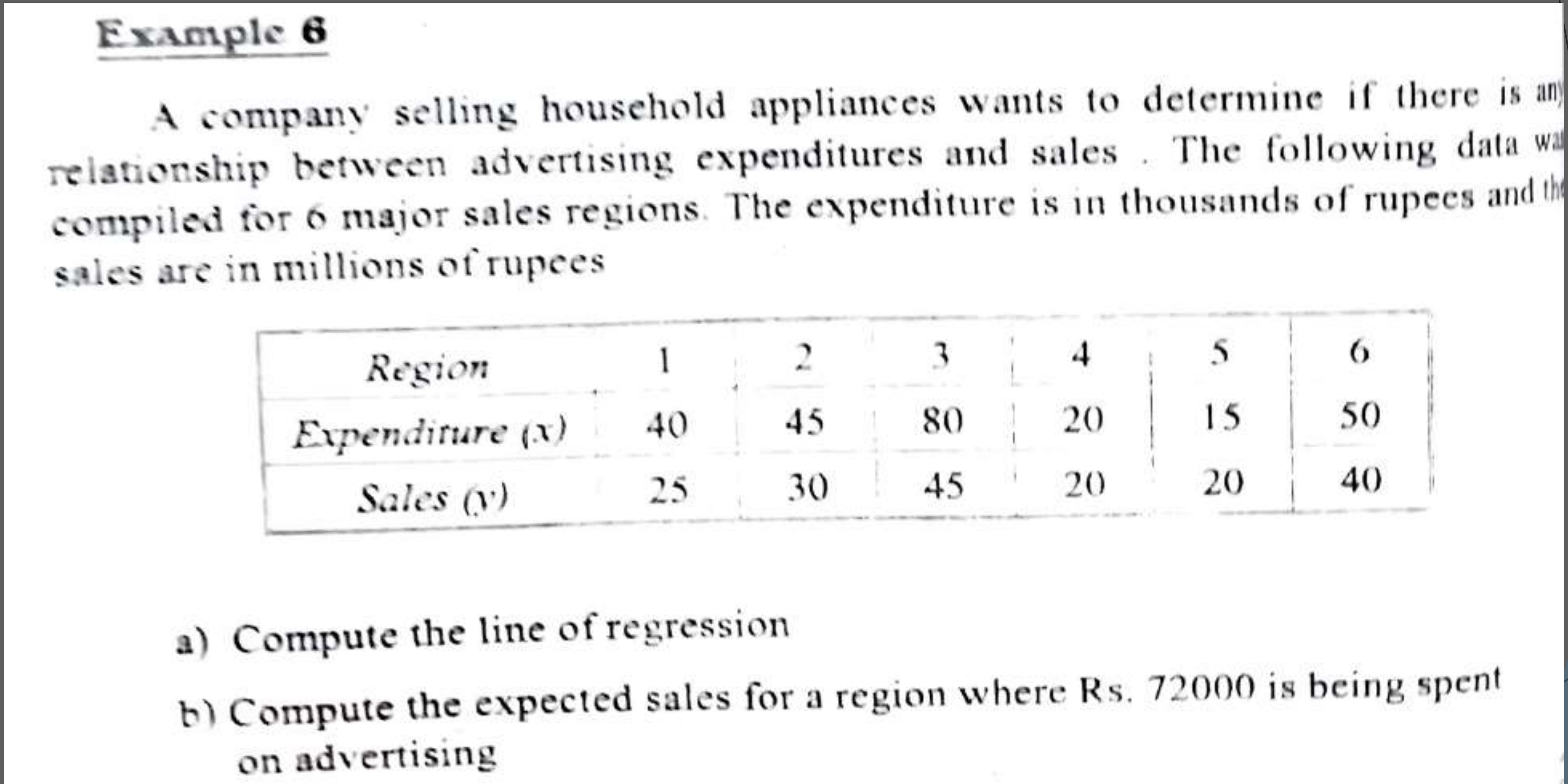
and $\bar{x} = \frac{\sum x}{n} = \frac{30}{6} = 5$

a = 116.33 - (0.89)(5) = 111.333, then the regression line of Y on X is

$$Y = 111.88 + 0.89X$$

Now estimate the yield when the amount of fertilizer is 3 units, therefore put X = 3 in regression equation we get

$$\hat{y} = 111.88 + (0.89)(3) = 114.6$$



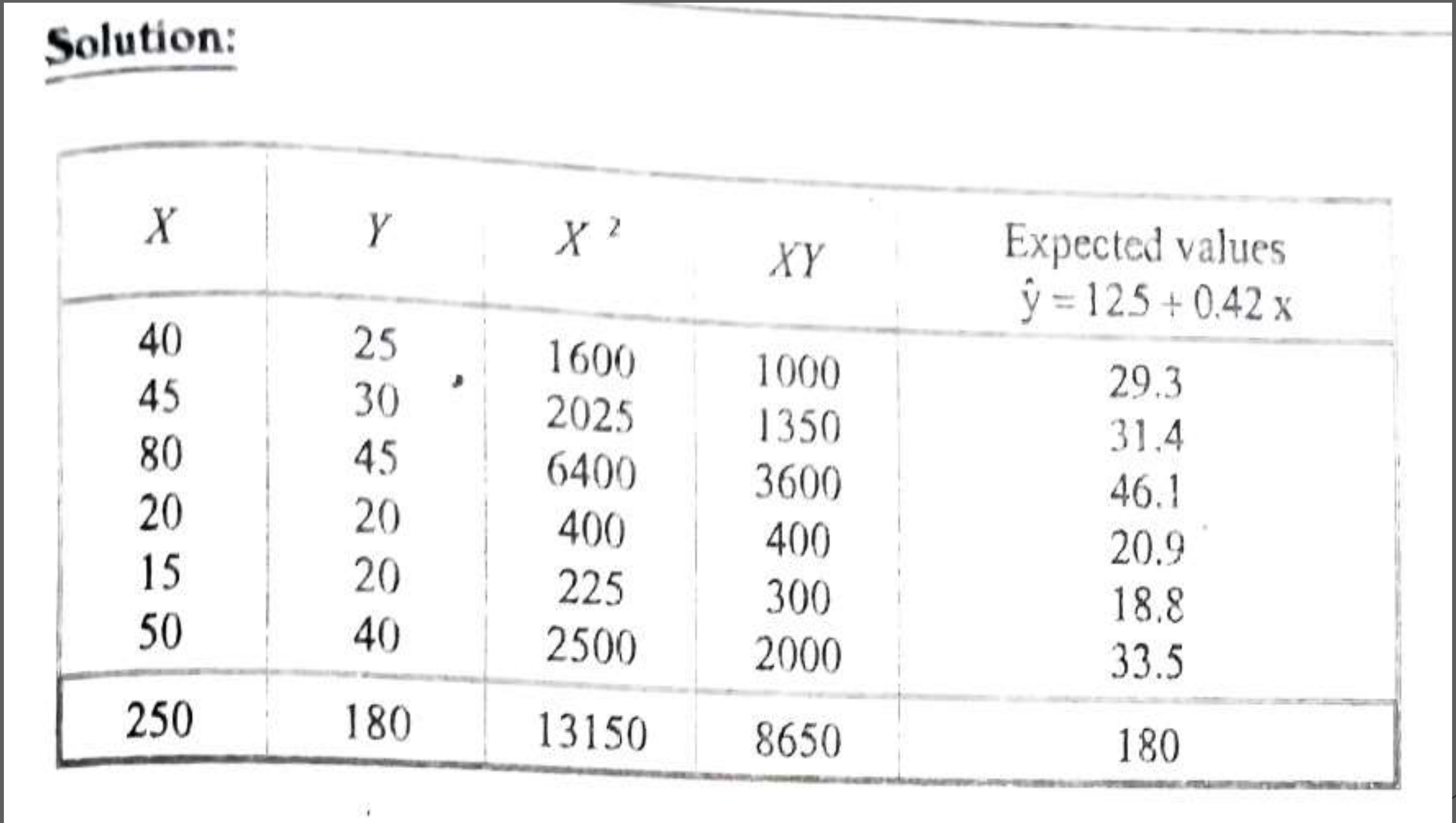
Example 6

A company selling household appliances wants to determine if there is any relationship between advertising expenditures and sales . The following data wa compiled for 6 major sales regions. The expenditure is in thousands of rupees and th sales are in millions of rupees

Region	1	2	3	4	5	6
Expenditure (x)	40	45	80	20	15	50
Sales (y)	25	30	45	20	20	40

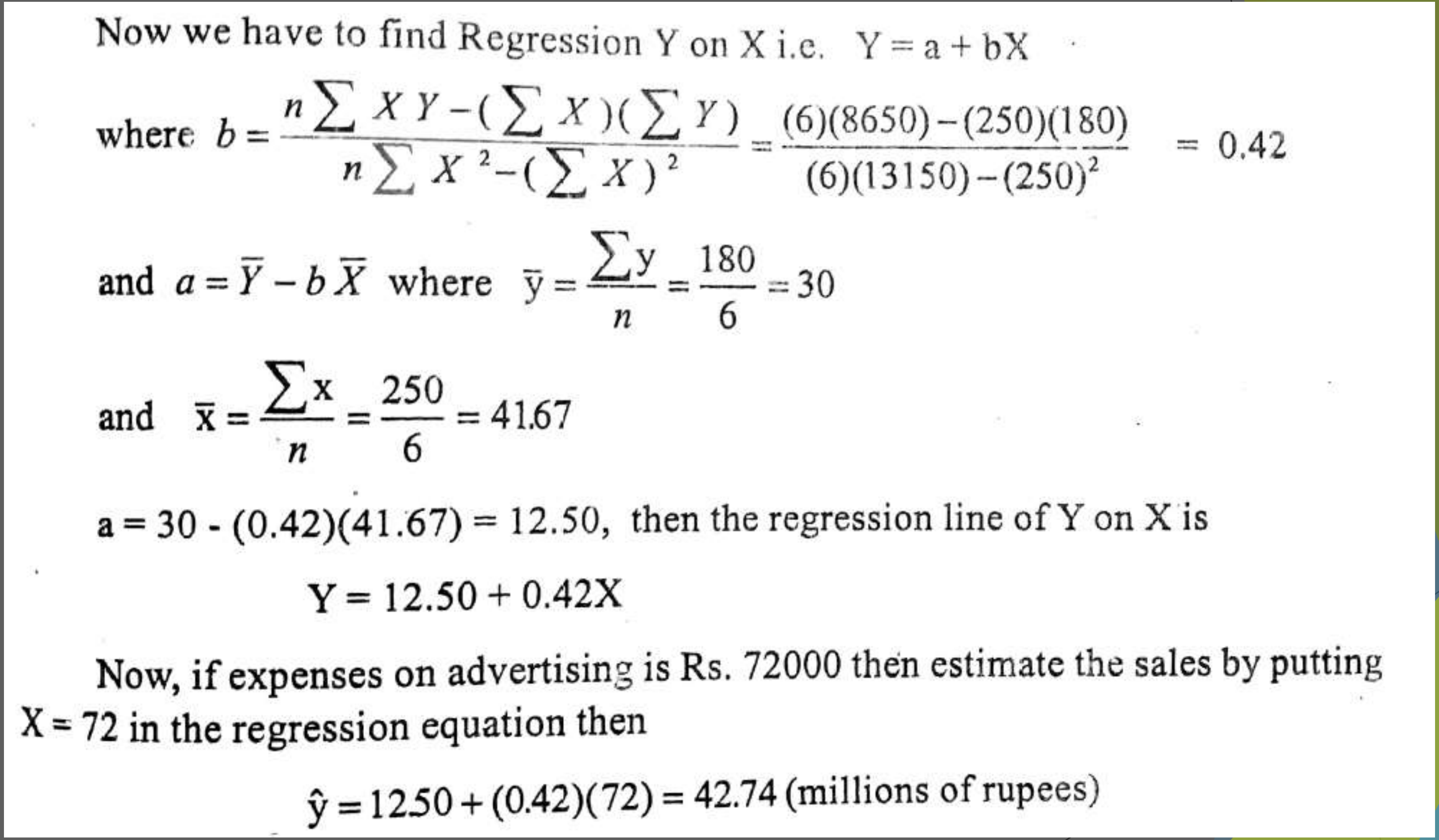
a) Compute the line of regression

b) Compute the expected sales for a region where Rs. 72000 is being spent on advertising



Solution:

X	Y	X^2	XY	Expected values $\hat{y} = 12.5 + 0.42\,x$
40	25	1600	1000	29.3
45	30	2025	1350	31.4
80	45	6400	3600	46.1
20	20	400	400	20.9
15	20	225	300	18.8
50	40	2500	2000	33.5
250	180	13150	8650	180



Now we have to find Regression Y on X i.e. $Y = a + bX$

where $b = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2} = \frac{(6)(8650) - (250)(180)}{(6)(13150) - (250)^2} = 0.42$

and $a = \bar{Y} - b\bar{X}$ where $\bar{y} = \frac{\sum y}{n} = \frac{180}{6} = 30$

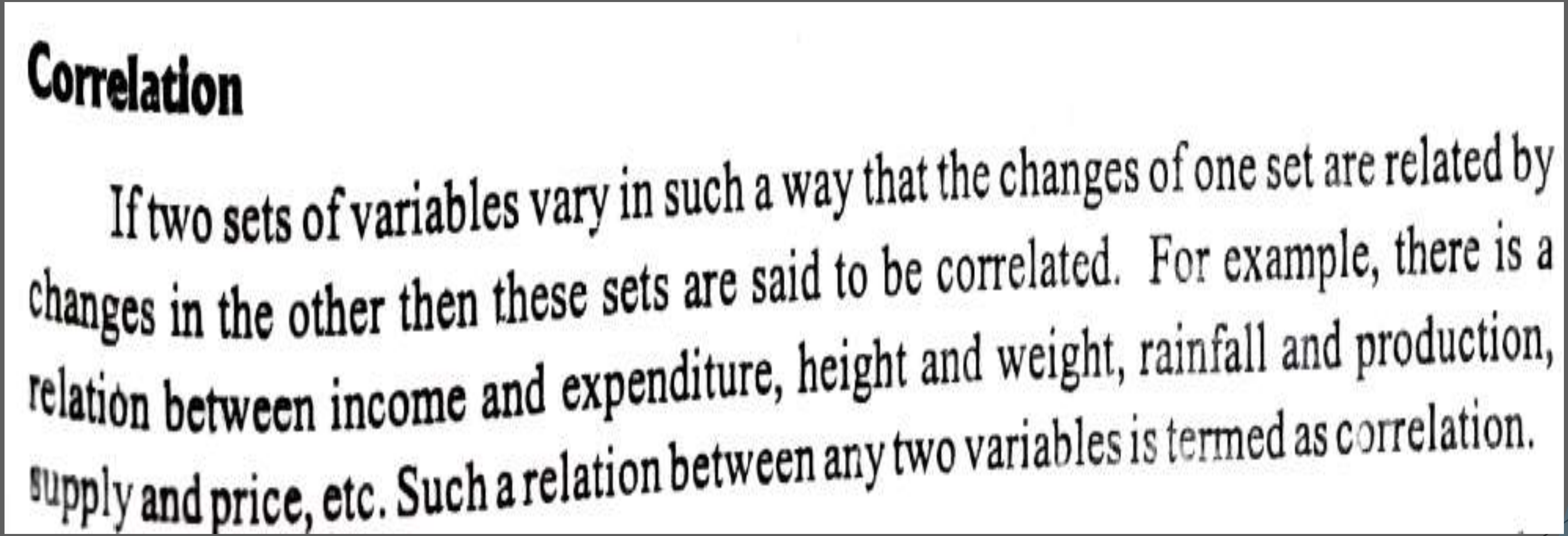
and $\bar{x} = \frac{\sum x}{n} = \frac{250}{6} = 41.67$

a = 30 - (0.42)(41.67) = 12.50, then the regression line of Y on X is

$$Y = 12.50 + 0.42X$$

Now, if expenses on advertising is Rs. 72000 then estimate the sales by putting X = 72 in the regression equation then

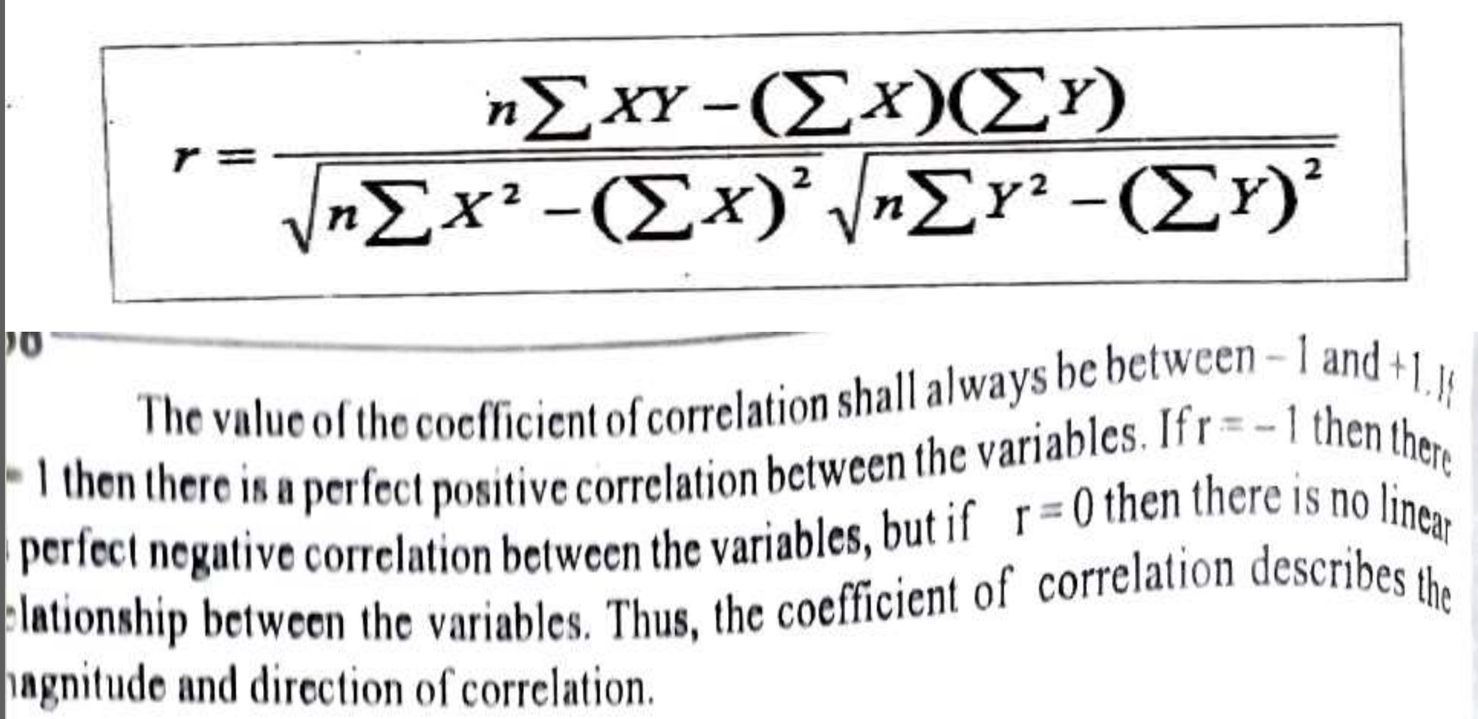
$$\hat{y} = 1250 + (0.42)(72) = 42.74 \text{ (millions of rupees)}$$



Correlation

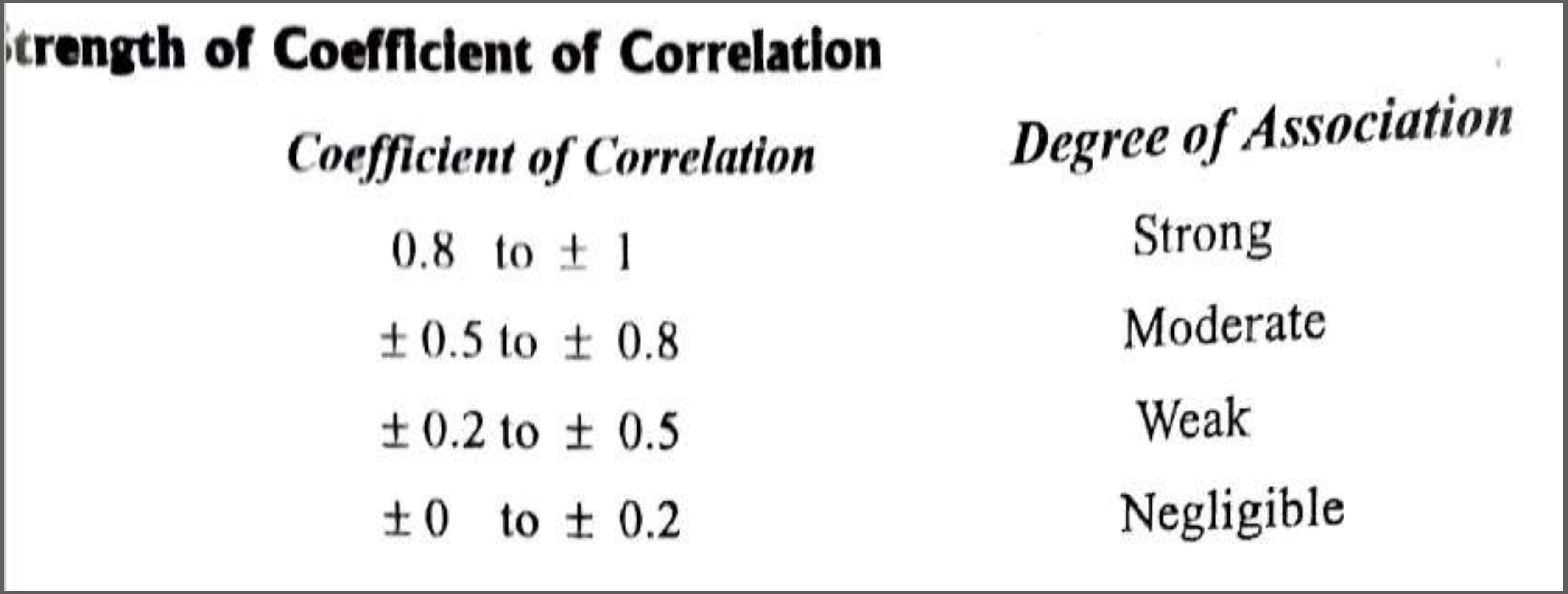
If two sets of variables vary in such a way that the changes of one set are related by changes in the other then these sets are said to be correlated. For example, there is a relation between income and expenditure, height and weight, rainfall and production, supply and price, etc. Such a relation between any two variables is termed as correlation.

Coefficient of Correlation



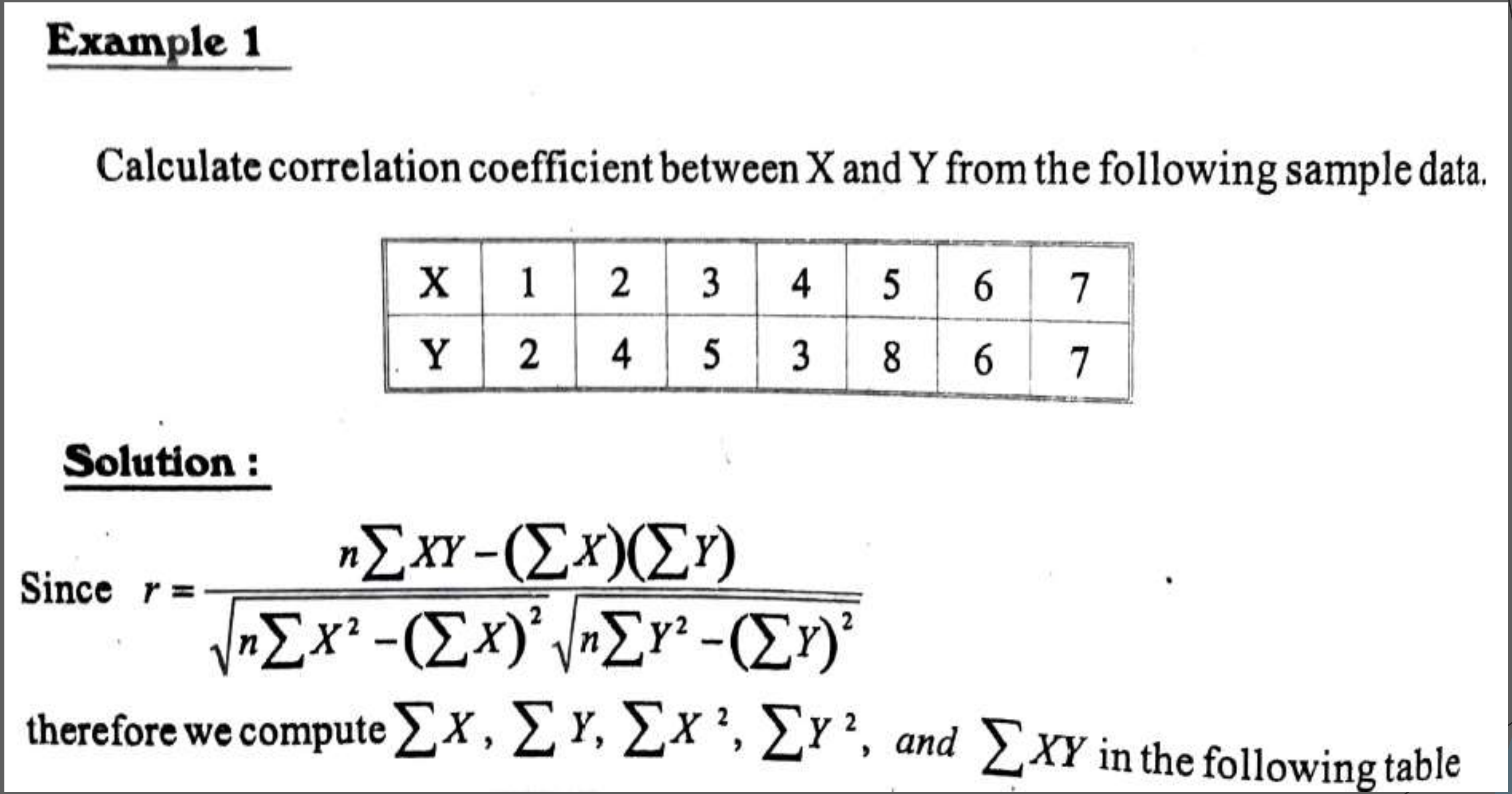
$$r = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{n\sum X^2 - (\sum X)^2}\sqrt{n\sum Y^2 - (\sum Y)^2}}$$

The value of the coefficient of correlation shall always be between – 1 and +1. If
1 then there is a perfect positive correlation between the variables. If r = – 1 then there
perfect negative correlation between the variables, but if r = 0 then there is no linear
lationship between the variables. Thus, the coefficient of correlation describes the
agnitude and direction of correlation.



trength of Coefficient of Correlation

Coefficient of Correlation	*Degree of Association*
0.8 to ± 1	Strong
± 0.5 to ± 0.8	Moderate
± 0.2 to ± 0.5	Weak
± 0 to ± 0.2	Negligible



Example 1

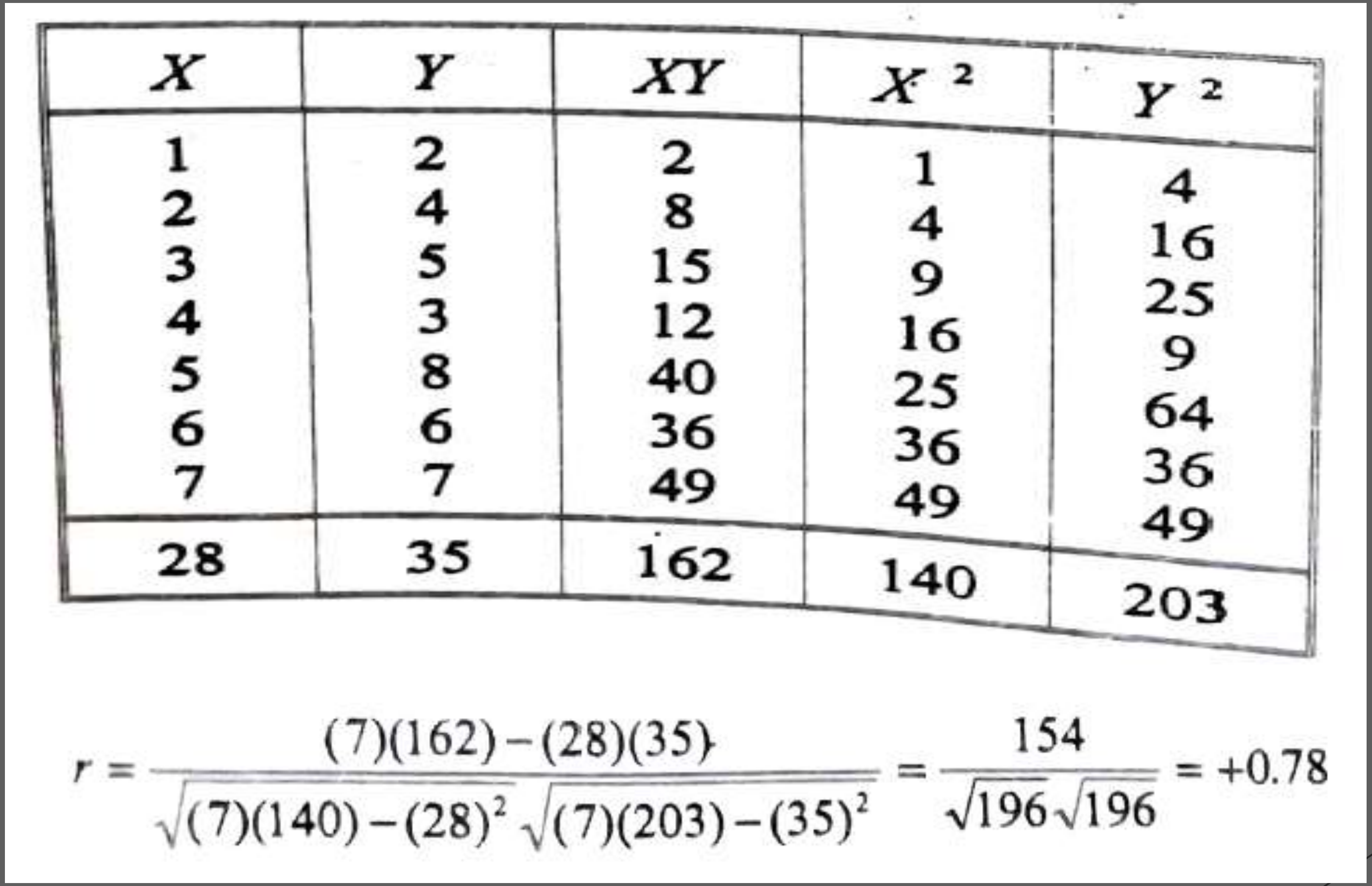
Calculate correlation coefficient between X and Y from the following sample data.

X	1	2	3	4	5	6	7
Y	2	4	5	3	8	6	7

Solution :

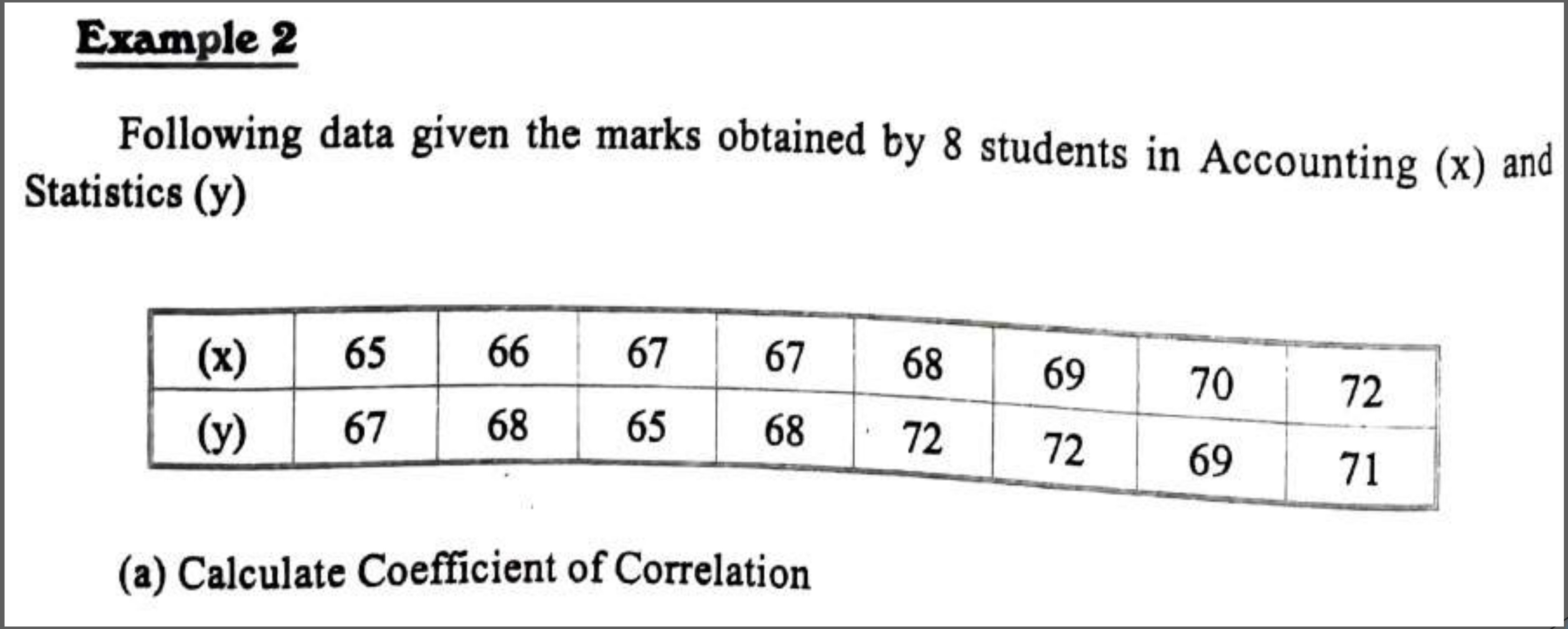
Since $r = \dfrac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{n\sum X^2 - (\sum X)^2}\sqrt{n\sum Y^2 - (\sum Y)^2}}$

therefore we compute $\sum X$, $\sum Y$, $\sum X^2$, $\sum Y^2$, *and* $\sum XY$ in the following table



X	Y	XY	X^2	Y^2
1	2	2	1	4
2	4	8	4	16
3	5	15	9	25
4	3	12	16	9
5	8	40	25	64
6	6	36	36	36
7	7	49	49	49
28	35	162	140	203

$$r = \frac{(7)(162)-(28)(35)}{\sqrt{(7)(140)-(28)^2}\sqrt{(7)(203)-(35)^2}} = \frac{154}{\sqrt{196}\sqrt{196}} = +0.78$$

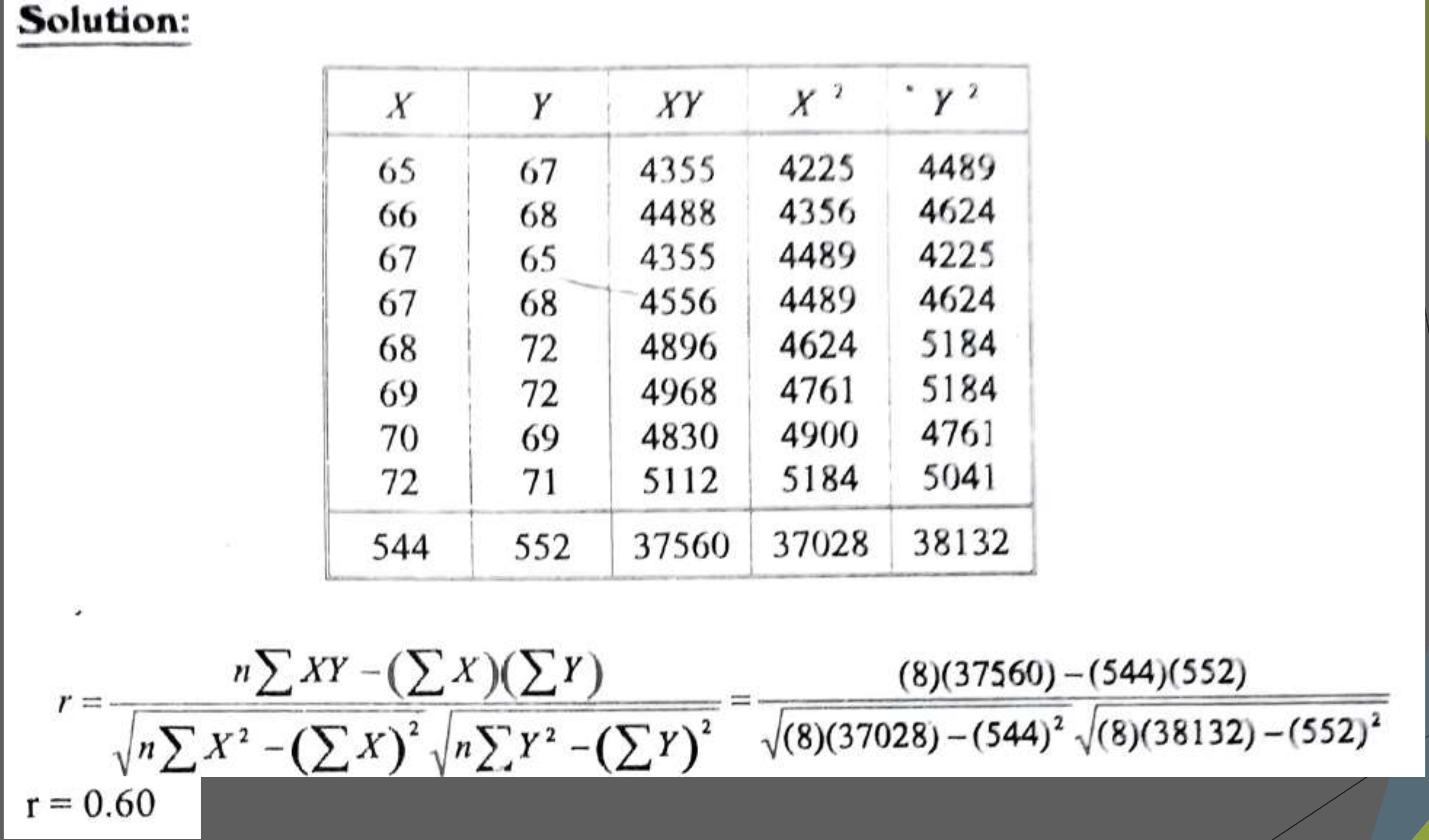


Example 2

Following data given the marks obtained by 8 students in Accounting (x) and Statistics (y)

(x)	65	66	67	67	68	69	70	72
(y)	67	68	65	68	72	72	69	71

(a) Calculate Coefficient of Correlation

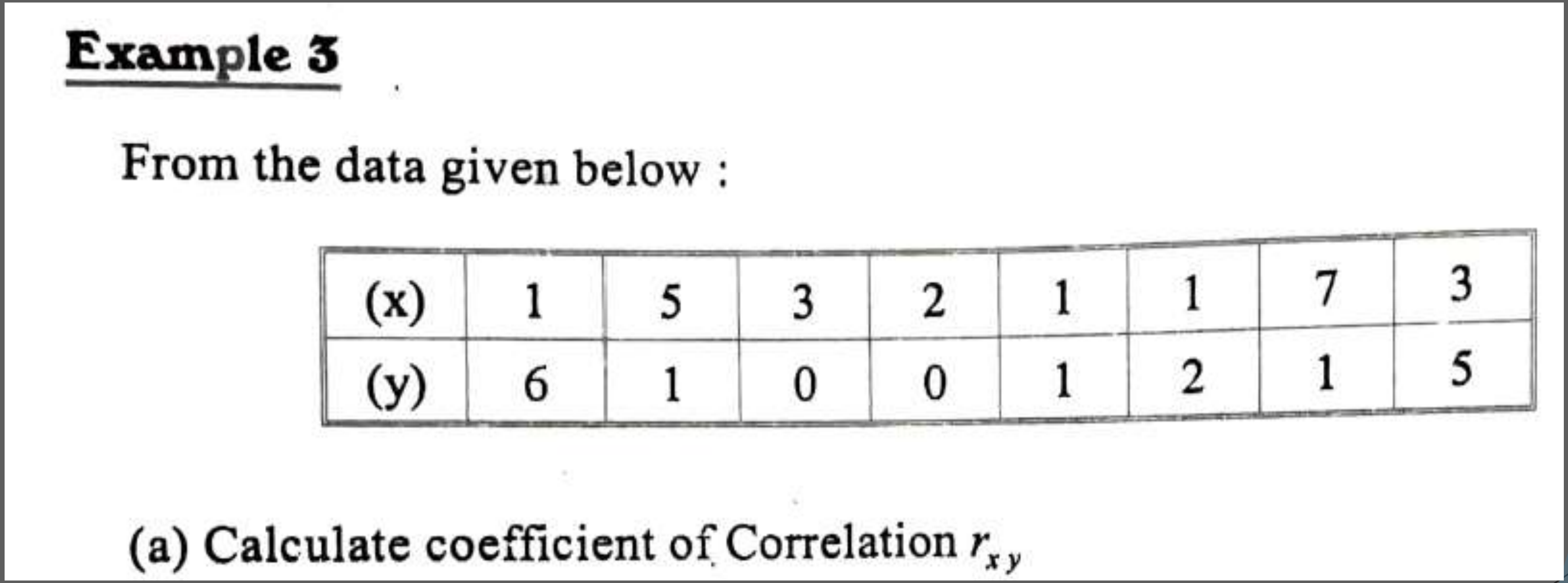


Solution:

X	Y	XY	X^2	Y^2
65	67	4355	4225	4489
66	68	4488	4356	4624
67	65	4355	4489	4225
67	68	4556	4489	4624
68	72	4896	4624	5184
69	72	4968	4761	5184
70	69	4830	4900	4761
72	71	5112	5184	5041
544	552	37560	37028	38132

$$r = \frac{n\sum XY - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{n\sum X^2 - \left(\sum X\right)^2}\sqrt{n\sum Y^2 - \left(\sum Y\right)^2}} = \frac{(8)(37560) - (544)(552)}{\sqrt{(8)(37028) - (544)^2}\sqrt{(8)(38132) - (552)^2}}$$

r = 0.60

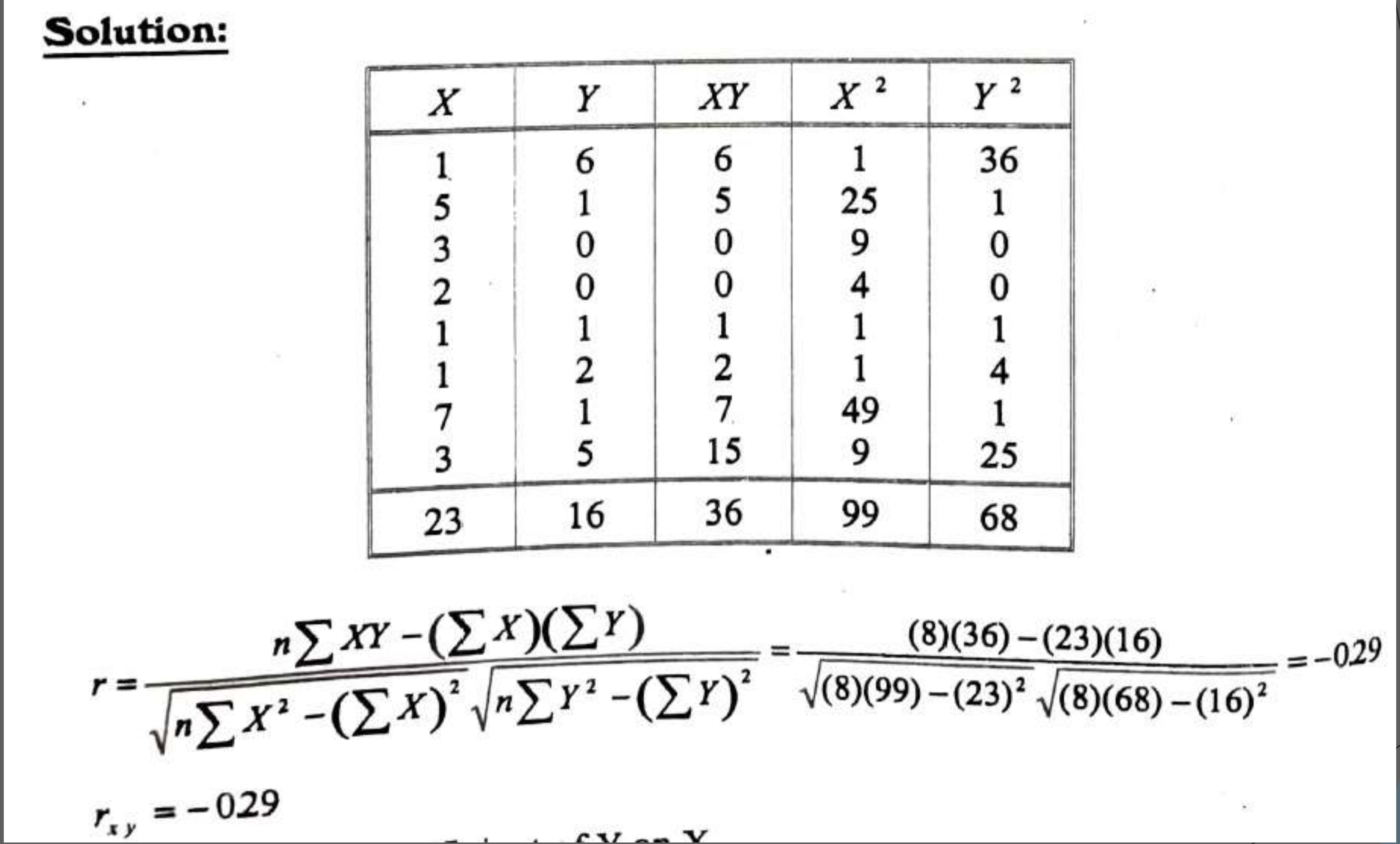


Example 3

From the data given below :

(x)	1	5	3	2	1	1	7	3
(y)	6	1	0	0	1	2	1	5

(a) Calculate coefficient of Correlation r_{xy}



Solution:

X	Y	XY	X^2	Y^2
1	6	6	1	36
5	1	5	25	1
3	0	0	9	0
2	0	0	4	0
1	1	1	1	1
1	2	2	1	4
7	1	7	49	1
3	5	15	9	25
23	16	36	99	68

$$r=\frac{n\sum XY-(\sum X)(\sum Y)}{\sqrt{n\sum X^2-(\sum X)^2}\sqrt{n\sum Y^2-(\sum Y)^2}}=\frac{(8)(36)-(23)(16)}{\sqrt{(8)(99)-(23)^2}\sqrt{(8)(68)-(16)^2}}=-0.29$$

$$r_{xy}=-0.29$$

Questions?