Module 3
Introduction to Regression

Contents

- Mathematical and Statistical Equation
- Meaning of Intercept and Slope
- Error term
- Measure for Model Fit
- R^2 | MAE | MAPE

- **Dependence of variable which take only numerical values.**
- **Establishing the actual relationship between two or more variables.**
- **This problem is dealt with regression.**
- **Degree of relationship between two or more variables.**
- **This problem is dealt with correlation analysis.**

Correlation is a statistical device which helps us in analyzing the covariation of two or more variables.

The problem of analyzing the relation between different series should be divided into three steps.

- **Determine whether a relation exists and if it does, measuring it.**
- **Testing whether it is significant.**
- **Establishing the cause and effect relation, if any.**

Correlation and Causation

The explanation of a significant degree of correlation may be anyone, or combination of the following reasons:

(i) The correlation may be due to pure chance, especially in a small sample:

We may get a high degree of correlation between two variables in a sample but in the universe, there may not be any relationship between the variables at all.

This is especially so in case of small samples. Such a correlation may arise either because of pure random sampling variation or because of the bias of the investigator in selecting the sample.

The following example will illustrate the point:

Income(rs)	**20,000**	**25,000**	**30,000**	**35,000**	**40,000**
Weight (lbs)	**120**	**140**	**160**	**180**	**200**

The above data show a perfect positive relationship between income and weight, i.e., as the income is increasing, the weight is increasing and the rate of change between two variables is the same.

(ii) *Both the correlated variables may be influenced by one or more other variables:*

It is just possible that a high degree of correlation between the variables may be due to some causes affecting each variable or different causes affecting each with the same effect.

For example, a high degree of correlation between the yield per acre of rice and tea may be due to the fact that both are related to the amount of rainfall. But none of the two variables is the cause of the other.

Instead, both the variables move together because both are influenced by a third variable.

(iii) *Both the variables may be mutually influencing each other so that neither can be designated as the cause and the other the effect:*

There may be a high degree of correlation between the variables, but it may be difficult to pinpoint as to which is the cause and which is the effect.

For example, such variables as demand and supply, price and production, etc. mutually interact.

As the price of a commodity increases its demand goes down and so price is the cause and demand the effect.

But it is also possible that increased demand of a commodity due to growth of population or other reasons may exercise an upward pressure on price. Now, the cause is the increased demand, the effect the price.

TYPES OF CORRELATION

Correlation is described or classified in several different ways. Three of the most important ways of classifying correlation are as follows:

1. Positive or negative correlation.
2. Simple and multiple correlation
3. Linear and non-linear (Curvilinear)

Positive and Negative Correlation

- **Whether correlation is positive (direct) or negative (inverse) would depend upon the direction of change of the variables.**
- **If both the variables are varying in the same direction, i.e., if as one variable is increasing the other, on an average, is also increasing or, if as one variable is decreasing, the other, on an average, is also decreasing, correlation is said to be positive.**
- **The on the other hand, the variables are varying in opposite directions, i.e., as one variable is increasing, the other is decreasing or vice versa, correlation is said to be negative.**

following examples would illustrate the difference between positive and negative correlation.

I. Positive Correlation

X : 10 12 15 18 20

Y : 15 20 22 25 37

II. Negative Correlation

X : 80 70 60 40 30

Y : 50 44 30 20 10

III. Positive Negative Correlation

X : 20 30 40 60 80

Y : 40 30 22 15 10

X : 100 90 60 40 30

Y : 10 20 30 40 50

Simple and Multiple Correlation

- **The distinction between simple and multiple correlation is based upon the number of variables studied.**
- **When only two variables are studied, it is a problem of simple correlation.**
- **When three or more variables are studied, it-is a problem of either multiple or partial correlation.**
- **In multiple correlation, three or more variables are studied simultaneously.**
- **For example, when we study the relationship between the yield of rice per acre and both the amount of rainfall and the amount of fertilizers used, it is a problem of multiple correlation.**

Linear and Non-Linear (Curvilinear) Correlation

- **The distinction between linear and non-linear correlation is based upon the constancy of the ratio of change between the variables.**
- **If the amount of change in one variable tends to bear constant ratio to the amount of change in the other variable, then the correlation is said to be linear.**

For example, observe the following two variables X and Y:

X: 10 20 30 40 50

Y: 70 140 210 280 350

- **The ratio of change between the two variables is the same.**
- **If such variables are plotted on a graph paper, all the plotted points would fall on a straight line.**
- **Correlation would be called non-linear or curvilinear if the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable.**

For example, if we double the amount of rainfall, the production of rice or wheat, etc., would not necessarily be doubled. we find a non-linear relationship between the variables.

METHODS OF STUDYING CORRELATION

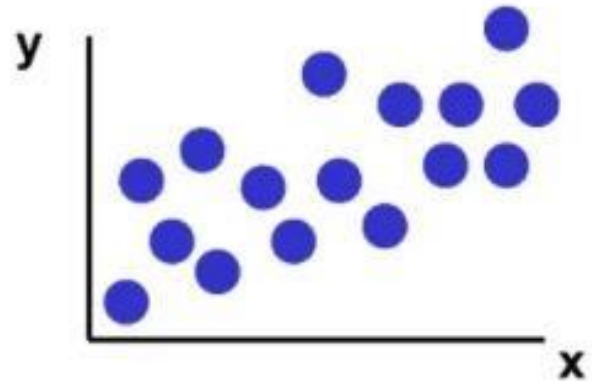
The various methods of ascertaining whether two variables are correlated or not are:

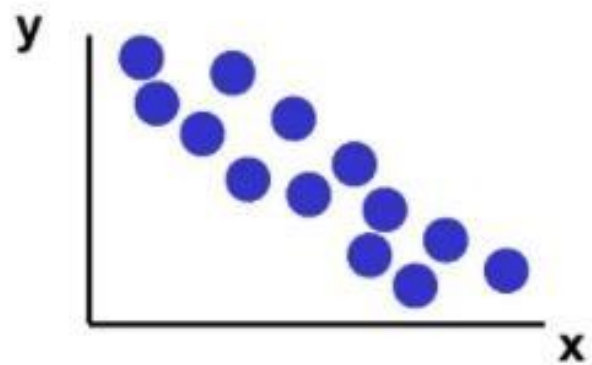
- **Scatter Diagram Method**
- **Graphic Method**
- **Karl Pearson's Coefficient of Correlation**
- **Concurrent Deviation Method**
- **Method of Least Squares**

- **The first two are based on the knowledge of diagrams and graphs.**
- **whereas the others are the mathematical methods.**

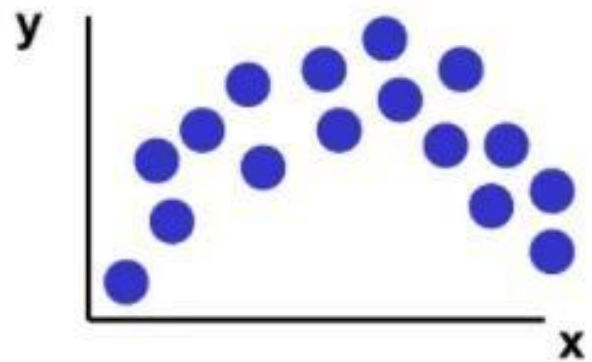
Scatter Plot Examples

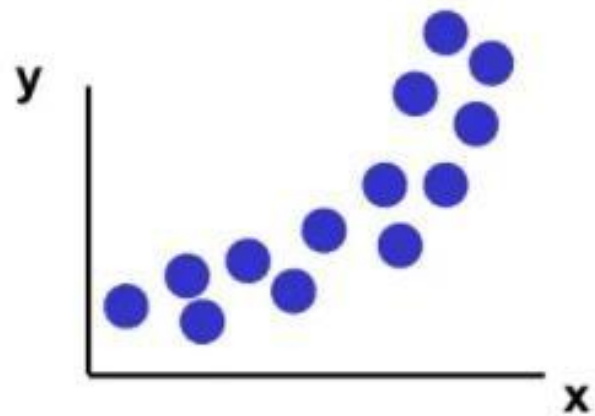
Linear relationships





Curvilinear relationships

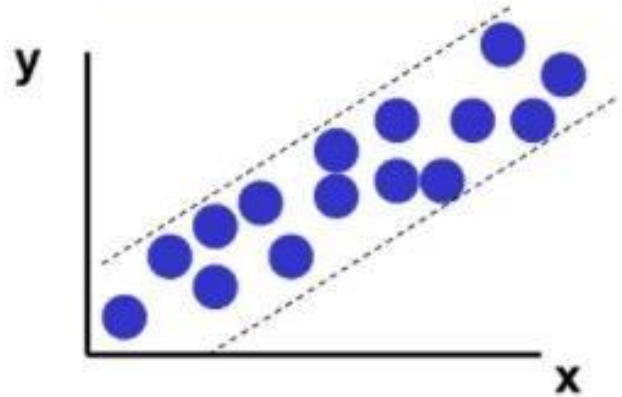


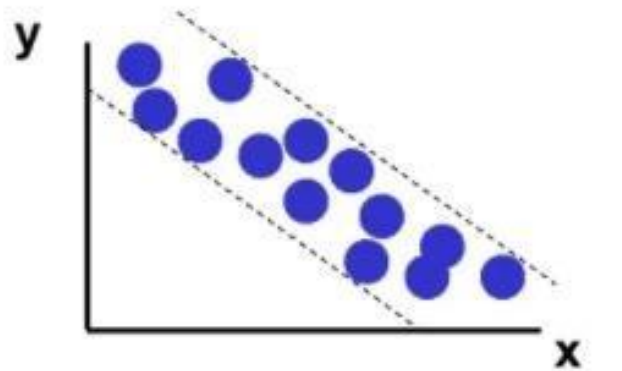


Scatter Plot Examples

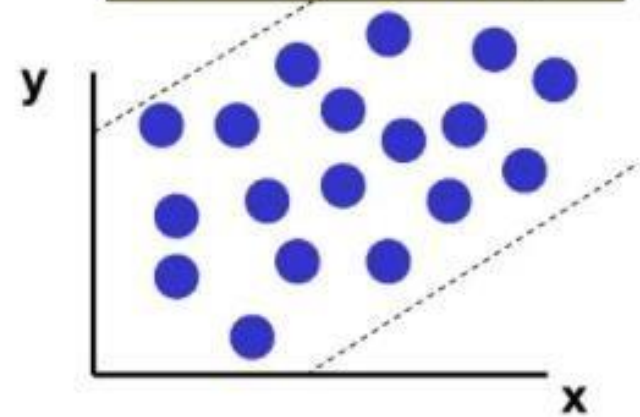
(continued)

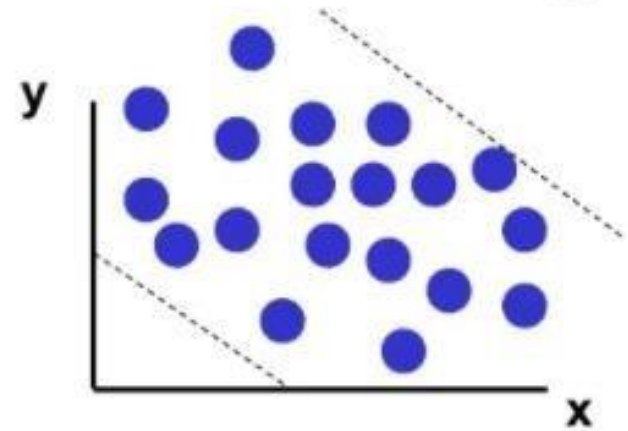
Strong relationships





Weak relationships

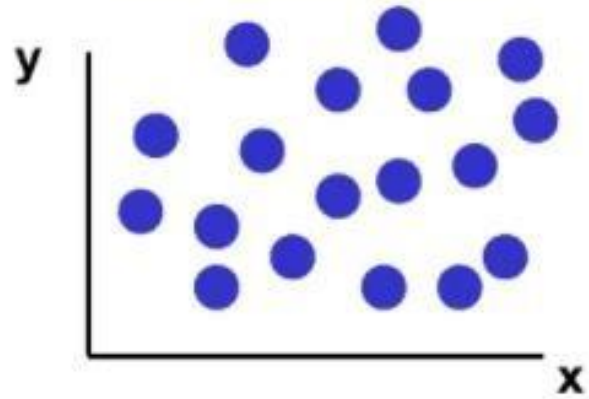


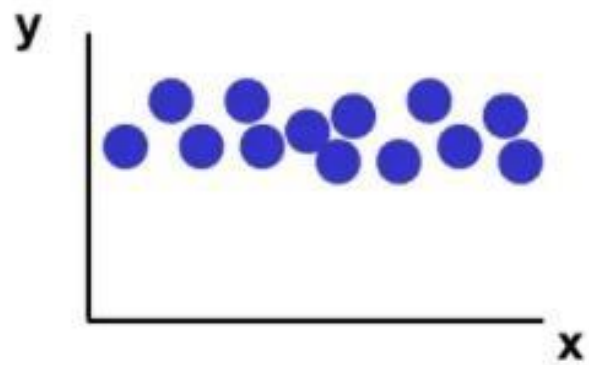


Scatter Plot Examples

(continued)

No relationship





Merits and Limitations of the Method

Merits:

- **It is a simple and non-mathematical method of studying correlation between the variables.**
- **As such it can be easily understood, and a rough idea can very quickly be formed as to whether or not the variables are related.**
- **It is not influenced by the size of extreme items whereas most of the mathematical methods of finding correlation are influenced by extreme items.**
- **Making a scatter diagram usually is the first step in investigating the relationship between two variables.**
- **If the variables are related, we can see what kind of line, or estimating equation describes this relationship.**

Limitations:

- **cannot establish the exact degree of correlation between the variables as is possible by applying the mathematical methods.**

KARL PEARSON'S COEFFICIENT OF CORRELATION

A mathematical method for measuring the intensity or the magnitude of linear relationship between two variable series was suggested by Karl Pearson (1867-1936), is most widely used in practice.

The Pearson coefficient of correlation is denoted by the symbol r.

It is one of the very few symbols that is used universally for describing the degree of correlation between two variables.

The formula for computing Pearsonian r is:

$$r = \sum xy \ / N\sigma_x\sigma_y \qquad \text{... (i)}$$

$$x = (X - \overline{X});\ \ y = (Y - \overline{Y})$$

Here,

σ_x= Standard deviation of series X

σ_y= Standard deviation of series Y

N = Number of pairs of observations

r = The (product moment) correlation coefficient

- **The value of the coefficient of correlation as obtained by the formula will always lie between +1& -1.**
- **When r = +1, it means that there is perfect positive correlation between the variables.**
- **When r = -1, it means that there is perfect negative correlation between the variables.**
- **When 0, it means there is no relationship between the two variables.**
- **However, in practice such values of r as +1,-1, and 0 are rare.**
- **values which lie between +1 and - 1 such as +0.8, -0.26, etc.**
- **The coefficient of correlation describes not only the magnitude of correlation but also its direction.**
- **Thus, +0.8 would mean that correlation is positive because the sign of r is +**

and the magnitude of correlation is 0.8. Similarly, -0.26 means low degree of negative correlation.

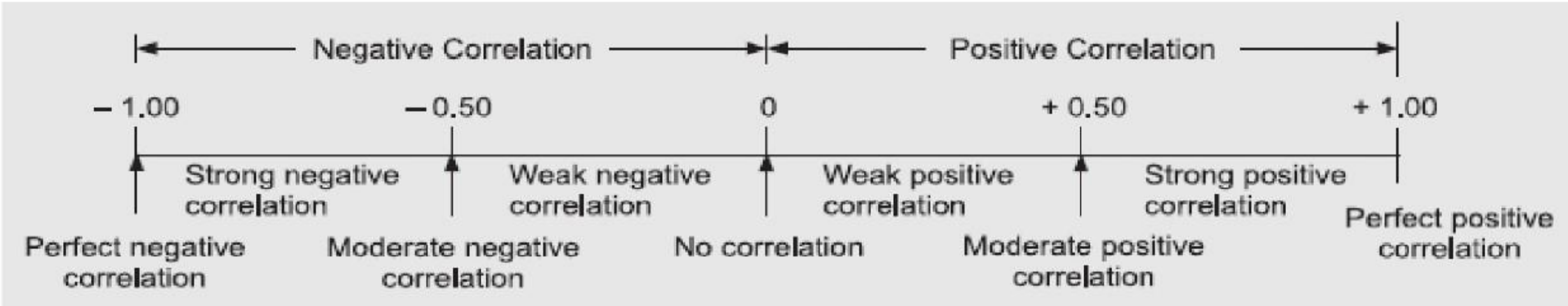
The above formula for computing Pearson's coefficient of correlation can be transformed to the following form which is easier to apply:

Where:

$$r = \Sigma xy / \sqrt{\sum x^2 \times \Sigma y^2}$$

.. (ii)

$$x = (X - \overline{X}) \text{ and } y = (Y - \overline{Y})$$


Negative Correlation
Positive Correlation
– 1.00
– 0.50
0
+ 0.50
+ 1.00
Strong negative correlation
Weak negative correlation
Weak positive correlation
Strong positive correlation
Perfect negative correlation
Moderate negative correlation
No correlation
Moderate positive correlation
Perfect positive correlation

The coefficient of correlation is said to be a measure of covariance between two series. The covariance of two series X and Y is written as:

Covariance = $\sum xy$ / N

where x and y stand for deviations of X and Y series from their respective means.

In order to find out the value of correlation coefficient, first we calculate covariance and then in order to convert it to a relative measure, we divide the covariance by the standard deviation of the two series. The ratio so obtained is called Karl Pearson's coefficient.

$r = \Sigma xy / N$, $\sigma_x = \sqrt{\Sigma x^2 / N}$, $\sigma_y = \sqrt{\Sigma y^2 / N}$

$r = \Sigma xy / \sqrt[N]{\Sigma x^2 / N \times \Sigma y^2 / N}$

$r = \Sigma xy / \sqrt{\Sigma x^2 \times \Sigma y^2}$

$$r = \text{cov}(X,Y) / \sigma_x \sigma_y$$

Where: cov (X,Y) = the covariance of variables X and Y

$$= 1/N \sum(X - \overline{X})(Y - \overline{Y})$$

$$= \sum xy / N$$

N = Number of pairs of observations

σ_x = Standard deviation of series X

$$\sigma x = \sqrt{(x - \overline{x})^2 / N}$$

σ_y = Standard deviation of series Y

$$\sigma y = \sqrt{(y - \overline{y})^2 / N}$$

so

$$r = 1/N \sum(x - x)(y - y) / \sqrt{(x - x)^2 / N} \sqrt{(y - y)^2 / N}$$

By simplifying

$$r = N\sum x\, y - \sum x \sum y / \sqrt{N \sum x^2 - (\sum x)^2}\ \sqrt{N \sum y^2 - (\sum y)^2}$$

Example:

Calculate Karl Pearson's coefficients of correlation from the following data and interpret the value.

Roll Number of students :	**1**	**2**	**3**	**4**	**5**
Marks in QA :	**48**	**35**	**17**	**23**	**47**
Marks in ML :	**45**	**20**	**40**	**25**	**45**

Example:

Making use of the data summarized below, calculate the coefficient of correlation, $r_{12.}$

Case :	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
X1 :	**10**	**6**	**9**	**10**	**12**	**13**	**11**	**9**
X2 :	**9**	**4**	**6**	**9**	**11**	**13**	**8**	**4**

Example:

From the following table calculate the coefficient of correlation by Karl Pearson's method.

X:	**6**	**2**	**10**	**4**	**8**
Y:	**9**	**11**	**?**	**8**	**7**

Arithmetic means of X and Y series are 6 and 8 respectively.

INTRODUCTION

After having established the fact that two variables are closely related, we may be interested in estimating (predicting) the value of one variable given the value of another.

For example:

advertising and sales are correlated, we find out expected amount of sales for a given advertising expenditure.

Similarly, the yield of rice and rainfall are closely related, we may find out the amount of rain required to achieve a certain production figure.

Regression analysis reveals average relationship between two variables and this makes possible estimation or prediction.

The dictionary meaning of the term 'regression' is the act of returning or going back.

The term 'regression' was first used by Sir Francis Galton (1822-1911), an English Scientist, in 1877 while studying the relationship between the height of fathers and sons.

What is Regression? Regression is a method to determine the statistical relationship between a dependent variable and one or more independent variables. The change dependent variable is associated with the change in the independent variables.

For example: that two variables, price (X) and demand (Y), are closely related, we can find out the most probable value of X for a given value of Y or the most probable value of Y for a given value of X.

The regression analysis attempts to accomplish the following:

1. **Regression analysis provides estimates of values of the dependent variable from values of the independent variable.**
2. **A second goal of regression analysis is to obtain a measure of the error involved in using the regression line as a basis for estimation.**
3. **With the help of regression coefficients, we can calculate the correlation coefficient. The square of correlation coefficient (r^2), called coefficient of determination, measures the degree of association of correlation that exists between the two variables.**

LINES OF REGRESSION

If we take the case of two variables X and Y, we shall have two regression lines as the regression of X on Y and the regression of Y on X. The regression line of Y on X gives the most probable values of Y for the given values of X and the regression line of X on Y gives the most probable values of X for given values of Y.

However, when there is either perfect positive or perfect negative correlation between the two variables ($r = \pm 1$), the regression lines will coincide, i.e., we shall have only one line.

REGRESSION EQUATIONS

Regression equations, also known as estimating equations, are algebraic expressions of the regression lines. Since there are two regression lines, there are two regression equations-the regression equation of X on Y is used to describe the variations in the values of X for the given changes in Y and the regression equation of Y on X is used to describe the variation in the values of Y for given changes in X.

Regression Equation of Y on X

The regression equation of Y on X is expressed as follows

$$Y = a + b\,X$$

It may be noted that in this equation, 'Y' is a dependent variable, i.e., its value depends on X. 'X' is independent variable, i.e., we can take a given value of X and compute the value of Y.

Regression Equation of X on Y

The regression equation of X on Y is expressed as follows

$$X = a + b\,Y$$

It may be noted that in this equation, 'X' is a dependent variable, i.e., its value depends on Y. 'Y' is independent variable, i.e., we can take a given value of Y and compute the value of X.

- **'a' is "Y-intercept" because its value is the point at which the regression line crosses the Y-axis, that is, the vertical axis. 'b' is the "slope" of line.**
- **It represents change in Y variable for a unit change in X variable. 'a' and 'b' in the equation are called numerical constants because for any given straight line, their value does not change.**
- **If the values of the constants 'a' and 'b' are obtained, the line is completely determined.**

Method of Least Squares

When the data represent a sample from a large population, the least squares line is the 'best' estimate of the population regression line.

"The general problem of finding equations of approximating curves fit given data is called curve fitting."

With a little algebra and differential calculus, it can be shown that the following two equations, if solved simultaneously, will yield values of the parameters a and b such that the least squares requirement is fulfilled.

Regression Equation of Y on X

$$\Sigma Y = N a + b \Sigma X$$

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

Regression Equation of X on Y

$$\Sigma X = N a + b \Sigma Y$$

$$\Sigma XY = a \Sigma Y + b \Sigma Y^2$$

These equations are known as the normal equations for estimating a and b. The quantities $\sum x$ $\sum x^2$, $\sum y$, $\sum y^2$, $\sum xy$ can be obtained from the given set of n points (x1, y1), (x2, y2), …, (xn, yn), and we can solve the equations simultaneously for a and b.

$$a = (\sum x^2)(\sum y) - (\sum x)(\sum xy) \;/\; n\sum x^2 - (\sum x)^2$$

$$b = n\sum xy - (\sum x)(\sum y) \;/\; n\sum x^2 - (\sum x)^2$$

Substituting these values of a and b in $y = a + bx$, we get the required equation of the line of regression of Y on X.

The equation of the line of regression of Y on X can be obtained in a much more systematic and simplified form in terms of $\bar{X}$ and $\bar{Y}$, σ_x, σ_y and $r = r_{xy}$.

Dividing both sides of $\sum y = na + b\sum x$ by n, the total number of pairs, we get

$$1 / n \sum y = a + b \,.\, 1 / n \sum x$$

$$\bar{y} = a + b\,\bar{x}$$

This implies that line of best fit, i.e., regression of Y on X passes through the point $(\bar{x}, \bar{y})$. Or in other words, the point $(\bar{x}, \bar{y})$ lies on the line of regression of Y on X.

Example :

From the following data, obtain the two regression equations.

X :	**6**	**2**	**10**	**4**	**8**
Y:	**9**	**11**	**5**	**8**	**7**

From $(b = n\sum xy - (\sum x)(\sum y) \,/\, n\sum x^2 - (\sum x)^2)$

we get: $b = \mathrm{Cov}(x, y) / \sigma_x^2$

We find that the equation ($y = a + bx$) is in the slope-intercept form, viz., $y = mx + c$.
Hence b represents the slope of the line of regression of y on x. Further, we have proved in ($\bar{y} = a + b\,\bar{x}$) that this line (i.e., line of regression of Y on X) passes through the point ($\bar{x}, \bar{y}$).

Hence, using the slope-point form of the equation of a line, the required equation of the line of regression of Y on X becomes:

$$y - \bar{y} = b\,(x - \bar{x})$$

Or $$y - \bar{y} = \mathrm{Cov}(x, y) / \sigma_x^2 \,.\, (x - \bar{x})$$

But $$r = r_{xy} = \mathrm{Cov}\,(x, y) / \sigma_x\sigma_y$$

$$\mathrm{Cov}\,(x, y) = r\,\sigma_x\sigma_y$$

Substituting in ($y - \bar{y} = \mathrm{Cov}(x, y) / \sigma_x^2 \,.\, (x - \bar{x})$),

Regression Equation of Y on X

$$y - \bar{y} = r\,\sigma_x\sigma_y \,/\, \sigma_x^2\,(x - \bar{x})$$

$$y - \bar{y} = r\,\sigma_y / \sigma_x\,(x - \bar{x})$$

Regression Equation of X on Y

$$x - \bar{x} = r\,\sigma_x\sigma_y \,/\, \sigma_y^2\,(y - \bar{y})$$

$$x - \bar{x} = r\,\sigma_x / \sigma_y\,(y - \bar{y})$$

$\bar{x}$ is the mean of X series; $\bar{y}$ is the mean of Y series.

COEFFICIENTS OF REGRESSION

$r\ \sigma_y / \sigma_x$ is known as the regression coefficient of Y on X.

$r\ \sigma_x / \sigma_y$ is known as the regression coefficient of X on Y.

Notations:

b_{yx} = Coefficient of regression of Y on X.

b_{xy} = Coefficient of regression of X on Y.

The coefficient of regression of Y on X is given by

$b_{yx} = \text{Cov}(x, y) / \sigma_x^2 = r\ \sigma_y / \sigma_x$ [$\because$ $\text{Cov}(x, y) = r\ \sigma_x\ \sigma_y$]

The coefficient of regression of X on Y is given by

$b_{xy} = \text{Cov}(x, y) / \sigma_y^2 = r\ \sigma_x / \sigma_y$

Accordingly, the equation of the line of regression of Y on X becomes

$$y - \bar{y} = b_{yx} (x - \bar{x})$$

and the equation of the line of regression of X on Y becomes $x - \bar{x} = b_{xy} (y - \bar{y})$

For numerical computations of the equations of line of regression of Y on X, and X on Y, the following formulae for the regression coefficients b_{yx} and b_{xy} are very convenient to use.

$$b_{yx} = \text{Cov}(x, y) / \sigma_x^2$$
$$= \sum (x - \bar{x})(y - \bar{y}) / \sum (x - \bar{x})^2$$
$$= n\sum xy - (\sum x)(\sum y) / n\sum x^2 - (\sum x)^2$$

$$b_{xy} = \text{Cov}(x, y) / \sigma_y^2$$
$$= \sum (x - \bar{x})(y - \bar{y}) / \sum (y - \bar{y})^2$$
$$= n\sum xy - (\sum x)(\sum y) / n\sum y^2 - (\sum y)^2$$

The correlation coefficient is the geometric mean between the regression coefficient.

$$r^2 = b_{xy} \times b_{yx}$$

$$b_{yx} = \text{Cov}(x, y) / \sigma_x^2 = r\,\sigma_y / \sigma_x \quad \text{and} \quad b_{xy} = \text{Cov}(x, y) / \sigma_y^2 = r\,\sigma_x / \sigma_y$$

Multiply them:

We get:

$$r^2 = b_{xy} \times b_{yx}$$

It should be noted that the under root of the product of two regression coefficients gives us the value of correlation coefficient.

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

Ex: From the following data, obtain the two regression equations :

Sales :	**91**	**97**	**108**	**121**	**67**	**124**	**51**	**73**	**111**	**57**
Purchases :	**71**	**75**	**69**	**97**	**70**	**91**	**39**	**61**	**80**	**47**

Ex. From the data given below find :

(a) The two regression coefficients.

(b) The two regression equations.

(c) The coefficient of correlation between the marks in Economics and Statistics.

(d) The most likely marks in Statistics when marks in Economics are 30.

Marks in Economics :	**25**	**28**	**35**	**32**	**31**	**36**	**29**	**38**	**34**	**32**
Marks in Statistics :	**43**	**46**	**49**	**41**	**36**	**32**	**31**	**30**	**33**	**39**

Example :

You are given the following data:

	X	Y
Arithmetic mean	**36**	**85**
Standard Deviation	**11**	**8**

Correlation coefficient between X and Y **0.66**

1) **Find the two Regression Equations.**
2) **Estimate the value of X when Y =75**

Ex: **Two regression lines of a sample are $X + 6Y = 6$ and $3X + 2Y = 10$.**
Find the correlation coefficient.

Ex: **By using the following data, find out the two lines of regression and from them, compute the Karl Pearson's coefficient of correlation.**
$\sum X = 250$, $\sum Y = 300$, $\sum XY = 7900$, $\sum Y^2 = 10000$, $\sum X^2 = 6500$, $N = 10$.

Difference between the Correlation and Regression:

- **Whereas coefficient of correlation is a measure of degree of covariability between X and Y.**
- **The objective of regression analysis is to study the 'nature of relationship' between the variables so that we may be able to predict the value of one on the basis of another.**

- **Correlation is merely a tool to determine the degree of relationship between two variables and, therefore, we cannot say that one variable is the cause and other the effect.**
- **In regression analysis, one variable is taken as dependent while the other as independent, thus making it possible to study the cause and effect relationship.**

- **In correlation analysis, r_{xy} is a measure of direction and degree of linear relationship between two variables X and Y. r_{xy}, and r_{yx}, are symmetric ($r_{xy} = r_{yx}$), i.e., it is immaterial which of X and Y is dependent variable and which is independent variable.**

- **In regression analysis the regression coefficients b_{xy}, and b_{yx}, are not symmetric, i.e., $b_{xy} \neq b_{yx}$, and hence it definitely makes a difference as to which variable is dependent and which is independent.**

- **Correlation coefficient is independent of change of scale and origin.**
- **Regression coefficients are independent of change of origin but not of scale.**

Standard Error of the Estimate

- The standard error of the regression (S), also known as the standard error of the estimate, represents the average distance that the observed values fall from the regression line.
- Conveniently, it tells you how wrong the regression model is on average using the units of the response variable.
- Smaller values are better because it indicates that the observations are closer to the fitted line.

$$S_{yx} = \sigma_y \sqrt{1 - r^2} \qquad \text{or} \qquad S_{yx} = \sqrt{(Y - Y_c)^2}$$

$$S_{xy} = \sigma_x \sqrt{1 - r^2} \qquad \text{or} \qquad S_{xy} = \sqrt{(X - X_c)^2}$$

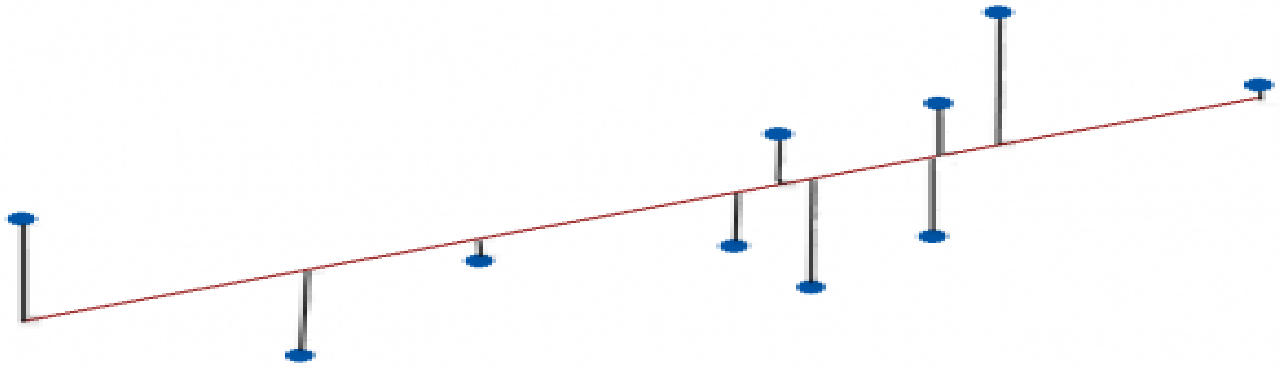
Example :

From the following data, obtain the two regression equations and calculate the standard error of estimates.

X :	**6**	**2**	**10**	**4**	**8**
Y:	**9**	**11**	**5**	**8**	**7**

Example: **In fitting of a regression of Y on X to a bivariate distribution consisting of 9 observations, the explained and unexplained variations were computed as 24 and 36 respectively.**
Find
(i) the coefficient of determination and
(ii) standard error of the estimate of Y on X.

•**The standard error of the regression provides the absolute measure of the typical distance that the data points fall from the regression line. S is in the units of the dependent variable.**
•**R-squared provides the relative measure of the percentage of the dependent variable variance that the model explains. R-squared can range from 0 to 100%.**



As R-squared increases and S decreases, the data points move closer to the line

formula (3)

$$b_{xy} = \frac{cov(x,y)}{(\sigma y)^2} \quad \text{or} \quad b_{yx} = \frac{cov(x,y)}{(\sigma x)^2}$$

use any one formula coeff of regression eqn of y on x / x on y.

(eg) in the estimation of regression eq^n of two variable x & y, the following results were obtained.

$\Sigma x = 900$, $\Sigma y = 700$, $\Sigma xy = 10$;

& $\Sigma x^2 = 6390$ & $\Sigma y^2 = 2860$, $\Sigma xy = 3900$.

where x & y are deviations from respective means. Obtain the two regression eq^n.

Soln: The coefficients of regression of y on x & x on y are given by.

$$b_{yx} = \frac{cov(x,y)}{(\sigma x)^2} = \frac{\Sigma (x-\bar{x})(y-\bar{y})}{\Sigma (x-\bar{x})^2}$$

$$= \frac{\Sigma xy}{\Sigma x^2} = \frac{3900}{6390}$$

$$= 0.6132$$

& $b_{xy} = \frac{cov(x,y)}{(\sigma y)^2}$

$= \frac{\Sigma xy}{\Sigma y^2} = \frac{3900}{2860}$

$\boxed{b_{xy} = 1.3636}$

$\bar{x} = \frac{\Sigma x}{n} = \frac{900}{10} = 90$

$\bar{y} = \frac{\Sigma y}{n} = \frac{700}{10} = 70$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{900}{10} = 90$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{700}{10} = 70$$

① Req eqn of y on x

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 70 = 0.6132(x - 90)$$

$$y = 0.6132x - 0.6132(90) + 70$$

$$\boxed{y = (0.6132x) + 19.812}$$

(2) Req eqn of x on y.

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

~~$\bar{x} = 90$~~

$$x - 90 = 1.3636(y - 70)$$

$$x = 1.3636y - 1.3636(70) + 90$$

$$\boxed{x = 1.3636y - 5.452}$$

(eg) The regression lines of a sample are $x + 6y = 6$ & $3x + 2y = 10$

find (i) sample mean $\bar{x}$ & $\bar{y}$

(ii) the coefficient of correlation between x & y

(iii) find the value of y at $x = 12$

Soln : The regression line pass thro
the point $(\bar{x}, \bar{y})$.

So, the regression lines of as sample are

$\bar{x} + 6\bar{y} = 6$ $\quad 3\bar{x} + 2\bar{y} = 10$.

$x + 6y = 6$ ① ; $3x + 2y = 10$ ⑪

multiply eqⁿ ① by 3.

$$3x + 18y = 18$$

$$\begin{array}{r} 3x + 18y = 18 \\ 3x + 2y = 10 \\ - \quad - \quad - \\ \hline 16y = 8 \end{array}$$

$$\boxed{y = \frac{1}{2}} \text{ or } \boxed{y = \frac{1}{2}}$$

Substituting the value of y in eq^n (i)

$$x + 6y = 6$$

$$x + \overset{3}{\cancel{6}} \times \frac{1}{\cancel{x}} = 6$$

$$x + 3 = 6$$

$$\boxed{x = 3} \quad \text{or} \quad \boxed{\overline{x} = 3}$$

(ii) consider the line :

let $x+6y=6$ be the regression line of y on x. So,

$$6y = 6 - x$$

$$= \frac{6}{6} - \frac{x}{6}$$

$$y = 1 - \frac{1}{6}x$$

$$b_{yx} = -\frac{1}{6}$$

x on y

consider the line:

$$3x+2y=10$$

$$3x=10-2y$$

$$x=\frac{10}{3}-\frac{2}{3}y$$

$$b_{xy}=-\frac{2}{3}$$

(iii) from eqn (ii) at $x = 12$

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

$$\therefore y = -\frac{1}{6}(12^2) + 1$$

$$y = -2 + 1$$

$$\boxed{y = -1}$$

(eg) After 9/11 attack on world trade center, a company could partially recover the following record on analysis of correlation.

& Variance of $X = 9$

Regression eq^n : $8x - 10y + 66 = 0$

$40x - 18y = 214$

Find on the basis of the above information.

i) mean values of X & Y.

ii) coefficient of correlation bet^n X & Y and

iii) Standard deviation of Y.

iii) Standard deviation

Sol^{n} : i) Finding the mean values of x & y.

$8x - 10y = -66$ — (i)

~~multiply by 5~~

multiply eq^{n} (i) by 5

$40x - 50y = -330$.

$40x - 18y = 214$

$- \quad + \quad -$

$-32y = -544$

$y = \frac{-544}{-32}$

$y = 17$ or $\overline{y = 17}$

Substituting the value of y in eqn (1)

~~8x - 10x~~

$$8x - 10 \times 17 = -66$$

$$8x = -66 + 170$$

$$8x = 104$$

$$x = \frac{104}{8}$$

$$\boxed{x = 13} \text{ or } \boxed{\overline{x} = 13}$$

(ii) For finding out the correlation coefficient we will have to find out the regression coefficient. Since we donot know which of the two regression eqn is the eqn of X on Y, we make an assumption.

let us take eqn (i) as the regression eqn of X on Y.

$$8X = -66 + 10Y$$

$$X = \frac{-66}{8} + \frac{10}{8}Y$$

$$b_{xy} = \frac{10}{8} = 1.25$$

from eqn (ii) we can calculate

$$b_{yx} = 40x - 18y = 214$$

$$-18y = 214 - 40x$$

$$y = \frac{214}{18} + \frac{40}{18}x$$

$$\boxed{b_{yx} = 2.22} \quad \text{or} \quad \boxed{b_{yx} = \frac{40}{18}}$$

Since both the regression coefficient are exceeding 1. our assumption is wrong

Hence, the 1st eqn is eqn of y on x

from eqn (i) $-10y = -8x - 66$

$$y = \frac{-8}{10}x + \frac{66}{-10}$$

$$y = -\frac{8}{10}x + 6.6$$

$$\boxed{b_{yx} = \frac{8}{10}} \quad \text{or} \quad \boxed{0.8}$$

from eqn (ii) $b x y = \frac{18}{40}$

$= 0.45$

i.e $40x - 18y = 214$

$40x = 214 + 18y$

$x = \frac{214}{40} + \frac{18}{40} y$

$b x y = \frac{18}{40}$ or 0.45

$$r = \sqrt{0.8 \times 0.45}$$

$$= \sqrt{0.36}$$

$$\boxed{r = 0.6}$$

(iii) SD of Y: $\sigma_x = \sqrt{9}$

$= 3$

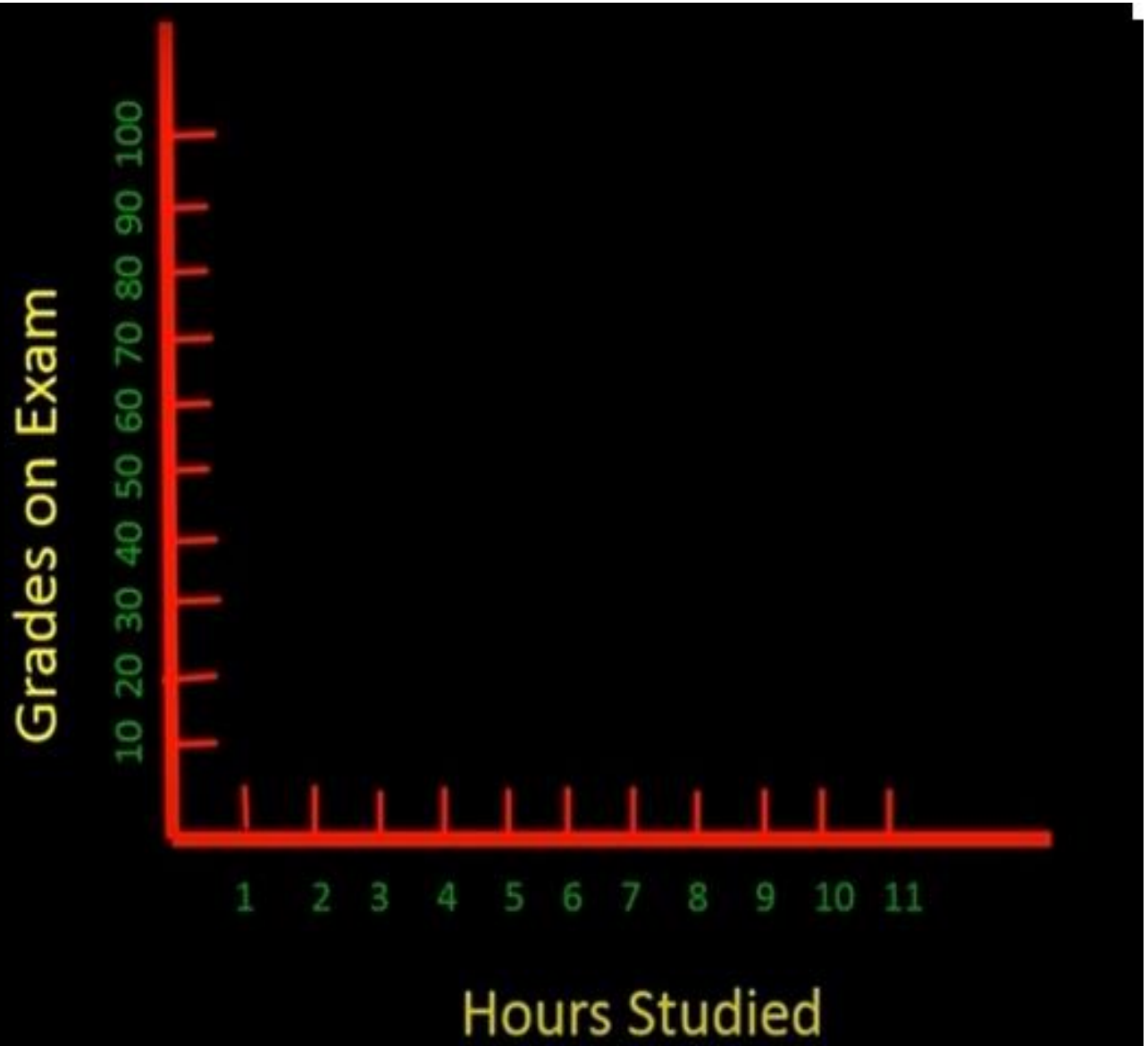
$$b_{xy} = r\frac{\sigma_x}{\sigma_y} \Rightarrow 0.45 = 0.6\frac{3}{\sigma_y}$$

or $0.45\sigma_y = 1.8$ or $\sigma_y = \frac{1.8}{0.45} = 4$

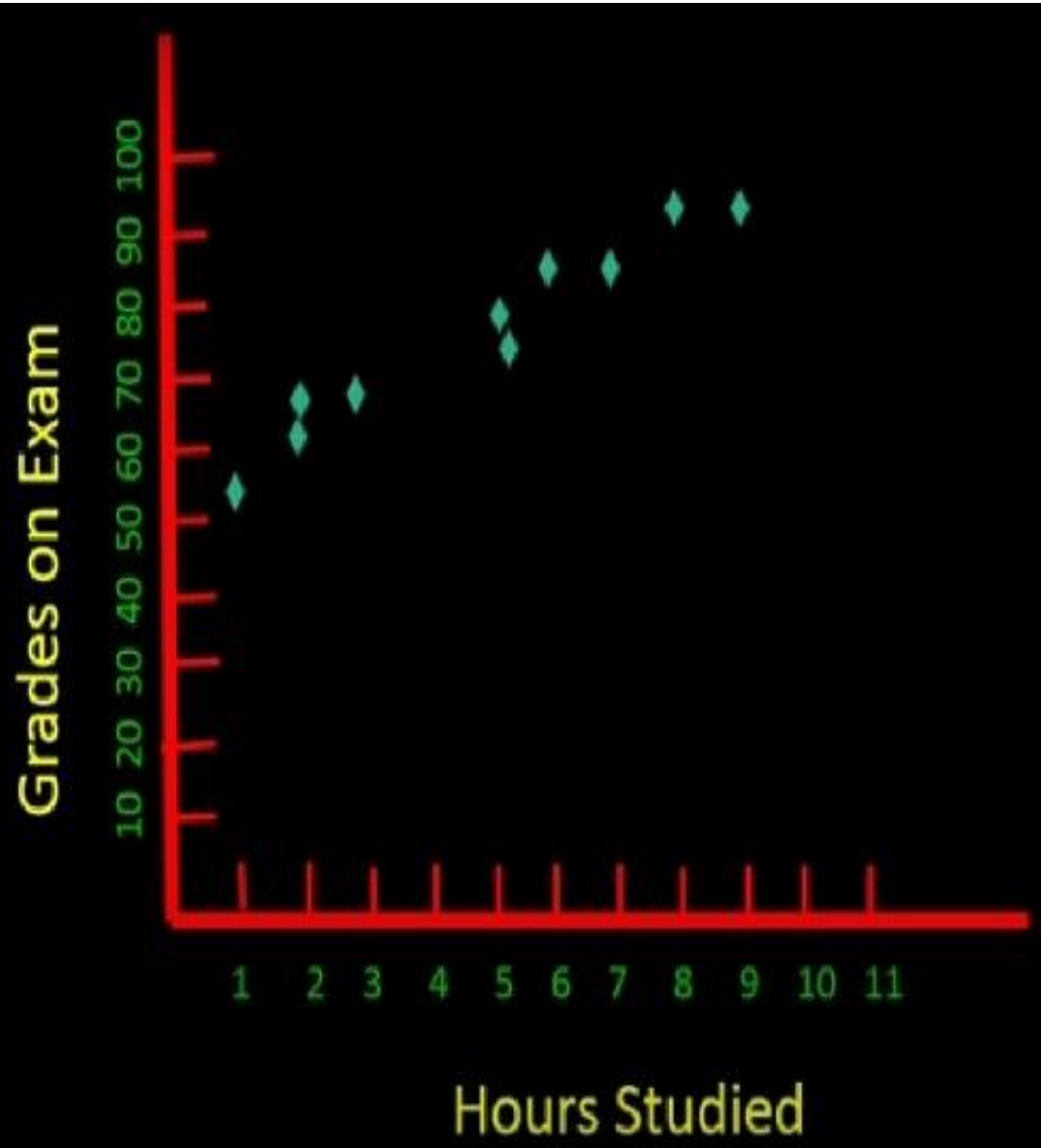
$\boxed{\text{SD of } Y = 4}$

The R-squared value tells us the proportion of variance of the response variable that is explained by the explanatory variable in a regression model

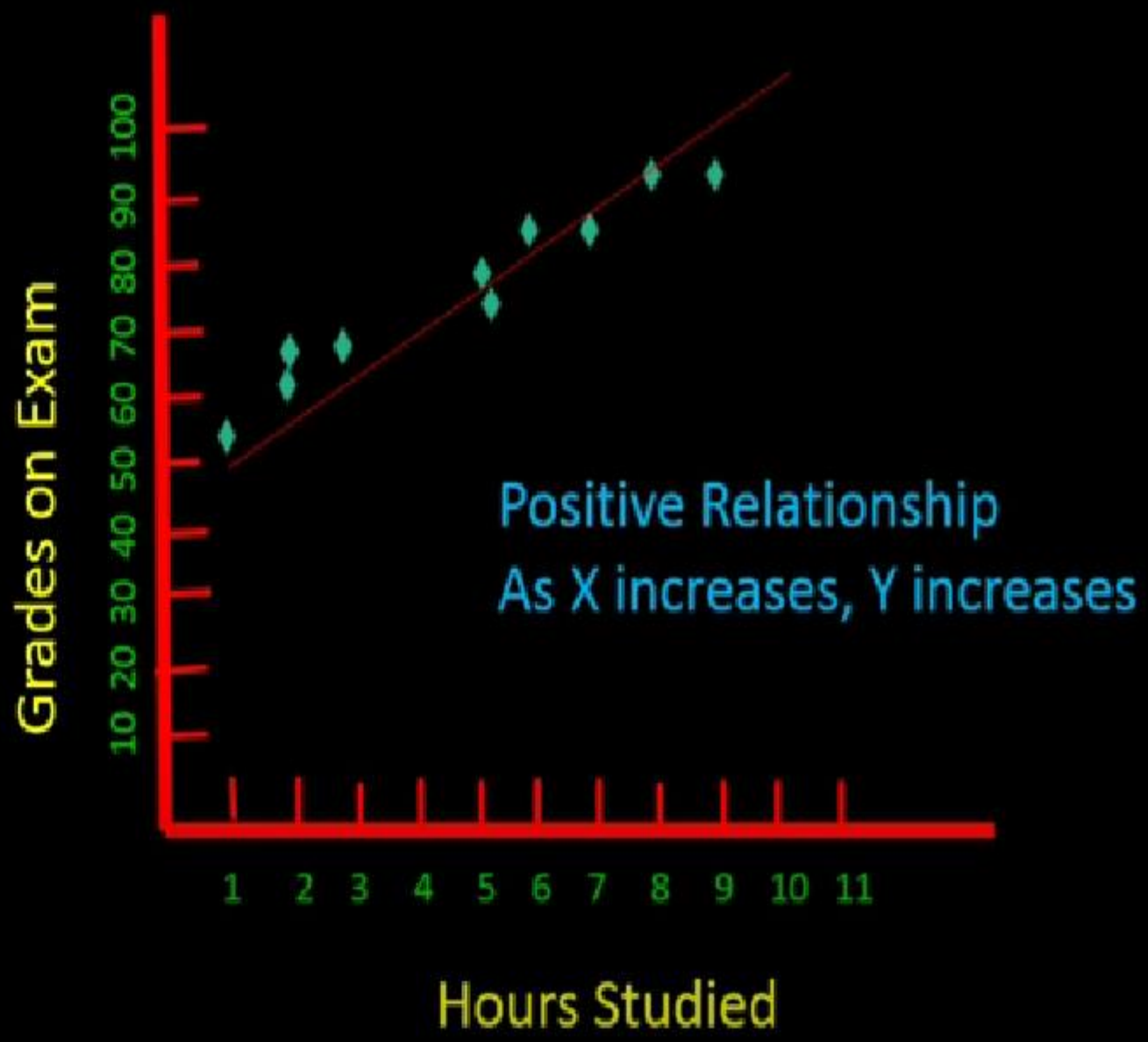
Hours Studied	Grade on Exam
2.00	69.00
9.00	98.00
5.00	82.00
5.00	77.00
3.00	71.00
7.00	84.00
1.00	55.00
8.00	94.00
6.00	84.00
2.00	64.00



Hours Studied	Grade on Exam
2.00	69.00
9.00	98.00
5.00	82.00
5.00	77.00
3.00	71.00
7.00	84.00
1.00	55.00
8.00	94.00
6.00	84.00
2.00	64.00



Hours Studied	Grade on Exam
2.00	69.00
9.00	98.00
5.00	82.00
5.00	77.00
3.00	71.00
7.00	84.00
1.00	55.00
8.00	94.00
6.00	84.00
2.00	64.00



Least Squares Method:

$$\min \Sigma(y_i - \hat{y}_i)^2$$

where:

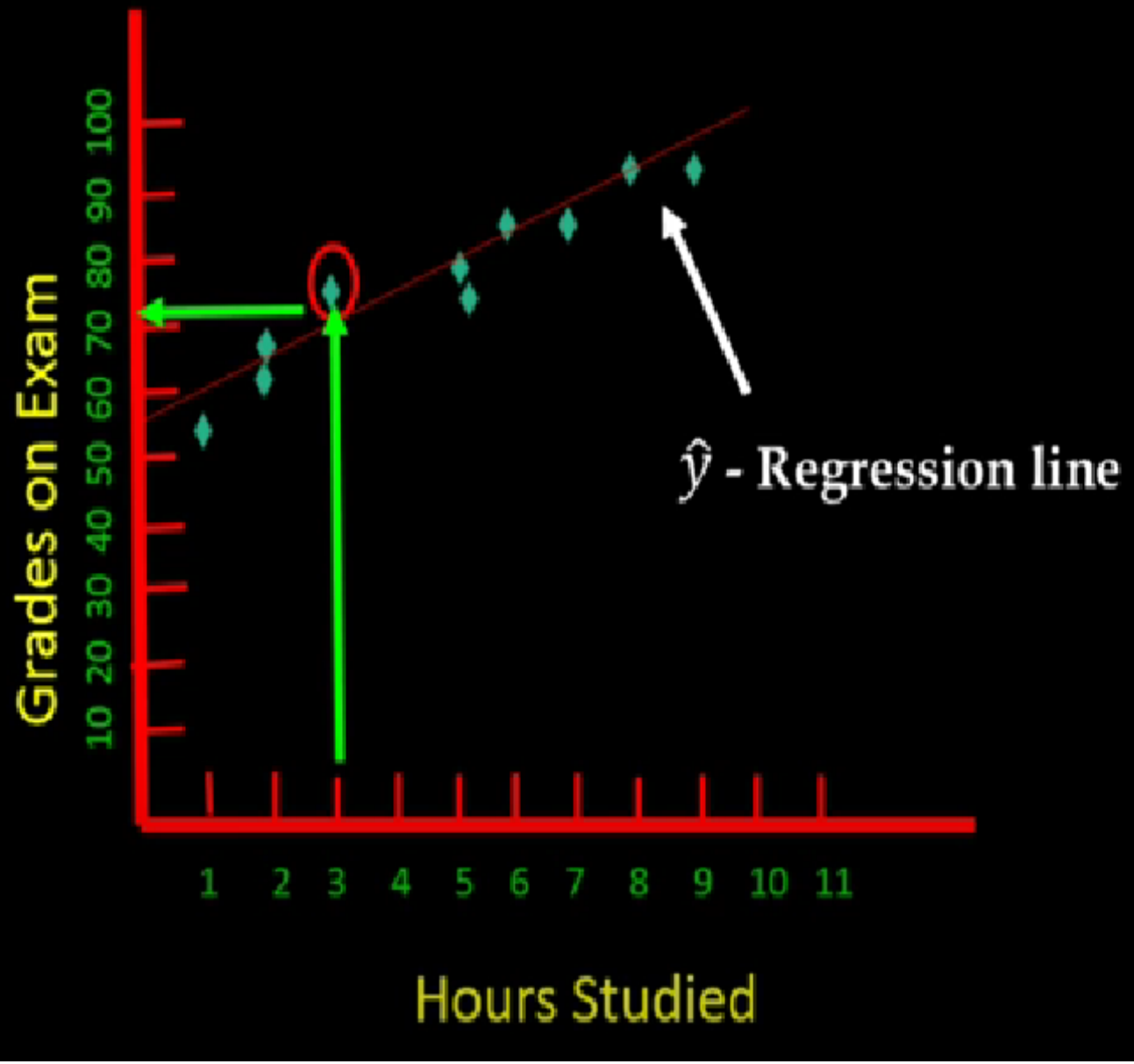
y_i = observed value of the dependent variable for the *i*th observation

$\hat{y}_i$ = predicted value of the dependent variable for the *i*th observation

Example: x=3 hours studied

$\hat{y}_i$ = approx. 69

y_i = 71



Least Squares Method:

$$\min \sum (y_i - \hat{y}_i)^2$$

where:

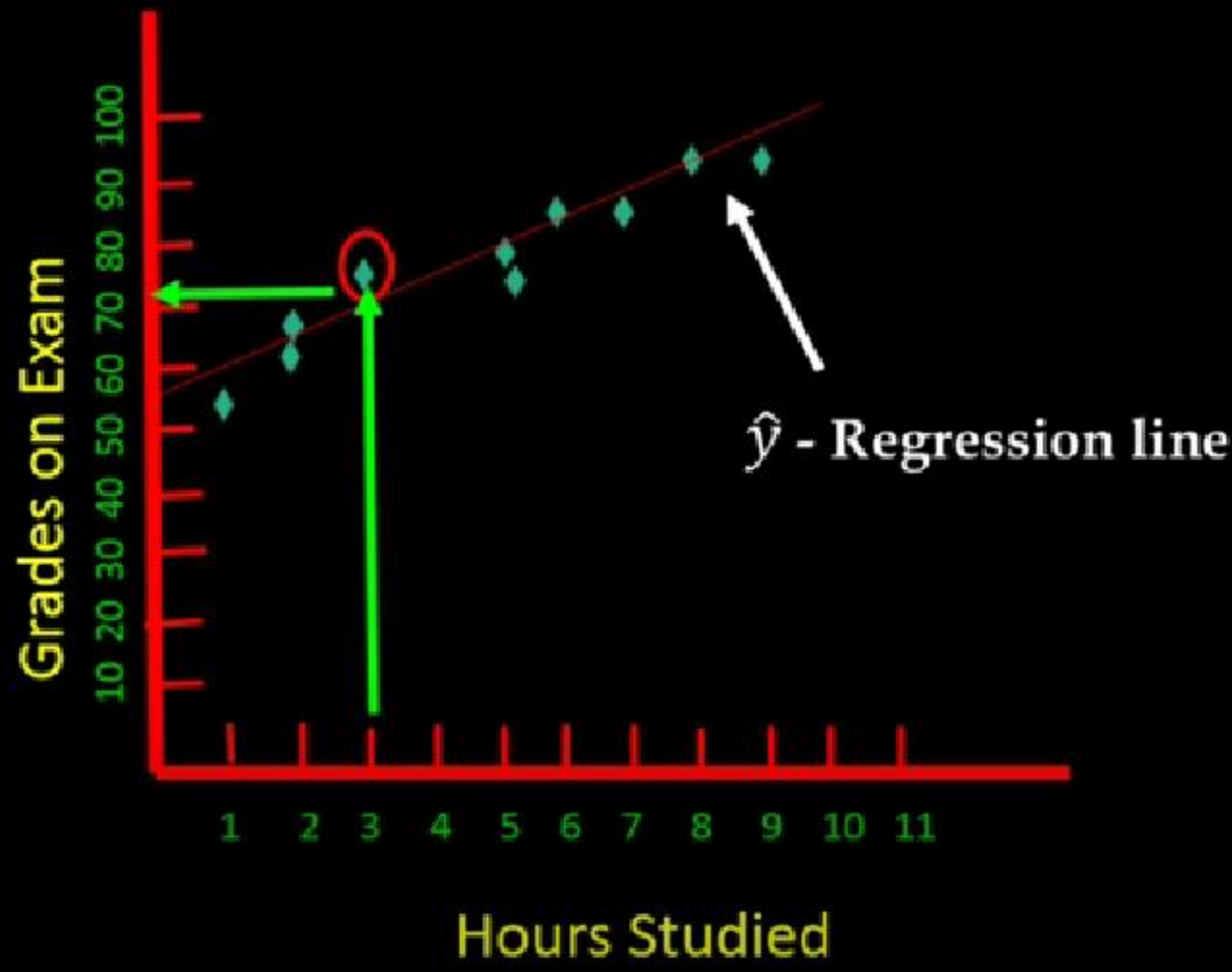
y_i = observed value of the dependent variable for the *i*th observation

$\hat{y}_i$ = predicted value of the dependent variable for the *i*th observation

Example: x=3 hours studied

$\hat{y}_i$ = approx. 69

y_i = 71



minimize sum of the squares of the deviations between observed and predicted

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
2	69	-2.8	-8.8	24.64	7.84
9	98	4.2	20.2	84.84	17.64
5	82	.2	4.2	.84	.04
5	77	.2	-.8	-.16	.04
3	71	-1.8	-6.8	12.24	3.24
7	84	2.2	6.2	13.64	4.84
1	55	-3.8	-22.8	86.64	14.44
8	94	3.2	16.2	51.84	10.24
6	84	1.2	6.2	7.44	1.44
2	64	-2.8	-13.8	38.64	7.84
$\Sigma x_i = 48$	$\Sigma y_i = 778$			320.6	67.6
$\bar{x} = 48/10$	$\bar{y} = 778/10$			$\Sigma(x_i - \bar{x})(y_i - \bar{y})$	$\Sigma(x_i - \bar{x})^2$
$= 4.8$	$= 77.8$				

Coefficient of Determination:

How well does the regression line fit the data?

$$r^2 = \text{SSR}/\text{SST}$$

where:

SSR = sum of squares due to regression = $\sum(\hat{y}_i - \bar{y})^2$

SST = total sum of squares = $\sum(y_i - \bar{y})^2$

SSE = sum of squares due to error = $\sum(y_i - \hat{y}_i)^2$

$$\text{SST} = \text{SSR} + \text{SSE}$$

x_i	y_i	Predicted Grades $\hat{y}_i = 55.048 + 4.74x_i$	Error $y_i - \hat{y}_i$	Squared Error $(y_i - \hat{y}_i)^2$	Deviation $y_i - \bar{y}$	Squared Deviation $(y_i - \bar{y})^2$
2	69	64.528	4.472	19.9988	-8.8	77.44
9	98	97.708	.292	.0852	20.2	408.04
5	82	78.748	3.252	10.5755	4.2	17.64
5	77	78.748	-1.748	3.0555	-.8	.64
3	71	69.268	1.732	2.9998	-6.8	46.24
7	84	88.228	-4.228	17.8759	6.2	38.44
1	55	59.788	-4.788	22.9249	-22.8	519.84
8	94	92.968	1.032	1.0650	16.2	262.44
6	84	83.488	.512	.2621	6.2	38.44
2	64	64.528	-.528	.2788	-13.8	190.44

SSE =79.1215

SST = 1599.6

Coefficient of Determination:

$r^2 = SSR/SST$ $= 1520.4785/1599.6$ $= .9505$

SST = SSR + SSE

SSR = SST - SSE

SSR = 1599.6 - 79.1215

= 1520.4785

r^2 = percent of variability in y can be explained by x

r^2 = 95.05% of the variability in grades can be explained by the number of hours studied

$$\hat{y} = b_0 + b_1 x$$

$$\hat{y} = 55.048 + 4.74x$$

$\bar{y} = 778/10 = 77.8$

$r^2 = \text{SSR}/\text{SST}$

$\text{SST} = \sum(y_i - \bar{y})^2$

$\hat{y} = 55.048 + 4.74(7)$

$\hat{y} = 88.228$

$\text{SSR} = \sum(\hat{y}_i - \bar{y})^2$

$\text{SSE} = \sum(y_i - \hat{y}_i)^2$

r^2 = Explained Variation / Total Variation