UNIT II
Regression

Programme Educational Objectives

Our program will create graduates who:

1. *Will be recognized as a creative and an enterprising team leader.*
2. *Will be a flexible, adaptable and an ethical individual.*
3. *Will have a holistic approach to problem solving in the dynamic business environment.*

Research Methodology & Quantitative Techniques Course Outcomes

CO1-Given a managerial problem and associated frequency distribution data, the student manager will be able to apply descriptive and inferential statistics to facilitate quick and rationale managerial decision making.

CO2-Given the data for two or more variables, the student manager will be able to estimate the strength of the relationship between two variables using 'Karl Pearson' and 'Spearman's Rank' correlation coefficient.

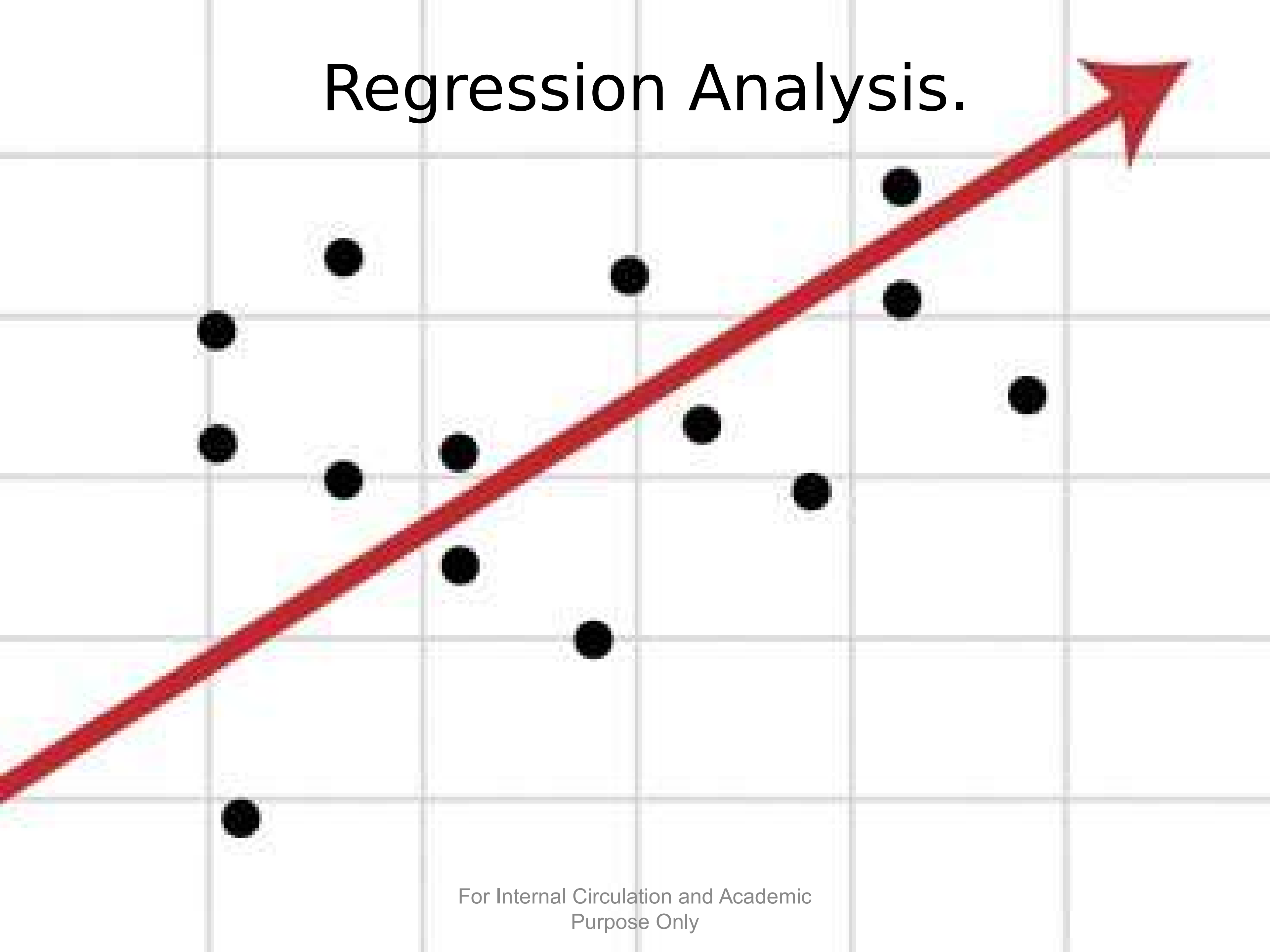
CO3-Given the data for two or more variables, the student manager will be able to predict / forecast using as moving averages, regression and time series analysis.

CO4-Given a managerial problem, the student manager will be able to formulate it as 'research problem' and also will be able to suggest suitable research methodology to identify workable solutions.

CO5-Given a business Problem/situation, the student manager will be able to develop methods and instruments (questionnaire/ interview schedule) for collection and measurement of qualitative as well as quantitative data using primary and secondary sources from a given sampling framework.

CO6-Given the sample statistics, the student manager will be able to apply Z, t and Chi-square tests to accept or reject the stated hypotheses for making sound decisions.

Regression Analysis.



Learning Objectives

The main **objective** of **regression analysis** is to explain the variation in one variable (called the dependent variable), based on the variation in one or more other variables (called the independent variables).

Correlation & Regression coefficients.

$$r = \sqrt{b_{xy} \, X \, b_{yx}}$$

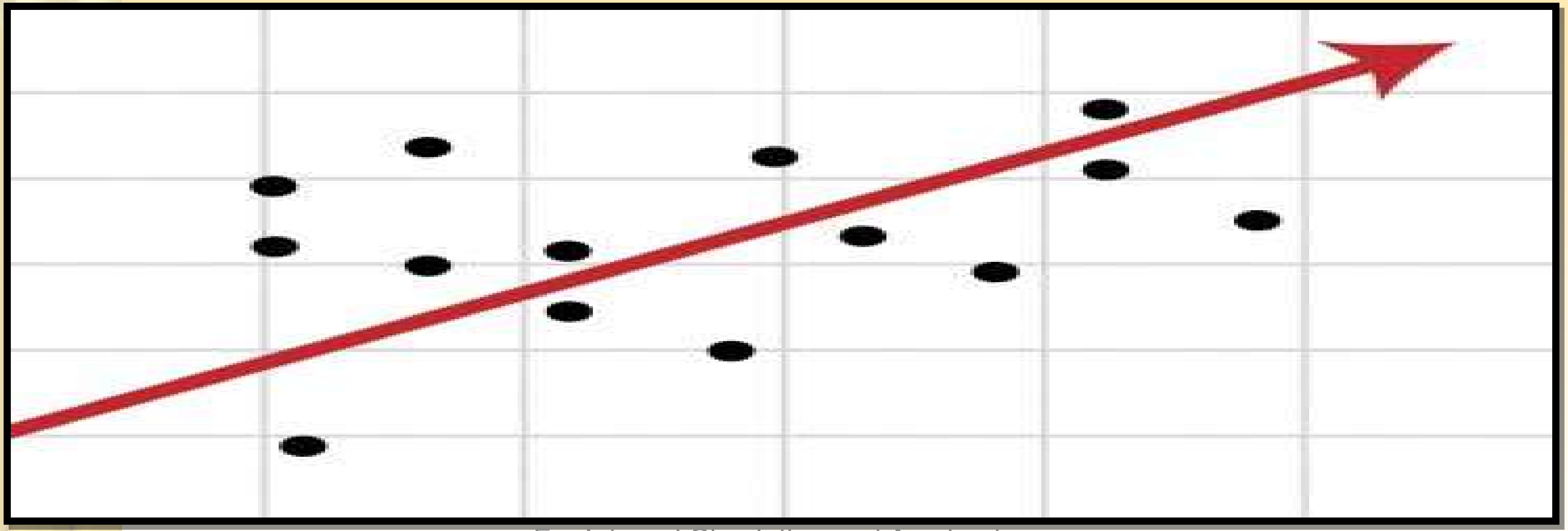
Find the value of the correlation coefficient if regression coefficients of Y on X and X on Y are 0.46 and 0.8 respectively.

Answer : r = 0.606

Method of LEAST SQUARES.

The method of least squares is a mathematical technique.

It is used to obtain the equation of a line which best fits the given data.



Method of LEAST SQUARES.

Equation of a straight line is $y = a + bx$

Normal Equations for obtaining the values of a and b are as follows

(i) $\sum y = Na + b \sum x$

(ii) $\sum xy = a \sum x + b \sum x^2$

EXAMPLE.

Fit a straight line of Y on X from the following data:

X	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Y	**2**	**1**	**3**	**2**	**4**	**3**	**5**

SOLUTION.

	X	Y	X^2	XY
	0	2	0	0
	1	1	1	1
	2	3	4	6
	3	2	9	6
	4	4	16	16
	5	3	25	15
	6	5	36	30
SUM	21	20	91	74

$\sum y = Na + b \sum x$

$\sum xy = a \sum x +$

$20 = 7a + 21b$

$74 = 21a +$

$20 = 7a + 21b$

Multiply by 3

$60 = 21a + 63b$

SOLUTION.

74 = 21a + 91b
minus
60 = 21a + 63b

14 = 28 b
b = 0.5

Putting value of b = 0.5 in above equations;
a = 1.357

Y = 1.357 + 0.5 X
Equation of the line which fits the given data.

EXAMPLE.

Fit a straight line of Y on X from the following data:

X	**0**	**2**	**4**	**6**	**8**	**10**	**12**
Y	**4**	**2**	**6**	**4**	**8**	**6**	**10**

EXAMPLE.

Fit a straight line of X on Y from the following data:

X	**0**	**2**	**4**	**6**	**8**	**10**	**12**
Y	**4**	**2**	**6**	**4**	**8**	**6**	**10**

Normal Equations for obtaining the values of a and b are as follows

(i) $\sum X = Na + b \sum Y$

(ii) $\sum XY = a \sum Y + b \sum Y^2$

EXAMPLE.

From the following data obtain the Regression equations of Y on X and X on Y

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

Answer:

$$Y = 11.9 - 0.65\,X$$

$$X = 16.4 - 1.3\,Y$$

Derivation of Lines of Regression directly from the data. (DEVIATION FROM ACTUAL MEAN)

The equation of regression line of X on Y is

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

$$X - \bar{X} = \frac{r\,\sigma x}{\sigma y}(Y - \bar{Y})$$

The equation of regression line of Y on X is

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

$$Y - \bar{Y} = \frac{r\,\sigma y}{\sigma x}(X - \bar{X})$$

Derivation of Lines of Regression directly from the data. (DEVIATION FROM ACTUAL MEAN)

Regression Coefficient of X on Y is

$$b_{xy} = \frac{\sum xy}{\sum y^2}$$

Regression Coefficient of Y on X is

$$b_{yx} = \frac{\sum xy}{\sum x^2}$$

In these formulas x and y are deviations from mean.

Regression Coefficients directly from the data.

The regression Coefficient of X on Y is

$$b_{xy} = \frac{r\,\sigma x}{\sigma y} = \frac{\sum XY - N\ \bar{X}\ \bar{Y}}{\sum Y^2 - N\,(\bar{Y})^2}$$

The regression Coefficient of Y on X is

$$b_{yx} = \frac{r\,\sigma y}{\sigma x} = \frac{\sum XY - N\ \bar{X}\ \bar{Y}}{\sum X^2 - N\,(\bar{X})^2}$$

EXAMPLE.

From the following data obtain the Regression equations using deviation of means method.

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

Answer:

$Y = 11.9 - 0.65\ X$

$X = 16.4 - 1.3\ Y$

X	x	x^2	Y	y	y^2	xy
6	0	0	9	1	1	0
2	-4	16	11	3	9	-12
10	4	16	5	-3	9	-12
4	-2	4	8	0	0	0
8	2	4	7	-1	1	-2
30	0	40	40	0	20	-26

X on Y

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

$$b_{xy} = \frac{\sum xy}{y^2}$$

$$b_{xy} = \frac{-26}{20} = -1.3$$

X – 6 = -1.3 (Y – 8)

X – 6 = -1.3Y + 10.4

X = -1.3Y + 16.4

$$\bar{X} = \frac{\sum X}{N} = \frac{30}{5} = 6$$

$$\bar{Y} = \frac{\sum Y}{N} = \frac{40}{5} = 8$$

EXAMPLE.

Following data relate to the scores of 9 salesmen of a company in an intelligence test and weekly sales in thousands. If the intelligence test score of a salesman is 65 what would be his weekly expected sales?

Test Score	**50**	**60**	**50**	**60**	**80**	**50**	**80**	**40**	**70**
Weekly Sales	**30**	**60**	**40**	**50**	**60**	**30**	**70**	**50**	**60**

Answer: Rs. 53750

EXAMPLE.

Following table shows ages (X) and blood pressure (Y) of 8 persons. Find the expected blood pressure of a person whose age is 49 years.

X	**52**	**63**	**45**	**36**	**72**	**65**	**47**	**25**
Y	**62**	**53**	**51**	**25**	**79**	**43**	**60**	**33**

Answer: 49.5

EXAMPLE.

In a correlation study the following values are obtained. Find the two regression equations that are associated with above values.

	X	Y
Mean	**65**	**67**
Std. Deviation	**2.5**	**3.5**
Correlation Coefficient	**0.8**	

Answer:

$X = 26.72 + 0.57\ Y$

$Y = -5.8 + 11.2\ X$

Properties of Regression Coefficients.

1. Both regression coefficients will have the same sign i.e. positive or negative
2. Square root of the product of two regression coefficients will give Correlation coefficient.
3. If one regression coefficient is more than 1 then the other will be less than 1.
4. The regression coefficients will have the same sign as that of the correlation coefficient.
5. Arithmetic mean of Regression coefficient is greater than correlation coefficient.
6. The two regression lines intersect at coordinates which are mean of series X and series Y.

Derivation of Lines of Regression directly from the data. (DEVIATION FROM ASSUMED MEAN)

Regression Coefficient of X on Y is

$$b_{xy} = r\frac{\sigma x}{\sigma y} = \frac{N\sum dx\,dy - (\sum dx\; X\; \sum dy)}{N\sum dy^2 - (\sum dy)^2}$$

Regression Coefficient of Y on X is

$$b_{yx} = r\frac{\sigma y}{\sigma x} = \frac{N\sum dx\,dy - (\sum dx\; X\; \sum dy)}{N\sum dx^2 - (\sum dx)^2}$$

In these formulas dx and dy are deviations from ASSUMED mean.

EXAMPLE.

From the following data obtain the Regression equations by taking deviation of series X from 5 and of series Y from 7.

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

Answer:

$$Y = 11.9 - 0.65\,X$$

$$X = 16.4 - 1.3\,Y$$

EXAMPLE.

A panel of two judges P and Q graded eight dance performances as below. However, Judge Q was absent during the 7th performance. What might have been his ranking if judge Q had been present?

P	**46**	**42**	**44**	**40**	**43**	**41**	**37**	**45**
Q	**40**	**38**	**36**	**35**	**39**	**37**	**?**	**41**

Answer:
$Y = +0.75\ X + 5.75$
33.5 marks

EXAMPLE.

From the following regression equations find the mean values of X and Y series.

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

$$40X - 18Y = 214$$

Hint: Regression lines (X on Y and Y on X) cut each other at the point of mean.

Answer:
Mean of X series = 13
Mean of Y series = 17

EXAMPLE.

In a partially destroyed laboratory record of analysis of correlation data, only the following results are legible:

Variance of X = 9

Regression equations

$8X - 10Y + 66 = 0$

$40X - 18Y = 214$

Find out :

1. The mean values of X and Y
2. Coefficient of correlation between X and Y
3. Standard Deviation of Y

Solution.

To find the mean values of X and Y we will solve the given equations:

Regression equations given:

(i) $8X - 10Y = -66$

(ii) $40X - 18Y = 214$

Multiplying equation (i) by 5 we get

(i) $40X - 50Y = -330$

Subtracting equation (ii) from equation (i) we get:

$-32\ Y = -544$

$Y = -544 / -32 = 17$ (Mean of Y = 17)

Solution.

Substituting the value of Y=17 in equation (i)

(i) $8X - 10Y = -66$

$8X - 10(17) = -66$

$8X - 170 = -66$

$8X = 104$

X = 104 / 8 = 13 (Mean of X = 13)

To find the Correlation coefficient (r) we must find the regression coefficients (b_{xy} and b_{yx}). However, we don't know which equation is X on Y and which equation is Y on X. So lets assume equation (i) as X on Y

Solution.

Assuming equation (i) as X on Y we will try to find bxy

(i) $8X – 10Y = -66$

$8X = -66 + 10Y$

$X = (-66/8) + (10/8) Y$

$X = a + bxy\ Y$

bxy= 10 / 8 = 1.25

From equation (ii) we can find out byx

(ii) $40X – 18Y = 214$

$Y = -(214/18) + (40/18) X$

bxy= 40 / 18 = 2.22

Solution.

Here we can see that both the regression coefficients are greater than 1 which is not possible. Therefore, our assumption that equation (i) is equation of X on Y is wrong.

Equation (i) is actually equation of Y on X. So we can write the equation as:

$-10\,Y = -\,8X - 66$

$Y = (8/10)\,X + 6.6$ which means that $byx = 0.8$

From equation (ii) i.e. X on Y we get

$40X = 214 + 18Y$ or $X = (214/40) + (18/40)\,Y$ which means that $bxy = 18/40 = 0.45$

Solution.

We know that Correlation coefficient is square root of product of Regression coefficients:

r = square root of 0.8 X 0.45 = 0.6

We can further calculate standard deviation of Y using the formula of Regression Coefficient.

Standard Deviation of Y = 4

STANDARD ERROR OF ESTIMATES.

- The standard error of estimate measures the accuracy of the estimated figures.
- Smaller the value of standard error, closer will be the dots to the regression line and better will be the estimates.
- If standard error of estimate is zero then the is no variation about the line and correlation is perfect.

$$S_{xy} = \sqrt{\frac{\sum (X - X_c)^2}{N}}$$

$$S_{xy} = \sigma x \sqrt{1 - r^2}$$

$$S_{xy} = \sqrt{\frac{\sum X^2 - a\sum X - b\sum XY}{N}}$$

$$S_{yx} = \sqrt{\frac{\sum (Y - Y_c)^2}{N}}$$

$$S_{yx} = \sigma y \sqrt{1 - r^2}$$

$$S_{yx} = \sqrt{\frac{\sum Y^2 - a\sum Y - b\sum XY}{N}}$$

Uses of Regression analysis.

1. Regression line facilitates to predict the values of dependent variable from the given value of independent variable.
2. Standard Error facilitates to obtain a measure of error involved in using the regression line as a basis of estimation.
3. Regression coefficients help us to calculate coefficient of correlation and coefficient of determination,
4. Regression analysis is a highly useful tool in Economics and business.

References and Suggested Readings

Fundamentals of Statistics by S.C. Gupta

Statistics Methods by S.P.Gupta