SECTION 2 CORRELATION & REGRESSION

Structure **Page Nos.**

2.0 INTRODUCTION

The objective of this unit is to enable you to investigate the interdependence of variables in terms of Correlation and Regression analysis through hands-on experience in using MS-Excel and its tools viz., Data Analysis Tool pack. However, prior to this you should go through *BCS- 040 Block 3 Unit 9, Regression Analysis*, which is a prerequisite.

Whenever you are going to conduct a study or experiment or research, irrespective of the discipline, it is desired to analyze the dependence of one variable over other(s). Correlation and Regression are used to analyse collected sample data to investigate the relationship between the variables to answer associated questions such as:

- Is there any relationship between the variables under study?
- How strongly the variables are related to each other?
- Can we use the relationship to estimate or forecast the value of one of the variables (dependent variable)?

This section provides a practical orientation in the light of your understanding of BCS-040.

2.1 OBJECTIVES

After going through this unit, you will be able to perform:

- correlation analysis through Excel;
- multiple Correlation analysis through Excel;
- linear Regression analysis through Excel; and
- multiple Regression analysis through Excel.

2.2 CORRELATION

Let us start by describing the following simple example: In any computer system, there are various components like Memory, Processor, Motherboard etc, and say you want to study how the performance of any computer system varies with different permutation and combination of some constituent components. For example, you may want to know “is there any relationship between the size of Random Access Memory - RAM and the performance of Computer system” OR “is there any relationship

between Hard disk storage capacity and the performance of Computer system" etc. in order to address such queries, you are required to conduct a statistical experiment that entails in the collection and tabulate the data as discussed earlier. So, collect and tabulate data under heading say "RAM-Size" and "Performance-Status" OR "Hard Disk-Capacity" and "Performance-Status".

Thus, it is required to carry out a statistical investigation of the fact that a Computer System's performance improves on increasing the RAM size. Since it is known that there is non-linear relationship between the two variables, a linear regression (which involves linear relationship between variables), will "not be a good approximation" in the study of relationship between system performance and RAM size. Consequently, a non-linear relationship of appropriate type, between computer system performance and RAM size will only give a better approximation (read chapter 8, Book 3). Since, Non Linear regression is not within the scope of this course, we will restrict our discussion only to Correlation and Linear Regression. Recall that a linear relationship between variables x (independent variable) and y (dependent variable) can be put in the form

$$y = a + b\,x.$$

The objective of our study is to investigate the question

> "Is there any relation between the size of Random Access Memory (RAM) and the Performance of Computer System?"

Steps:

- From the objective of the problem, identify the two variables, which in this case are "size of Random Access Memory (RAM)" and the "performance of Computer system".
- In order to proceed with our study, we are required to collect data, for which we are require to have RAM of different sizes. Further, for the sake of simplicity, let us restrict the study to the processor existing in the system. It is an exhaustive exercise, in which we have to mount different RAM on the Motherboard slot repeatedly, restart the system and monitor the system performance parameters. Say, we record the Percent (%) variation in RAM size and the percent (%) variation in the System performance in the Excel spread sheet.
- Since we are investigating the effect on System performance, the influence of variation in RAM size, i.e., the % variation in RAM is the independent variable x and % variation in system performance y is the dependent variable.
- Notice that the importance of identifying the two variables rests on the fact that there are *two different lines of regression* viz., y on $x$and x on y. (see Chapter 8, Book 3)

Now to calculate the correlation between the two continuous variables in excel, tabulate the recorded observations into the excel worksheet. Excel enables us to investigate the question stated above in three different ways, viz., through

1.Excel Formula 2. Standard formula - Correl(x,y)3. Data Analysis ToolPak

As an exercise you are required to observe the consistency of the results obtained through all three methods. This will help you to identify the mistakes, which might have occurred while writing your own formula.

In the screen shot given below you can identify that all three options leads to same result.

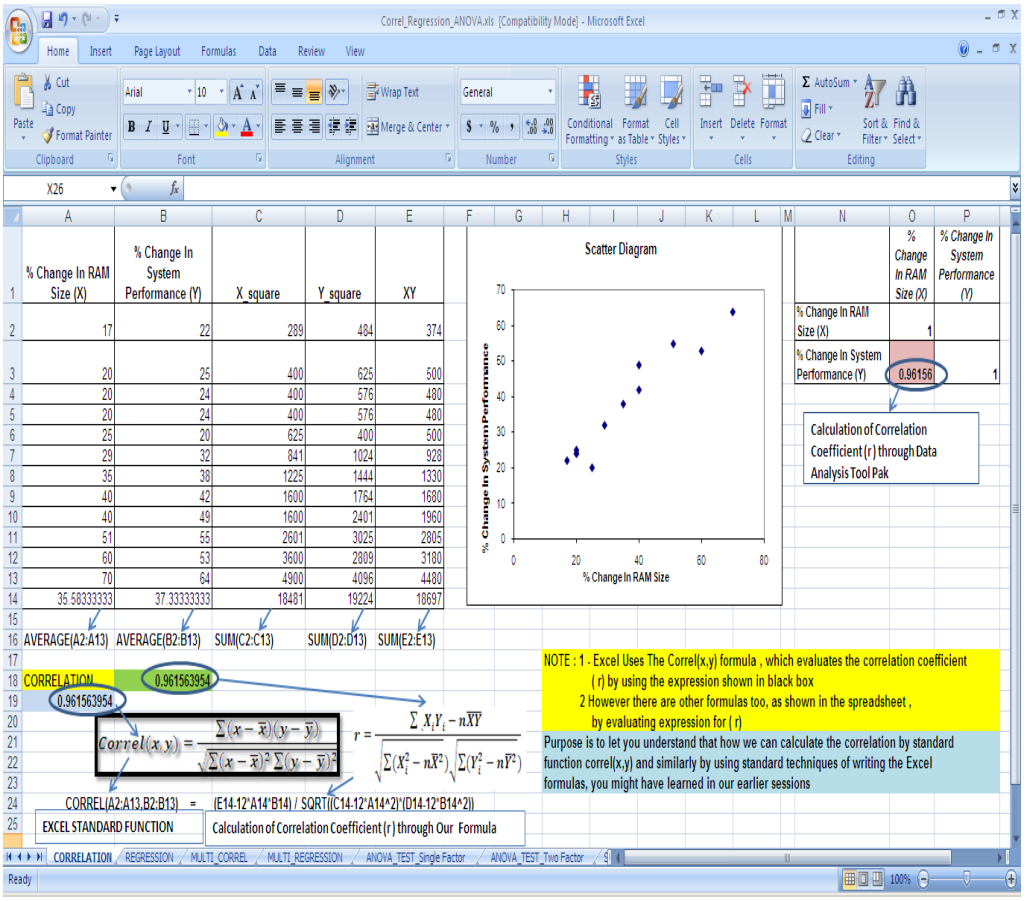


Figure 1: Correlation

Now, let us explore each of the option to calculate the correlation coefficient (r) i.e., using own Excel Formula, Standard formula - ***correl(x,y)*** and Data Analysis ToolPak

1. **Own Excel Formula**: There are different forms of the formula to calculate correlation coefficient (r).Standard function described in excel i.e. ***correl(x,y)*** uses the formula in the form

$$r = \mathrm{r}(\mathrm{x},\mathrm{y}) = \frac{\sum(\mathrm{x}-\bar{\mathrm{x}})(\mathrm{y}-\bar{\mathrm{y}})}{\sqrt{\sum(\mathrm{x}-\bar{\mathrm{x}})^2 \sum(\mathrm{y}-\bar{\mathrm{y}})^2}}.$$

However, we will calculate the same using the form

$$\mathrm{r} = \frac{\sum \mathrm{X_i Y_i} - \mathrm{n}\,\bar{\mathrm{X}}\bar{\mathrm{Y}}}{\sqrt{\sum \mathrm{X_i^2} - \mathrm{n}\,\bar{\mathrm{X}}^2}\sqrt{\sum \mathrm{Y_i^2} - \mathrm{n}\,\bar{\mathrm{Y}}^2}},$$

in our excel implementation.

To implement it, tabulate the data as shown in the above screen shot and write the formula **(E14-12*A14*B14)/SQRT((C14-12*A14^2)*(D14-12*B14^2))** to calculate correlation coefficient (r). You might have learned the method of formula writing in excel in earlier courses. Another form useful in computations is

$$\mathrm{r} = \frac{n\sum \mathrm{X_i Y_i} - (\sum \mathrm{X_i})(\sum \mathrm{Y_i})}{\sqrt{n\sum \mathrm{X_i^2} - (\sum \mathrm{X_i})^2}\sqrt{n\sum \mathrm{Y_i^2} - (\sum \mathrm{Y_i})^2}}.$$

2. **Standard formula - correl(x,y)** :

   a. Tabulate the collected data in column A & B as shown in above screen shot

   b. Select any cell in which you wish to have result of correlation coefficient (r)

c. In that cell write CORREL(A2:A13,B2:B13), where A2:A13 and B2:B13 are the columns related to the tabulated/recorded data, i.e., the values of variables x and y.

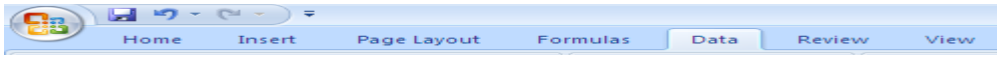
You can also use Excel Help by pressing function F1 key and type the function or formula you want to understand. Say, press F1 and type **correl** in search option you will get all related details of that formula. CORREL(X,Y) is the Excel formula to calculate correlation coefficient (r).

Here, $\boldsymbol{Correl(x,y)} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$ where $\bar{x}$&$\bar{y}$ are average values of variable x&y respectively.

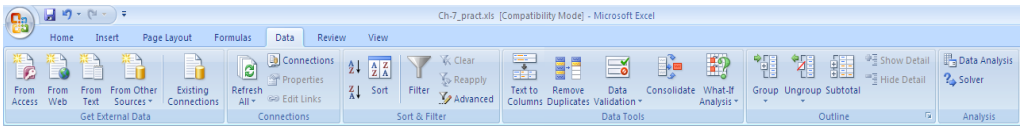
3. **Using Data Analysis ToolPak** : You may calculate the correlation coefficient (r) by using the Data Analysis Toolpak utility of Excel, which provides a set the tools necessary for statistical Data Analysis.

Steps to find Correlation coefficient (r) using Data Analysis ToolPak:

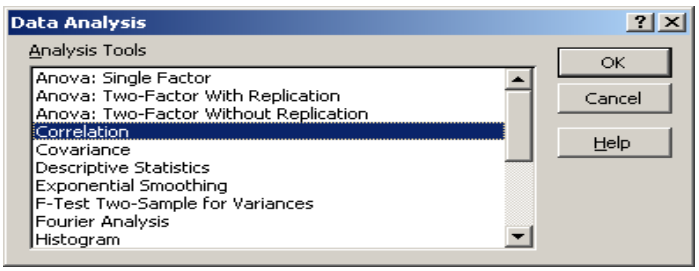
a. Click the Data tab



b. Click the Data Analysis tab



c. Select correlation option and click OK



d. Select the input range,

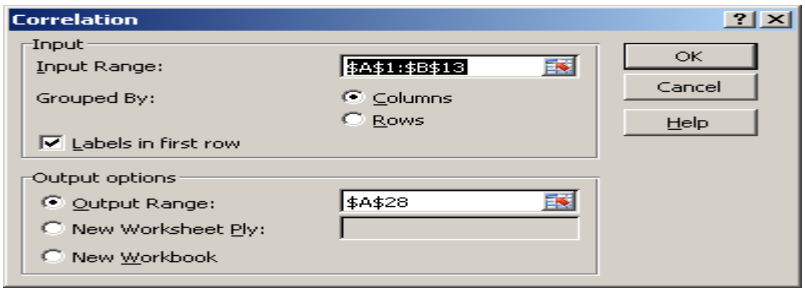
i. Which in our case is A1 to B13, relating to % variation in RAM and % variation in system performance.

ii. Choose the columns option as the data is tabulated in the columnar manner.

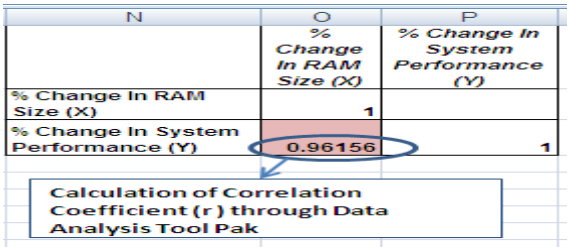
iii. Check the Option Labels in first row, as in A1 and B1 we are having headings for respective columns.

iv. Finally, choose the cell location where you want to have the output result.

v. press OK



e. Output:



	N	O	P
		% Change In RAM Size (X)	% Change In System Performance (Y)
% Change In RAM Size (X)		1	
% Change In System Performance (Y)		0.96156	1

Calculation of Correlation Coefficient (r) through Data Analysis Tool Pak

NOTE : We will prefer to use Data Analysis Toolpak for other applications subsequently and would like to put more emphasis on data interpretation.

Screen Shot of Final Outcome

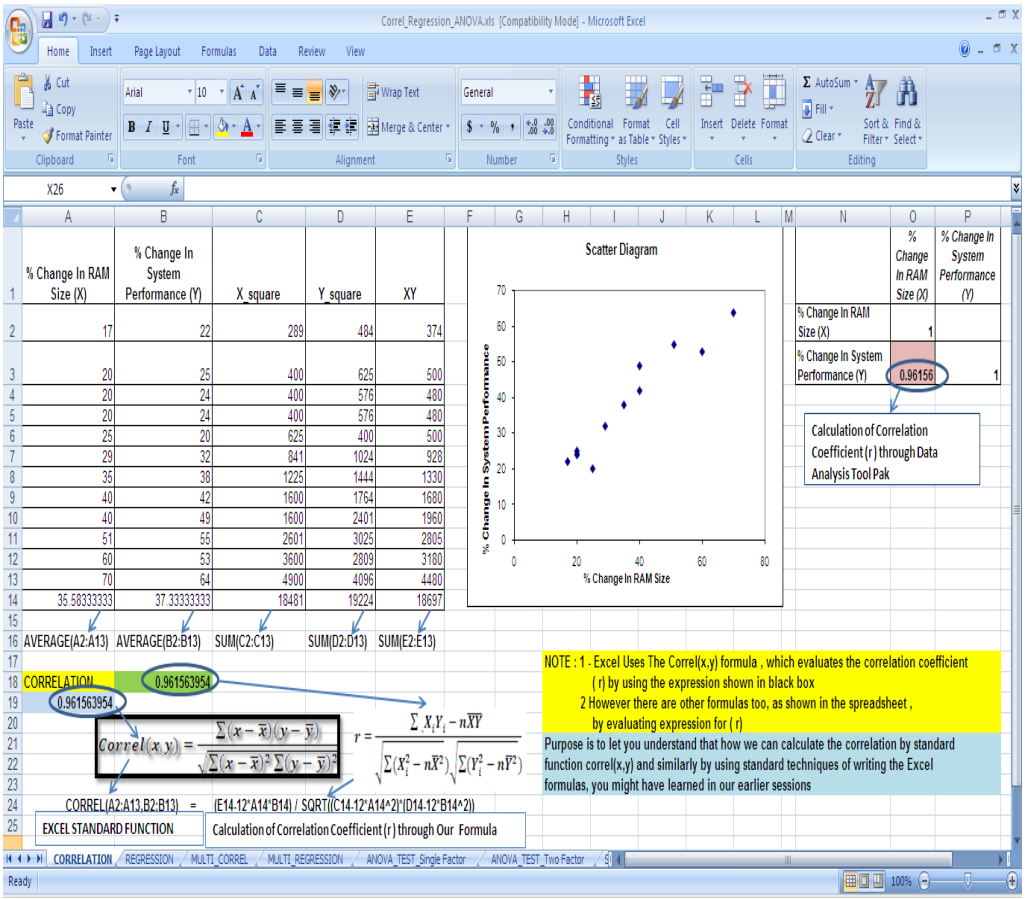


Figure 2 : Correlation final outcome

Remark on Data Interpretation

- **Which curve to fit?**
  - Between two continuous variables the relationship can be ascertained by first plotting a scatter diagram.
  - Then, simply looking at the trend of the resultant plot, it could be linear or curvilinear.
  - You are required to apply your understanding of our earlier sessions to plot scatter diagram etc. in Excel.

- **Why r ?**
  - If the resulting scatter diagram exhibits a linear relationship between x and y, the extent can be quantified using correlation coefficient (r),which "measures the extent of linearity" present in the data.
  - Value of r always lies between -1 and $+1$.
  - Closer the value of r towards ± 1, the stronger is the linear relationship between the variables.
  - You can verify in the subsequent section on simple linear regression that depending on the value of r being $+ve$or $-ve$, the slope of the line of regression will also be $+ve$ or $-ve$ respectively. In other words, when r is $+ve$ or $-ve$, an increase in the value of "independent variable (x)" will result in an increase or decrease in the value of the "dependent variable (y)".
  - If the relationship is found to be significantly strong, the line of "best fit "passing through the bi-variate data can be obtained and it is discussed in section 2.4.

☞ Check your progress 1

Try this: Analyze the screen shot and use the data interpretation tips given above, to answer following questions:

a) Study the scatter diagram only, what can we conclude about the data?

.....

.....

.....

b) Analyze the correlation coefficient (r) and comment on the relationship between RAM size variation and System performance.

.....

.....

.....

c) Can we use the collected data for the purpose of forecasting?

.....

.....

.....

2.3 MULTIPLE CORRELATION

In this section, we extend our discussion to the case of coefficient of multiple correlation, which measures the extent to which a dependent variable can be predicted using the *linear function of a set of independent variables*. To explain the concept let us take another example from the IT sector, companies generally have a monitoring index called the technical index(TECH INDEX).The monitoring of this index goes on monthly basis, and the indexing is performed to identify growth of an independent company with respect to the industry requirements. The data related to the variation between industry parameter called TECH INDEX and the monthly rate of return is gathered for few companies viz. Google, Yahoo, Microsoft, Apple; the same is to be analyzed. Since variation of multiple companies in coordination with one parameter is to be analyzed, we take recourse to the concept of Multiple Correlation which you studied in BCS 040 Block 3. Denoting the companies by x_1, x_2, x_3, x_4 and industry index by y, a multiple regression model that describes this relationship can be put in the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \in$$

You are required to perform the following:

1. Plot the graph for all collected data.

   This will help you to visually identify the variation in variable synchronization.

2. Run Correlation analysis on all the variables simultaneously.

Given below is the data on various IT sector companies stated above:

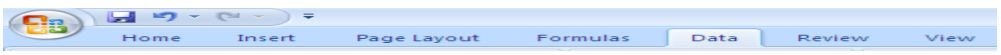
MONTHLY RATE OF RETURNS FOR IT SECTOR COMPANIES

		COMPANY			
DATE	**TECH INDEX**	**GOOGLE**	**YAHOO**	**MICROSOFT**	**APPLE**
1-Apr-11	0.8799	0.7541	2.1407	-4.6296	-18.8406
1-May-11	7.5187	14.9701	-2.5948	18.986	6.6964
1-Jun-11	5.558	11.9792	7.7869	-1.7226	-3.3473
1-Jul-11	1.3716	7.907	-8.5551	-0.5535	5.8442
1-Aug-11	-1.6289	-5.1724	1.2474	6.679	1.9427
1-Sep-11	2.4171	3.4091	0.8214	1.8261	2.1063

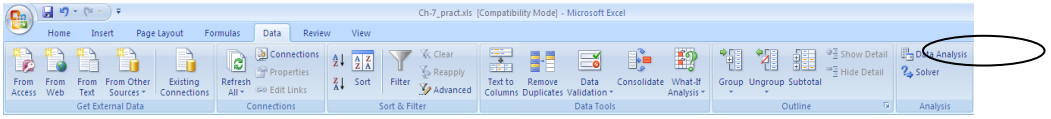
As the first step of analysis, we are interested in visualizing the data graphically and to this end, obtaining a scatter plot is simple as well as most appropriate. Notice that in the case of multiple correlation and regression with only *two independent variables*, scatter diagram can be plotted in the three dimensions. Thus, for the sake of completeness, we can generate a plot with date along the x-axis and the series values along the same y-axis to get an insight into the given data. (see Output Screen Shot given below)

Steps for Multiple Correlation using Data Analysis ToolPak

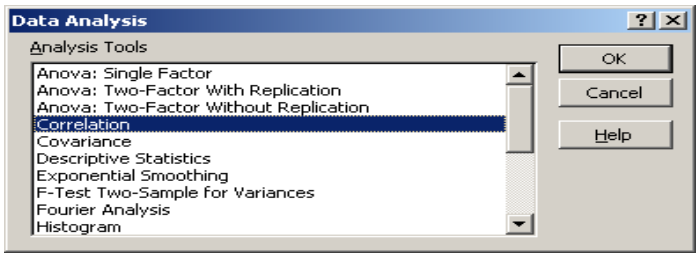
1. Click the Data tab



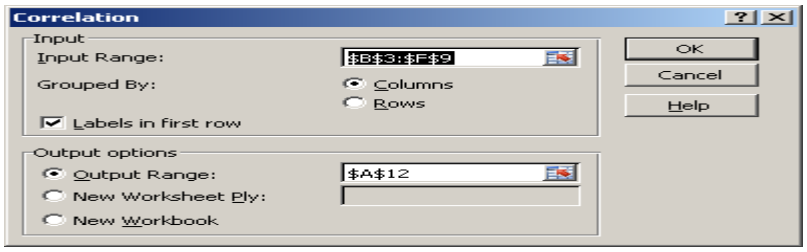
2. Click the Data Analysis tab



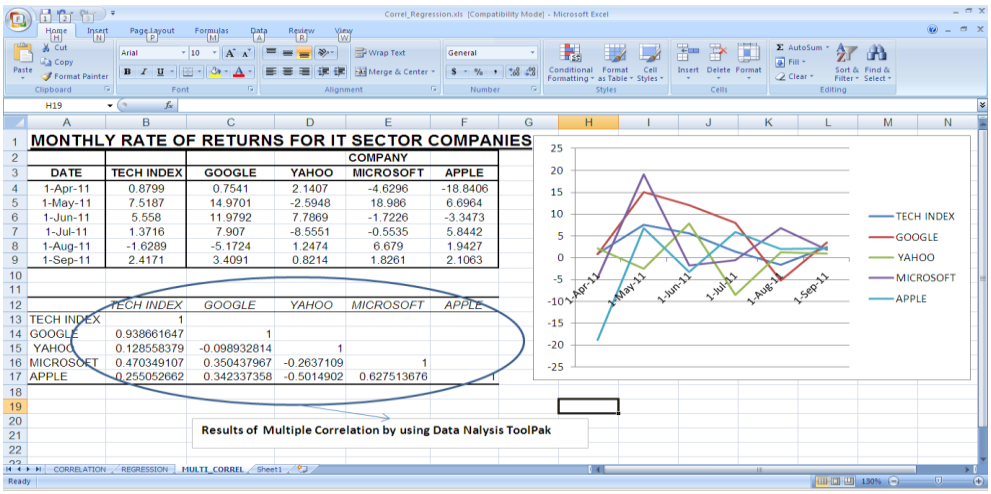
3. Select correlation option and click OK



4. Select the input range,
   a. Which in our case is $B$3:$F$9, related to TECH Index and Companies data understudy.
   b. Choose the columns option as the data is tabulated in the columnar manner.
   c. Check the Option Labels in first row, as from B3 to F3 we are heaving headings for respective columns.
   d. Finally, opt for the cell location where you want to have the output result
   e. press OK



5. Output



DATA INTERPRETATION

	TECH INDEX	*GOOGLE*	*YAHOO*	*MICROSOFT*	*APPLE*
TECH INDEX	1				
GOOGLE	0.938661647	1			
YAHOO	0.128558379	-0.098932814	1		
MICROSOFT	0.470349107	0.350437967	-0.2637109	1	
APPLE	0.255052662	0.342337358	-0.5014902	0.627513676	1

The table shown above contains correlation coefficients discussed in the previous section, computed between pairs of variables for the data under study. Notice that it is in the form of a matrix, which is known as *matrix of correlation coefficients* and it is symmetric about the diagonal (why?). Clearly, the elements along the diagonal are correlation coefficients computed between values of a variable with itself, which is *unity* (refer to section 2.2 and state why?).Further, our study requires us to analyze the relation between industry parameter TECH INDEX and Monthly rate of return of various companies. Clearly,

1. There is high correlation (0.938661647) between Google's monthly rate of return data and the industry TECH INDEX, which is on the expected lines.
2. Similarly, for the Microsoft's monthly rate of return data the correlation coefficient is0.470349107 and it is reasonably high.
3. In contrast to these,the respective correlation coefficients in the case of Yahoo and Apple are 0.128558379 and 0.255052662 respectively, which is low.

Thus, we learned how to interpret the results obtained by using the correlation coefficient in Data analysis ToolPak. However, we will subsequently in section 2.4.2 learn how to compute the value of multiple correlation coefficient (denoted by R) and to this end, we will extend your knowledge of section 9.3/Unit 9.

2.4 REGRESSION

In section 2.2 above, you learned about correlation and the related excel tools and now we extend our discussion towards regression. Regression analysis enables us in estimation and forecasting of the value of dependent variable for given value(s) of the independent variable(s) and is extensively used in various disciplines. Clearly, while in simple linear regression there is a single independent variable; in multiple regression there are more than one independent variable. In other words, the former is confined to the study of two variables only; the latter is concerned with the study of more than two variables. Further, in case of simple regression if the relationship between the dependent and independent variables follows a straight line pattern, it is called *linear regression*. On the other hand, if the relation is expressed in the form of a curve, it is called curvilinear regression.

Task: Discuss with other students or counselors

- Example of curves.
- Example of curves that *can be* transformed into linear form (called tractable linear form). What about$y = a\ b^x$?
- Example of curves that *cannot be* transformed into linear form (called non-tractable form). What about$y = a + b^x$?
- Importance of scatter diagram in deciding a specific curve.

We will restrict our discussion to the simple linear regression and later extend it to the case of Multiple regression. The general problem of finding the equation that fits a given data set is called *curve fitting*.

Before starting with the practical problem solving through excel, we present below a discussion on regression.

2.4.1 Linear Regression

Here we are going to study the case of two variables X&Y. Thus,there are two lines of regression viz.,X on Y and Yon X. The regression line Y on$X$ gives the most probable values of Y for a given value of$X$ and the regression line X on Ygives the most probable values of X for a given value of$Y$. However, when there is a perfect correlation between X&Y i.e.,$r = \pm 1$, the two regression lines Y on X and X on Y coincides i.e. we will have one regression line(see chapter 8, Book 3). Otherwise, there are two lines of regression which coincide at $(\bar{x}, \bar{y})$.

The regression line Y on X is given by

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

which simplifies to,

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

Similarly, the regression line X on Y is given by

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

which simplifies to,

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

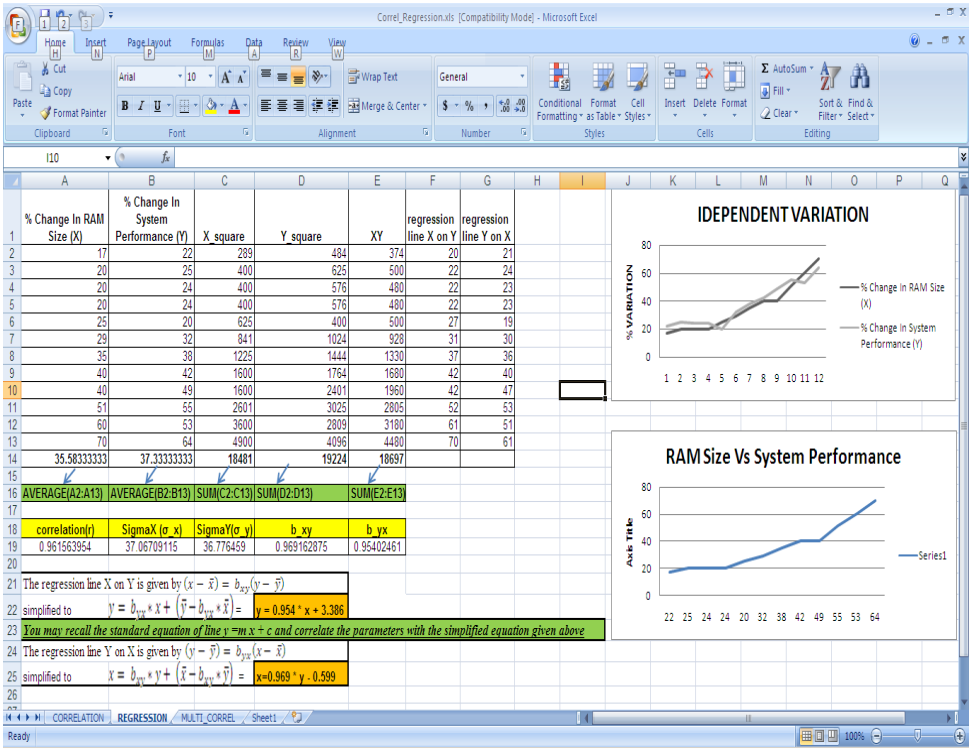
Where,

- $\bar{x}, \bar{y}$ and σ_x, σ_y are arithmetic mean and standard deviation of variables$x$ and y respectively,
- $b_{xy} = r \times \frac{\sigma_x}{\sigma_y}$is the regression coefficient of X on Y,
- $b_{yx} = r \times \frac{\sigma_y}{\sigma_x}$is the regression coefficient of Y on X,
- $r$is correlation coefficient.

Note:

1. The two equations above can re-written in a simple way in terms of standardized values in the form given below:
   $\frac{y-\bar{y}}{\sigma_y} = r\frac{x-\bar{x}}{\sigma_x}and\frac{x-\bar{x}}{\sigma_x} = r\frac{y-\bar{y}}{\sigma_y}$.
2. The two forms written above are the least-squares lines of regressions of Y on X and X on$Y$.
3. You may recall the equation of a straight line $y = m\,x + c$ and compare the parameters with the least-squares line of regressions given above to get $m = b_{yx}$and $c = (\bar{y} - b_{yx} \times \bar{x})$.
4. $b_{xy} \times b_{yx} = r^2$. (you may refer to chapter 8, Book 3 for details)
5. We recommend you to apply the skill you developed in earlier sessions and perform calculations by writing your own formula for correlation and regression.

However, we exhibit the utilization of Excel Data Analysis ToolPak utility, in our subsequent section on multiple regression.



	% Change In RAM Size (X)	% Change In System Performance (Y)	X_square	Y_square	XY	regression line X on Y	regression line Y on X
2	17	22	289	484	374	20	21
3	20	25	400	625	500	22	24
4	20	24	400	576	480	22	23
5	20	24	400	576	480	22	23
6	25	20	625	400	500	27	19
7	29	32	841	1024	928	31	30
8	35	38	1225	1444	1330	37	36
9	40	42	1600	1764	1680	42	40
10	40	49	1600	2401	1960	42	47
11	51	55	2601	3025	2805	52	53
12	60	53	3600	2809	3180	61	51
13	70	64	4900	4096	4480	70	61
14	35.58333333	37.33333333	18481	19224	18697		
16	AVERAGE(A2:A13)	AVERAGE(B2:B13)	SUM(C2:C13)	SUM(D2:D13)	SUM(E2:E13)		
18	correlation(r)	SigmaX (σ_x)	SigmaY(σ_y)	b_xy	b_yx		
19	0.961563964	37.06709115	36.776459	0.969162875	0.95402461		
21	The regression line X on Y is given by $(x-\bar{x}) = b_{xy}(y-\bar{y})$						
22	simplified to	$y = b_{yx} * x + (\bar{y} - b_{yx} * \bar{x})$ =		y = 0.954 * x + 3.386			
23	*You may recall the standard equation of line y =m x + c and correlate the parameters with the simplified equation given above*						
24	The regression line Y on X is given by $(y-\bar{y}) = b_{yx}(x-\bar{x})$						
25	simplified to	$x = b_{xy} * y + (\bar{x} - b_{xy} * \bar{y})$ =		x=0.969 * y - 0.599			

Figure 3 : Regression

Remark on Data Interpretation:

1. **Correlation coefficient** (r)
   From the above screen shot, one can observe that the value of correlation coefficient (r) is quite high, thus we may conclude that the X and Y are directly proportional.

2. **Regression coefficients** $b_{xy}$and b_{yx}
   $b_{xy} = 0.69$ implies that with unit increase in Y, the value of X increases by 0.69 times. Similarly $b_{xy} = 0.54$ implies that with unit increase in$X$, the value of Y increases by 0.54 times.

☞ Check your progress 2

Try this: Analyze the screen shot and use the data interpretation tips given above, to attempt following questions:

a) Comment on the interdependence of RAM size variation and System performance.

.....

.....

.....

.....

b) With an unit increase in RAM size, how much you expect the system performance to improve?

.....

.....

.....

.....

c) For unit improvement in system performance how much RAM size is expected to be altered?

.....

.....

.....

d) Can you forecast the system performance, if change in RAM size is 25%?

.....

.....

2.4.2 Multiple Regression

Did you complete the sections 2.3 and 2.4.1 before starting with this section? It is advised that you should complete these sections before starting with multiple regression analysis. You may ask why? There as on being, it will enable you to identify the continuity and extensions in the topics you are working on. As in the case of TECH INDEX parameter you studied in section 1.2.1 above, we firstly identified the companies which influence the composite industry index. Thus, we ought to include variables that can explain the variation in the composite industry index, viz. GOOGLE & MICROSOFT and not YAHOO & APPLE. Data on the first two companies considered for multiple regression analysis are listed below:

		COMPANY	
DATE	TECH INDEX	GOOGLE	MICROSOFT
1-Apr-11	0.8799	0.7541	-4.6296
1-May-11	7.5187	14.9701	18.986
1-Jun-11	5.558	11.9792	-1.7226
1-Jul-11	1.3716	7.907	-0.5535
1-Aug-11	-1.6289	-5.1724	6.679
1-Sep-11	2.4171	3.4091	1.8261

Screen shot is as follows

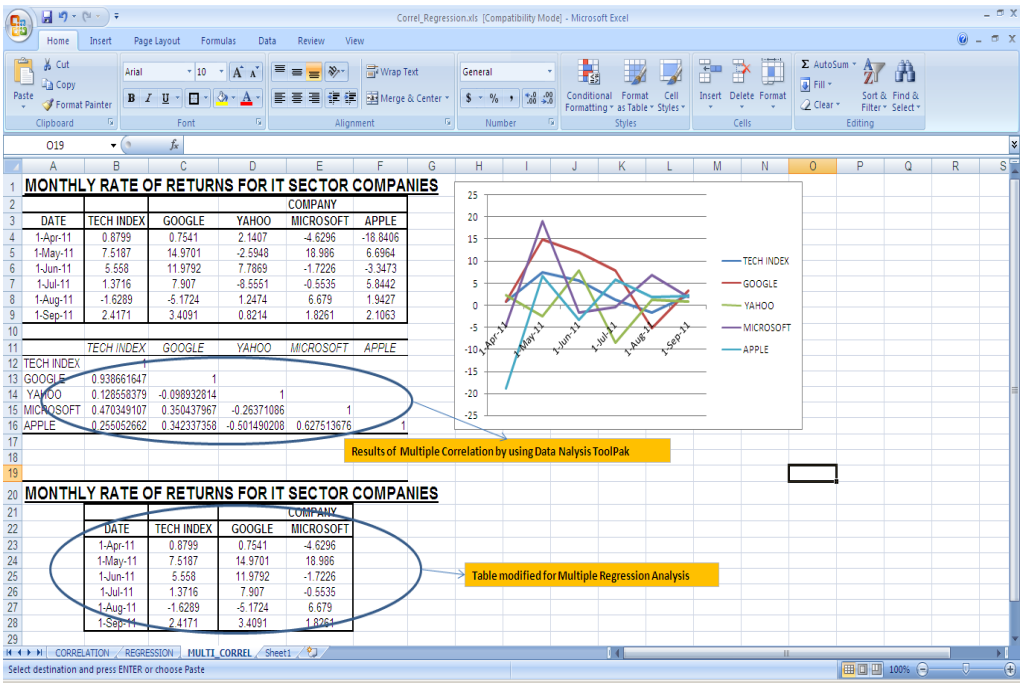
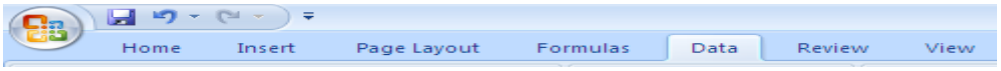


Figure 4: Multiple Regression

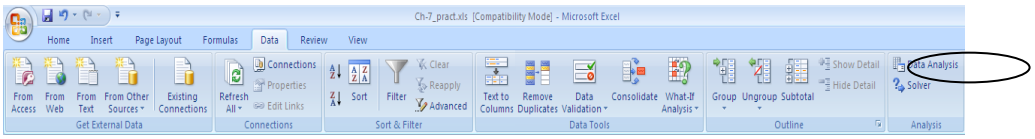
Now let us use the tool i.e., Data Analysis ToolPak's Regression facility to perform multiple regression analysis:

Steps for Multiple Regression using Data Analysis Tool Pak:

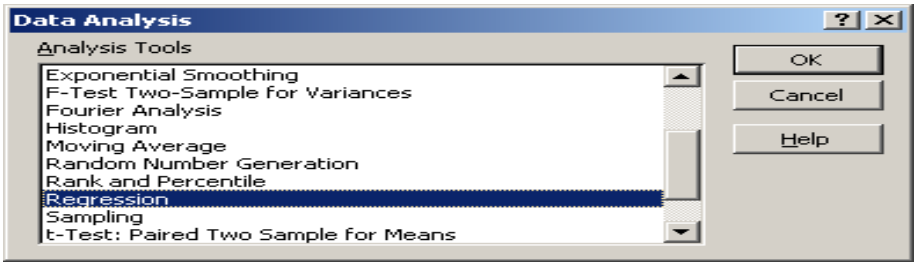
1. Click the Data tab



2. Click the Data Analysis tab



3. Select Regression option and click OK



4. Select the respective data range,

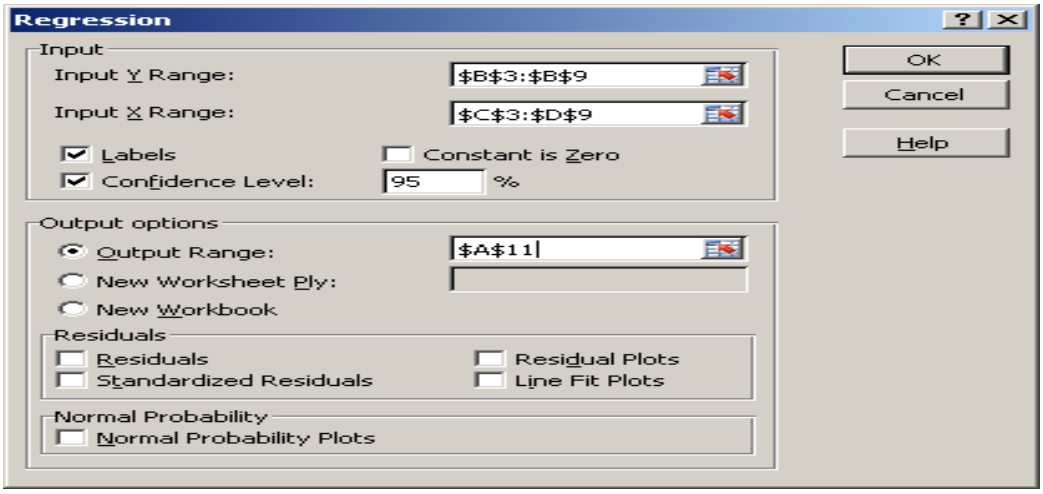
   a. Which in our case is $B$3:$B$9 as Input Y- Range, related to TECH Index and $C$3:$D$9 as Input X-Range, related to shortlisted Companies viz. Google and Microsoft.

   b. Choose the Output range, we opted for cell location $A$11

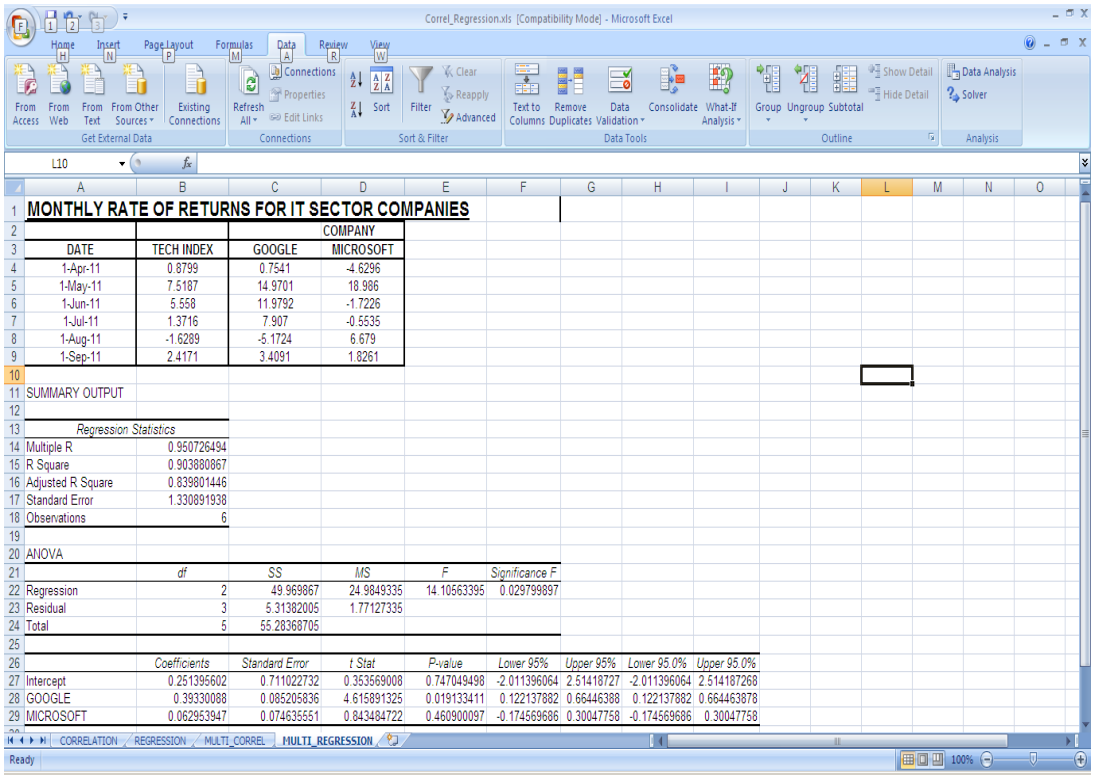
   c. Check the Option Labels and confidence level at which we wish to fix and this in our case is 95%.

   d. Finally, opt for the cell location where you want to have the output result

   e. Press OK.



5. Output



Correl_Regression.xls [Compatibility Mode] - Microsoft Excel

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	MONTHLY RATE OF RETURNS FOR IT SECTOR COMPANIES														
2				COMPANY											
3	DATE	TECH INDEX	GOOGLE	MICROSOFT											
4	1-Apr-11	0.8799	0.7541	-4.6296											
5	1-May-11	7.5187	14.9701	18.986											
6	1-Jun-11	5.558	11.9792	-1.7226											
7	1-Jul-11	1.3716	7.907	-0.5535											
8	1-Aug-11	-1.6289	-5.1724	6.679											
9	1-Sep-11	2.4171	3.4091	1.8261											
10															
11	SUMMARY OUTPUT														
12															
13	*Regression Statistics*														
14	Multiple R	0.950726494													
15	R Square	0.903880867													
16	Adjusted R Square	0.839801446													
17	Standard Error	1.330891938													
18	Observations	6													
19															
20	ANOVA														
21		*df*	*SS*	*MS*	*F*	*Significance F*									
22	Regression	2	49.969867	24.9849335	14.10563395	0.029799897									
23	Residual	3	5.31382005	1.77127335											
24	Total	5	55.28368705												
25															
26		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*						
27	Intercept	0.251395602	0.711022732	0.353569008	0.747049498	-2.011396064	2.514187268	-2.011396064	2.514187268						
28	GOOGLE	0.39330088	0.085205836	4.615891325	0.019133411	0.122137882	0.664463878	0.122137882	0.664463878						
29	MICROSOFT	0.062953947	0.074635551	0.843484722	0.460900097	-0.174569686	0.30047758	-0.174569686	0.30047758						

CORRELATION | REGRESSION | MULTI_CORREL | MULTI_REGRESSION

Ready

The Data Analysis ToolPak, gave exhaustive summary for the Data, out of which we will focus on the Regression statistics and subsequent sessions, we will extend our discussion to ANOVA.

Interpretation of Results

1. Following is the summary output obtained using Data analysis ToolPak, which we seek to interpret.

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.950726494
R Square	0.903880867
Adjusted R Square	0.839801446
Standard Error	1.330891938
Observations	6

2. The Multiple R is the multiple correlation coefficient and it measures the strength of the association. The range of Multiple R is -1 to +1, which in the present case yields to the value$R = 0.950726494$.This implies that the dependent variable Y which in our case is TECH INDEX has a high positive correlation with the combined effect of Google (X_1) and Microsoft (X_2).

3. The R Square value (R^2) is the square of the multiple correlation coefficient$R$. It measures the strength of the regression prediction compared with predicting solely by the response mean. Alternatively, the value of $R^2 = 0.903880867$,can also be interpreted as that90.3% of the variation in$Y$can be explained by variables $X_1 \& X_2$. It may further be noted the closer value of R to 1, the stronger is the regression prediction.

4. Defining multiple correlation as $R^2 = \frac{Explained\ Variation}{Total\ Variation}$ which is in terms of sums of squares (SS)can be expressed as

   $R^2 = \frac{SS\ Regression}{SS\ Total} = \frac{49.969867}{55.28368705} = 0.90388086733083$(for values see screen shot above).

5. Discuss with other students or counselors that in general, the value of R^2*will always increase* when an additional regressor / independent variable is added (say YAHOO).Details of which is beyond the scope of current content.

6. We present below the Regression coefficients from the screen shot along with summary statistics.

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0.251395602	0.711022732	0.353569008	0.747049498
GOOGLE	0.39330088	0.085205836	4.615891325	0.019133411
MICROSOFT	0.062953947	0.074635551	0.843484722	0.460900097

The column of coefficients contain the values of intercept (c) and the regression coefficients $m_1$andm_2 of the regression equation

$$Y = c + m_1X_1 + m_2X_2,$$

which in the present case is

$$Y = 0.251395602 + 0.39330088X_1 + 0.062953947X_2$$

The value $m_1$givesus how much will$Y$changefor each unit change inX_1, and X_2 is held constant. Similarly form_2.To understand this, consider the regression equation$Y = -39.62 - 0.5403X_1 + 13.4852X_2$. Here since the coefficient ofX_1 is negative, keeping X_2 constant, an unit change inX_1leads to a decrease in Y by 0.5403. Again, since the value of regression coefficientX_2 is 13.4852, keeping X_1 constant, an unit change in X_2 leads to an increase in$Y$ by 13.4852.

Read section 9.4.2 of Unit 9

☞ Check Your Progress 3

1) Analyze the Multiple Regression Summary output of Two independent variable $X_1 \& X_2$ and dependent variable Y given below, and determine what percentage of variation in$Y$ is explained by $X_1 \& X_2$.

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.840726494
R Square	0.705680867
Adjusted R Square	0.639801446
Standard Error	1.330891938
Observations	5

.....

.....

.....

.....

2) Consider the following multiple regression analysis results for two independent variable X_1 & X_2 and dependent variable Y,

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	- 5.51395602	0.711022732	0.353569008	0.747049498
X_1	- 39.330088	0.085205836	4.615891325	0.019133411
X_2	6.2953947	0.074635551	0.843484722	0.460900097

Determine the Multiple Regression equation and explain how X_1 & X_2 are related to Y.

.....

.....

.....

.....

☞ **Lab Sessions 5**

Perform the following using spreadsheet package.

1) Is there a relationship between moderate milk consumption and heart disease rate? The table underneath provides data from 6 developed countries from various cultures.

Country	A	B	C	D	E	F
Liters of milk per year per capita (x)	25	24	8	79	18	65
Deaths from hearth disease per 100,000 people per year (y)	211	191	297	107	167	86

2) The data in the table below show the percent of people who purchase their music from the internet (with 1997 corresponding to t = 0). Calculate the equation of the regression line and predict what percent of people will purchase their music from the internet in 2007 if this model is correct.

Year	0	1	2	3	4	5
Percent	0.3	1.1	2.4	3.2	2.9	3.4

3) Implement questions of Check Your Progress using spreadsheet, wherever possible.

2.5 SUMMARY

The practical sessions covered in this unit must have enabled you to broaden your understanding of correlation and regression, which you gained in BCS 040, through practical implementation using MS EXCEL Data Analysis ToolPak. It is important to understand that mere usage of MS Excel or any other software will not serve the purpose unless and until you possess an understanding of the subject. It is worth noting that regression analysis finds extensive application in a wide range of field of research.

1.7 ANSWERS TO CHECK YOUR PROGRESS

Check Your Progress-1

a) Scatter plot shows that data is highly positively correlated. Recall that the value $r$is approximately close to 0.9
b) Since $r = 0.96$, thus variation in one variable directly affects the other. In other words, increase or decrease in RAM size will result in a corresponding increase or decrease in system performance.
c) Yes, we can use the collected data to construct a linear regression model and hence use for the purpose of forecasting. This is justified from the high value of correlation coefficient$r = 0.96$.

Check Your Progress-2

a) Both factors are seen to be highly interdependent.
b) Unit increase in RAM size will lead to 0.69 times increase in systems performance.
c) Unit improvement in systems performance requires 0.54 times of alteration in RAM
d) 25 times of RAM change will lead to improve system performance by 0.69 X 25 times

Check Your Progress-3

1) 70.6% of variation in Y is explained by $X_1 \& X_2$.
2) $Y = -5.51 - 39.33\ X_1 + 6.295\ X_2$

Tips for solving Session 5

1)
The coefficient of correlation is given by $\mathrm{r} = \frac{n\sum \mathrm{X_i Y_i} - (\sum \mathrm{X_i})(\sum \mathrm{Y_i})}{\sqrt{n\sum \mathrm{X_i^2} - (\sum \mathrm{X_i})^2}\sqrt{n\sum \mathrm{Y_i^2} - (\sum \mathrm{Y_i})^2}}$

so that in our case we get$\mathrm{r} = \frac{6.29284 - 219\times 1059}{\sqrt{6.12055 - (219)^2}\sqrt{6.215945 - (1059)^2}} = -0.863$

2)
Let us calculate the coefficient of correlation using the formula in Exercise 1:

$$\mathrm{r} = \frac{6 \times 44.1 - 15 \times 13.3}{\sqrt{6 \times 55 - (15)^2}\sqrt{6 \times 37.27 - (13.3)^2}} = 0.929$$

Since there is high positive correlation between the variables, we calculate the coefficients of the least squares regression line as below:
Slope $b_{yx} = \frac{n\sum \mathrm{X_i Y_i} - (\sum \mathrm{X_i})(\sum \mathrm{Y_i})}{n\sum \mathrm{X_i^2} - (\sum \mathrm{X_i})^2} = \frac{6\times 44.1 - 15\times 13.3}{6\times 55 - (15)^2} = 0.62$ and
Intercept$c = \left(\bar{y} - b_{yx} \times \bar{x}\right) = \frac{1}{n}\left(\sum \mathrm{Y_i} - b_{yx}\sum \mathrm{X_i}\right) = \frac{1}{6}(13.3 - 0.62 \times 15) = 0.67.$

The equation of the least squares regression line thus becomes$y = 0.67 + 0.62\ x$.
We calculate the percentage of people that purchase their music on the internet in the year 2007 by substituting $x = 2007 - 1997 = 10$ into the least squares regression line. We get$y = 0.67 + 0.62 \times 10 = 6.9\%$.