Multiple Linear Regression

SESSION 12

Structure

12.1 INTRODUCTION

In Lab Session 11, you have learnt how to determine the linear relationship between a response variable and one predictor variable, obtain the coefficient of determination, test the significance of the regression coefficients, find the confidence interval, carry out residual analysis, and so on. In Unit 11 of MSTE-002 (Industrial Statistics-II), you have studied the manual fitting of multiple linear regression model when we have more than one predictor variables and other related statistical analysis.

Prerequisite

- Lab Session 11 of MSTL-002 (Industrial Statistics Lab).
- Units 11 of MSTE-002 (Industrial Statistics-II).

In this lab session, we generalise Lab Session 11 using two or more predictor variables with the help of MS Excel 2007. We explain how to construct the scatter matrix, fit the multiple regression model from the **Data Analysis ToolPak** and also use the matrix approach in Excel 2007. We also use the residual analysis to check the **linearity** assumption of the model and the normal probability plot to check the **normality** assumption just as we did for simple regression analysis.

Objectives

After performing the activities of this session, you should be able to:

- prepare the spreadsheet for multiple regression analysis in MS Excel 2007;
- estimate the model parameters using the method of least squares;
- fit the multiple linear regression model;
- test the significance of the regression parameters;
- determine the confidence interval of the regression parameters;
- construct the residual plot and normal probability plot;
- fit and analyse the multiple regression model using matrix approach; and
- interpret the results of multiple regression analysis.

12.2 PROBLEM DESCRIPTION

Suppose a juice manufacturing company wants to evaluate the effect of the factors such as advertisement cost and price of juice on monthly sales. For this purpose, the following data on the monthly sales, monthly advertisement cost and price per litre juice were obtained for 40 months to explore the relationship of sales with advertisement cost and price of juice.

Table 1: Data of monthly sales, advertisement cost and price of juice

S. No.	Sales (₹'000)	Advertisement Cost (₹'00)	Price (₹ per litre)	S. No.	Sales (₹'000)	Advertisement Cost (₹'00)	Price (₹ per litre)
1	15400	290	80	21	23900	350	75
2	27800	400	74	22	19000	350	78
3	21200	370	79	23	19500	390	81
4	31400	520	73	24	22100	400	76
5	35900	560	72	25	17500	340	80
6	31800	480	73	26	11200	280	89
7	21400	330	80	27	12400	280	87
8	15500	310	81	28	20700	390	79
9	11200	270	88	29	10900	250	89
10	32100	540	74	30	18400	360	82
11	22100	330	75	31	27400	400	76
12	17800	340	87	32	23000	360	84
13	26000	400	76	33	29100	420	74
14	23400	380	78	34	17400	270	82
15	27600	440	79	35	33400	510	72
16	26100	440	73	36	37400	570	70
17	24200	430	81	37	30200	500	82
18	26400	400	73	38	35500	560	74
19	20000	350	84	39	17700	350	78
20	24600	410	81	40	11500	210	87

Using this data,

- Draw a scatter diagram. Does this relationship appear to be linear?
- Do the regression analysis through Excel and interpret the results.
- Draw the fitted regression line on the scatter plot. Does a regression line appear to give a good fit here?
- Draw the residual curve and normal probability plot.

12.3 PROCEDURE FOR MULTIPLE REGRESSION ANALYSIS

In Unit 11 of MSTE-002, you have learnt the manual computation of the multiple regression analysis when we have two independent variables using the method of least squares. If we have more than two independent variables, we use the matrix approach in manual calculation, which will be discussed in Sec. 12.9. In this section, we consider two independent variables. This process consists of the following steps:

Step 1: Let us consider a multiple linear regression model with two independent variables X_1 and X_2 which are related with a response variable Y. The equation of the multiple regression line is given as

$$Y = B_0 + B_1X_1 + B_2X_2 + e \qquad \ldots (1)$$

where B_0 is the intercept, B_1 and B_2 are the partial regression coefficients corresponding to the independent variables X_1 and X_2, and e is a normally distributed random error component with mean zero and variance σ^2.

Step 2: The regression parameters can be estimated using the normal equations as discussed in Units 9 and 11 of MSTE-002. Using the method of least squares, we get the normal equations for two independent variables as follows:

✓ $$\sum Y_i = nB_0 + B_1\sum X_{1i} + B_2\sum X_{2i} \qquad \ldots(2)$$

✓ $$\sum X_{1i}Y_i = B_0\sum X_{1i} + B_1\sum X_{1i}^2 + B_2\sum X_{1i}X_{2i} \qquad \ldots(3)$$

✓ $$\sum X_{2i}Y_i = B_0\sum X_{2i} + B_1\sum X_{2i}X_{1i} + B_2\sum X_{2i}^2 \qquad \ldots(4)$$

Step 3: Therefore, the fitted multiple linear regression model is given as

$$\hat{Y} = \hat{B}_0 + \hat{B}_1X_1 + \hat{B}_2X_2 \qquad \ldots(5)$$

12.4 MATRIX PLOT IN EXCEL 2007

In the data given in Sec. 12.2, the monthly sales depend upon the advertisement cost and price. So the dependent variable is monthly sales and the independent variables are advertisement cost and price of the product.

Here Y is monthly sales, X_1, advertisement cost and X_2, price of the product. By following the steps given below, you can construct the matrix plot in Excel 2007:

Step 1: We enter the given data in Excel spreadsheet. The sheet with monthly sales data is shown partially in Fig. 12.1.

Dependent Y variable

Independent X variables, located together in the Excel sheet

	A	B	C	D
1	S.No.	Sales (₹ '000) Y	Advertisement Cost (₹ '00) X1	Price (₹ per litre) X2
2	1	15400	290	80
3	2	27800	400	74
4	3	21200	370	79
5	4	31400	520	73
6	5	35900	560	72
7	6	31800	480	73
8	7	21400	330	80
9	8	15500	310	81
10	9	11200	270	88
11	10	32100	540	74
12	11	22100	330	75
13	12	17800	340	87
14	13	26000	400	76
15	14	23400	380	78
16	15	27600	440	79
17	16	26100	440	73
18	17	24200	430	81
19	18	26400	400	73
20	19	20000	350	84
21	20	24600	410	81
22	21	23900	350	75
23	22	19000	350	78
24	23	19500	390	81
25	24	22100	400	76
26	25	17500	340	80

Multiple regression

Fig. 12.1: Partial screenshot of the spreadsheet for the given data.

The procedure for multiple regression analysis is the same as for simple linear regression analysis. Note that the only change from simple linear regression analysis is to include more than one column in the ***Input X Range***. In multiple regression, we need to put all regressors in **adjacent columns** (Columns C and D here). If they are not given in **adjacent columns**, you should always remember to keep all the regressors in the adjacent columns.

Step 2: Most of the statistical softwares have an option for the scatter matrix plot. In Excel, there is no built-in option for matrix plot. We need to put in extra effort to create a scatter matrix plot. In Excel 2007, we create separate scatter plots for (X_1, Y), (X_2, Y), (Y, X_1), (X_2, X_1), (Y, X_2) and (X_1, X_2). Then we put them together to obtain the scatter matrix plot.

For this purpose, we select the ***Scatter with only Markers*** from the ***Insert*** tab as shown in Fig. 12.2 without selecting the data.

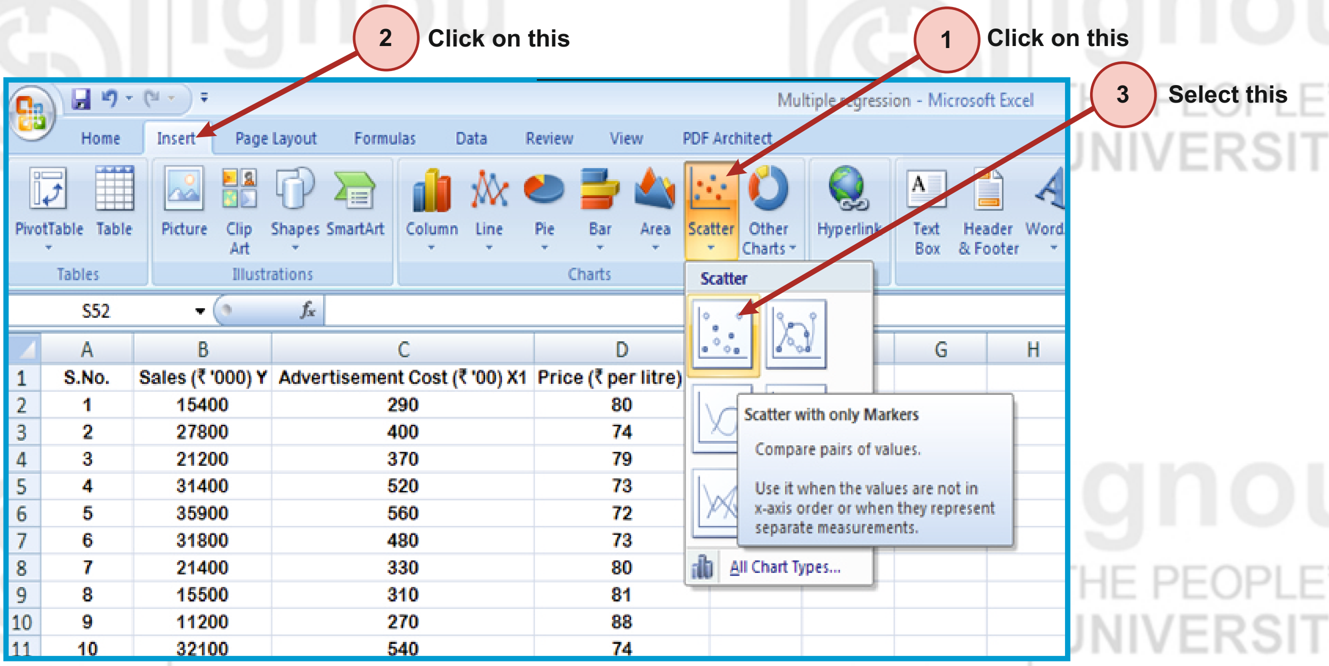


Fig. 12.2

If you click on any cell of the sheet, the tabs under ***Chart Tools*** will disappear. To get it back just click on the chart.

Step 3: We get a blank chart on the Excel sheet as shown in Fig. 12.3. We choose the ***Select Data*** option from ***Design*** tab under ***Chart Tools*** as shown in Fig. 12.3.

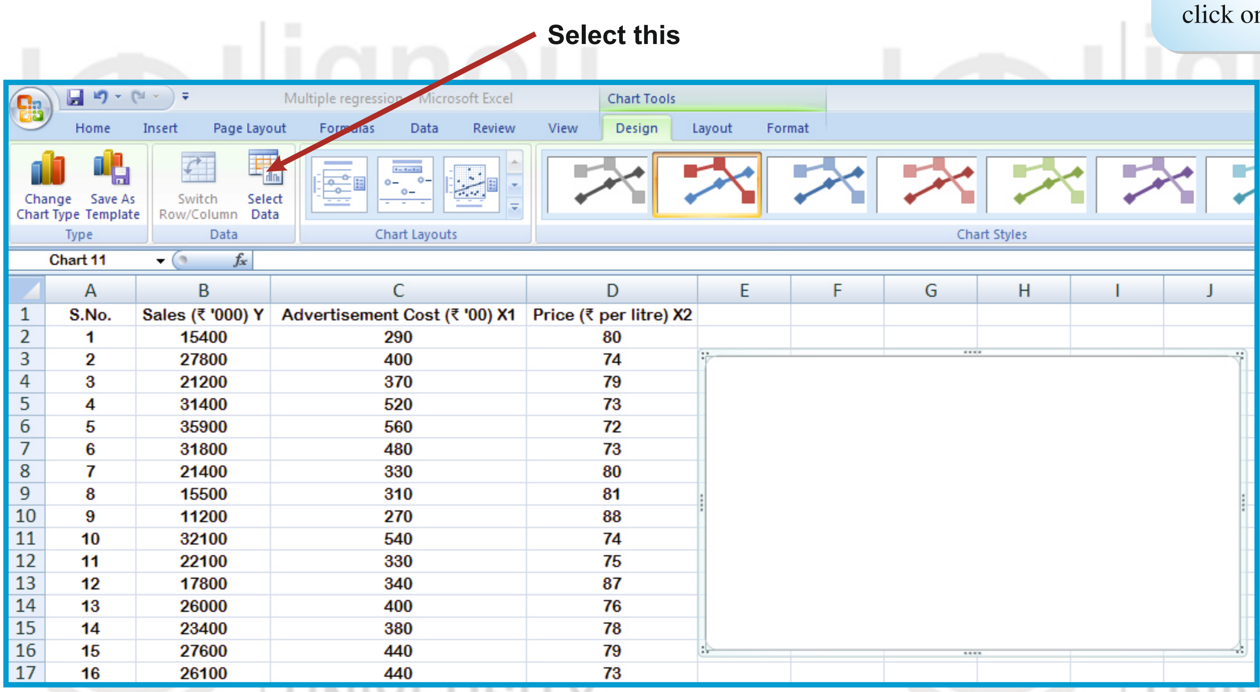


Fig. 12.3

Step 4: A new dialog box is opened (Fig. 12.4a). We now

1. click on ***Add*** button as shown in Fig. 12.4a,
2. select Cells C2:C41 as ***Series X values***,
3. select Cells B2:B41 as ***Series Y values***, and
4. click on ***OK*** as shown in Fig. 12.4b.

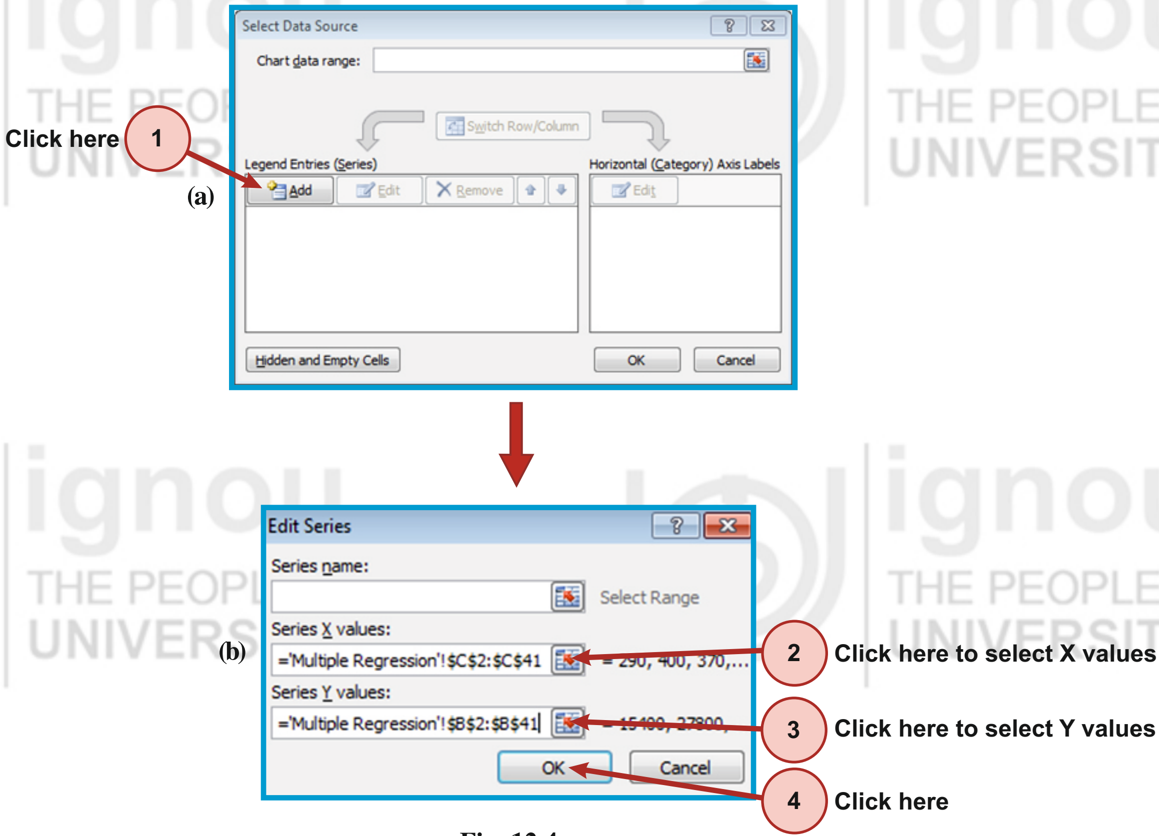


Fig. 12.4

Step 5: This gives us a scatter diagram having sales (Y) on the Y-axis and advertisement cost (X_1) on the X-axis. The resulting chart may appear as a "**Scatter with Lines**" as shown in Fig. 12.5.

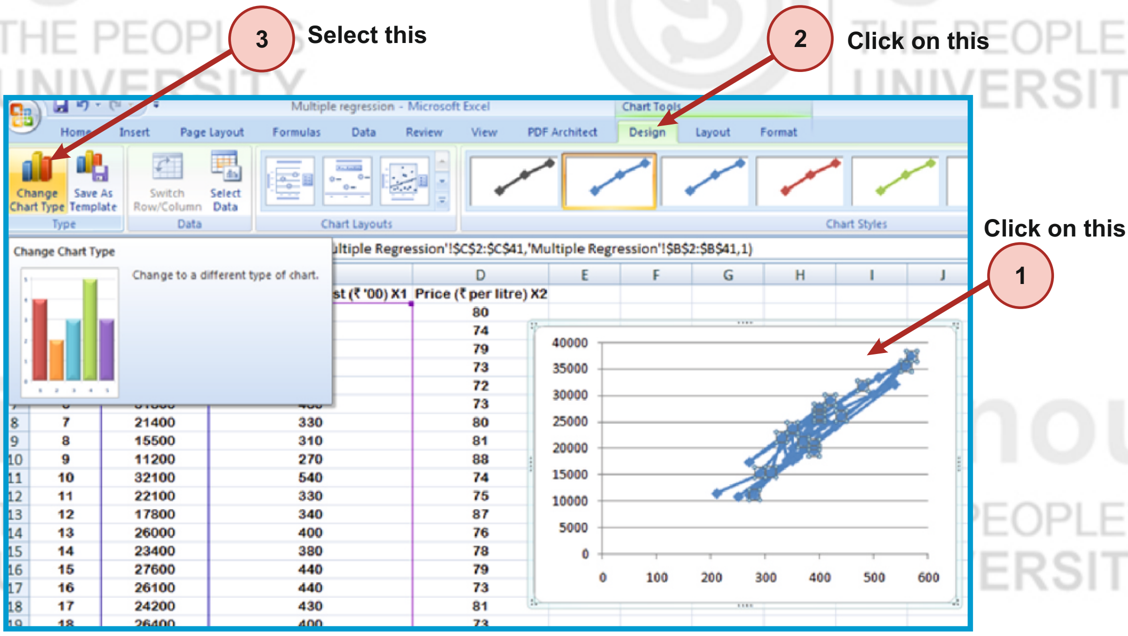


Fig. 12.5

Step 6: To make a chart **with only markers**, we

1. click on the chart shown in Fig. 12.5,
2. click on the ***Design*** tab,
3. select the ***Change Chart Type*** option as shown in Fig. 12.5, and
4. select ***Scatter with only Markers*** as shown in Fig. 12.6.

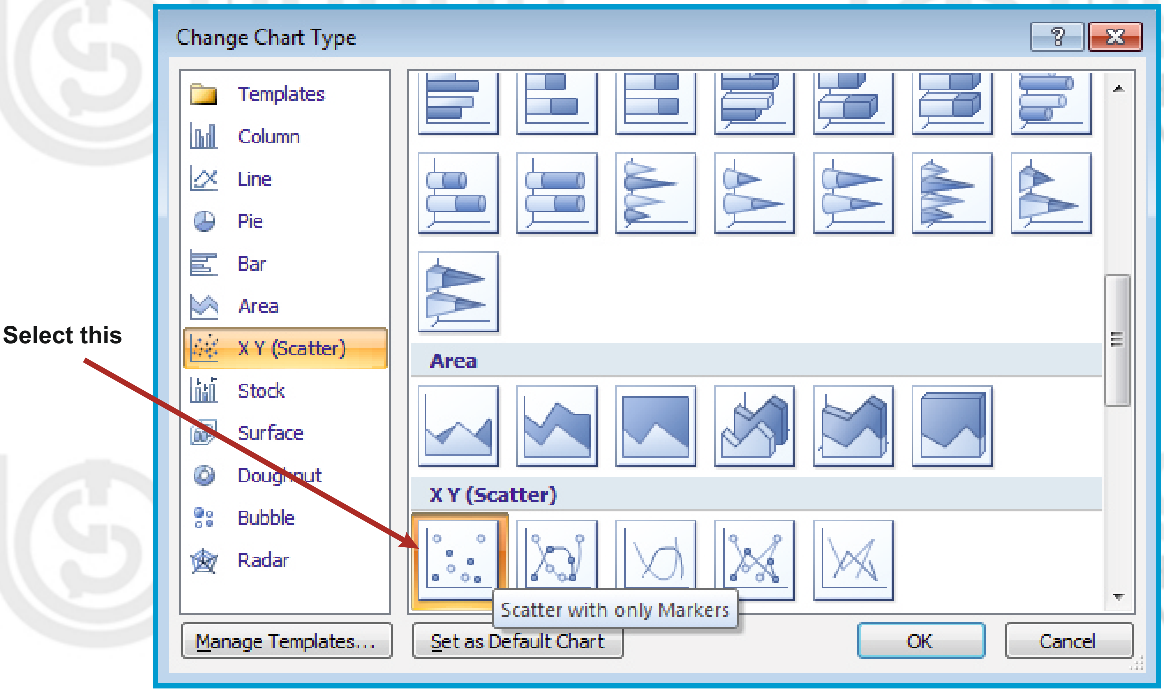


Fig. 12.6

Step 7: We format the chart as explained in Lab Session 3 of MSTL-001 (Basic Statistics Lab) as follows:

✓ Eliminate the grid lines and border around the chart.

✓ Change the marker option, size and colour.

✓ Also change the fonts, axes, background, etc., as desired.

The resulting scatter plot for (X_1, Y) is shown in Fig. 12.7.

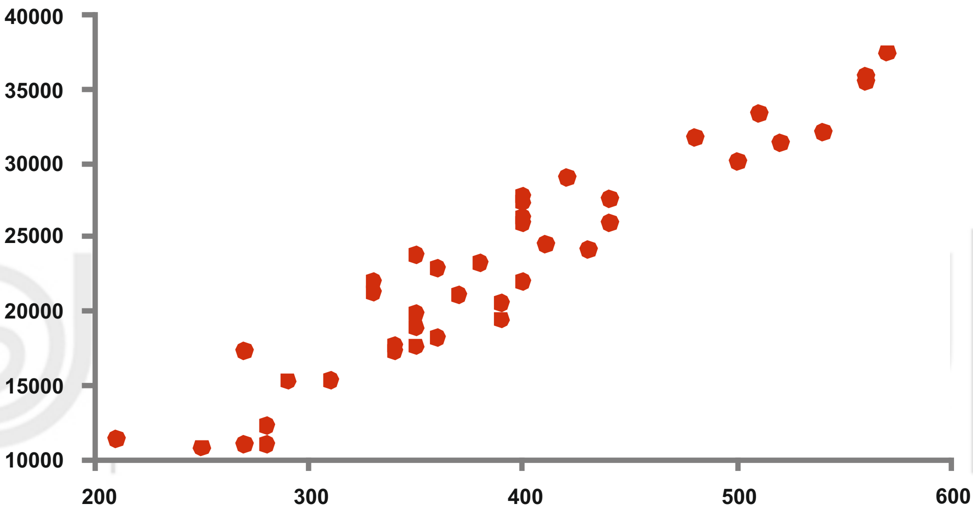


Fig. 12.7

Step 8: Similarly, we create the scatter plots for other combinations, i.e., (X_2, Y), (Y, X_1), (X_2, X_1), (Y, X_2) and (X_1, X_2) by considering the first variable as ***Series X values*** and the second variable as ***Series Y values***. After creating all scatter plots, we put them together as shown in Fig. 12.8 so that they look like a scatter plot

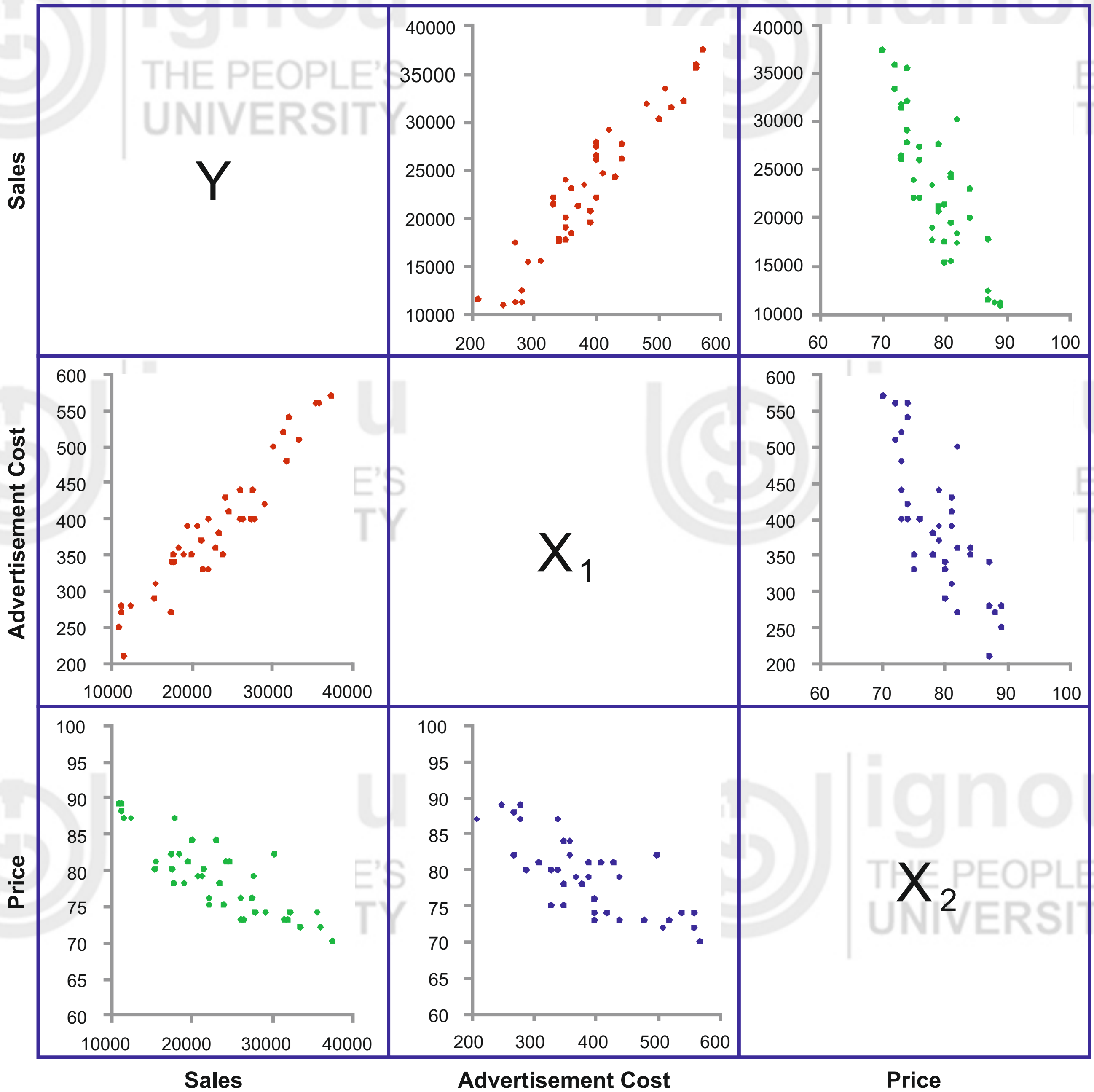


Fig. 12.8: Matrix of scatter plots for the monthly sales data.

Interpretation

Fig. 12.8 reveals that the scatter plot matrix is a two dimensional array (3×3 table) of the scatter plots where every cell except the diagonal cells contains the respective scatter plot. The first row gives you an idea about the relationship of monthly sales with advertisement cost and price in second and third columns, respectively. The second row shows the relationship of advertisement cost with the monthly sales and price in first and third columns, respectively, while the third row illustrates the relationship of price with sales and advertisement in first and second columns, respectively. Notice that the scatter plot matrix clearly shows the approximate linear relationship among all three variables.

In the next section, you will learn how to use MS Excel 2007 to fit the multiple regression model. This is far easier as compared to the manual calculation done in Unit 11 of MSTE-002.

12.5 FITTING AND ANALYSIS OF MULTIPLE REGRESSION MODEL IN EXCEL 2007

We use the ***Data Analysis ToolPak*** for multiple regression analysis in Excel. You have already carried out simple regression analysis using ***Data Analysis ToolPak*** with Excel 2007 in Lab Session 11.

We fit a multiple linear regression model on the given data by choosing ***Data Analysis*** under the ***Data*** tab and subsequently selecting ***Regression*** as discussed in Step 4 of Lab Session 11. Now refer to Fig. 12.9. We

- ✓ specify the data and label for monthly sales in ***Input Y Range***, i.e., (Cells B1:B41),
- ✓ specify the data and labels for the monthly advertisement cost and price in ***Input X Range***, i.e., (Cells C1:D41),
- ✓ check the ***Labels*** box (since we have included data labels in the input ranges),
- ✓ specify confidence interval under ***Confidence Level*** if we wish to calculate it for any value other than 95% (which is the default value). Here we have specified it as 98%,
- ✓ provide a new worksheet name under ***Output options*** (here we have used the name "**Analysis**" for the output sheet), and
- ✓ check the ***Residuals***, ***Standardised Residuals*** and ***Normal Probability Plot*** boxes and then click on ***OK***.

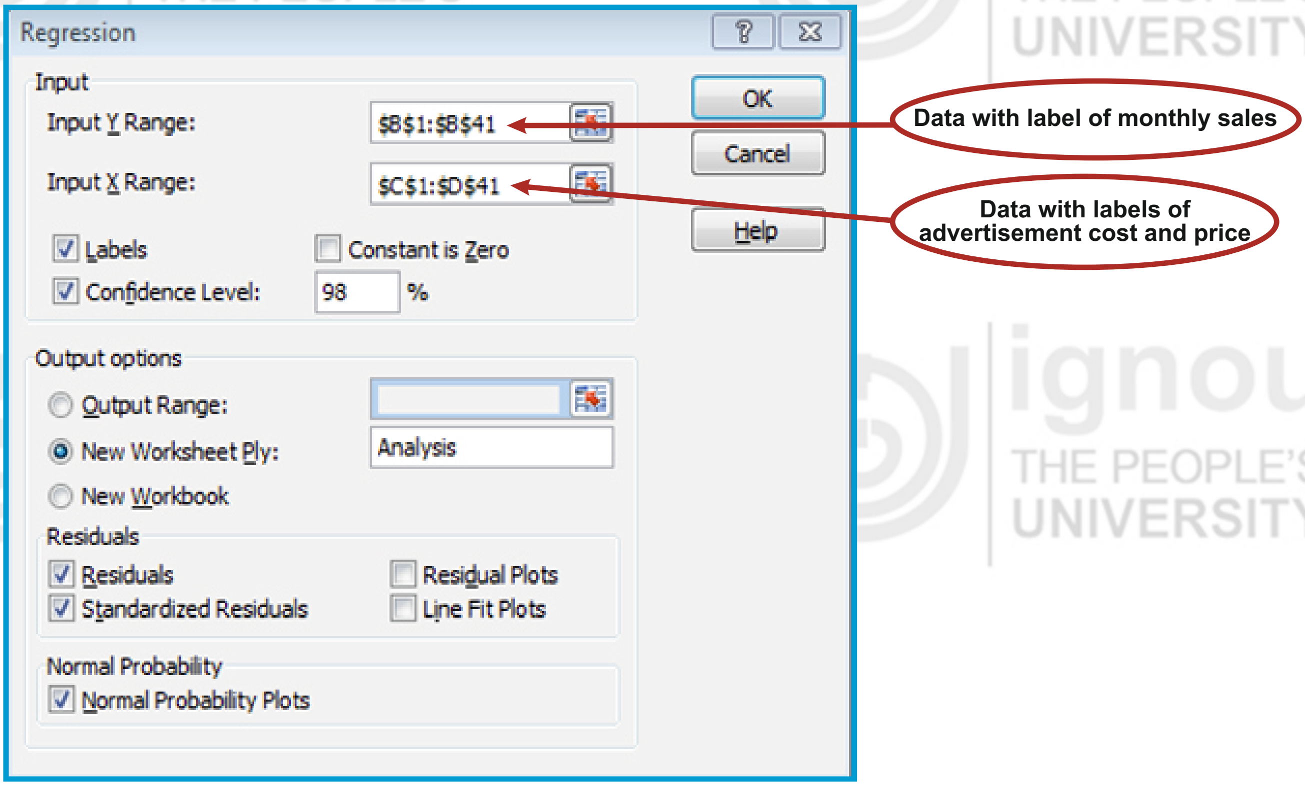


Fig. 12.9

The resulting worksheet is shown in Fig. 12.10.

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.9673							
5	R Square	0.9356							
6	Adjusted R Square	0.9321							
7	Standard Error	1878.6585							
8	Observations	40							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	2	1897523511.9554	948761755.9777	268.8199	0.0000			
13	Residual	37	130586238.0446	3529357.7850					
14	Total	39	2028109750.0000						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 98.0%*	*Upper 98.0%*
17	Intercept	31277.9623	8572.2838	3.6487	0.0008	13908.8656	48647.0591	10434.9051	52121.0195
18	Advertisement Cost (₹ '00) X1	58.9278	5.1102	11.5314	0.0000	48.5735	69.2820	46.5026	71.3530
19	Price (₹ per litre) X2	-394.8564	88.3289	-4.4703	0.0001	-573.8278	-215.8849	-609.6236	-180.0892
20									
21									
22									
23	RESIDUAL OUTPUT					PROBABILITY OUTPUT			
24									
25	*Observation*	*Predicted Sales (₹ '000) Y*	*Residuals*	*Standard Residuals*	*Ordered Residuals*	*Percentile*	*Sales (₹ '000) Y*		
26	1	16778.5047	-1378.5047	-0.7533	-3403.8836	1.25	10900		
27	2	25629.6976	2170.3024	1.1861	-2776.4252	3.75	11200		
28	3	21887.5826	-687.5826	-0.3758	-2739.9848	6.25	11200		

Analysis | Multiple Regression

Fig. 12.10: Partial screenshot of the output sheet "Analysis".

12.6 INTERPRETATION OF THE RESULTS OF MULTIPLE REGRESSION ANALYSIS

The interpretation of the results of regression analysis given in Fig. 12.10 is the same as discussed in Sec. 11.5 of Lab Session 11. The output of multiple regression also has four components as shown in Fig. 12.10:

1. Regression Statistics table
2. ANOVA table
3. Regression Coefficients table
4. Residual Output

We repeat the interpretation for all components so that you can grasp it better.

12.6.1 Interpretation of the Regression Statistics Table

The first part of the output given in Fig. 12.11 shows Multiple R = 0.9673 (Cell B4). It means that the multiple correlation coefficient between sales and the joint effect of advertisement cost and price is 0.9673. We can say that sales are highly positively correlated with both advertisement cost and price.

	A	B
1	SUMMARY OUTPUT	
2		
3	*Regression Statistics*	
4	Multiple R	0.9673
5	R Square	0.9356
6	Adjusted R Square	0.9321
7	Standard Error	1878.6585
8	Observations	40
9		

Fig. 12.11

R^2 = 0.9356 given in Cell B5 means that 93.56% of variation in Y is explained by the regressors X_1 and X_2, i.e., 93.56% of the variability in the monthly sales is accounted for by the regression model.

When we have more than one regressor, it is better to use R^2_{adj} than R^2. The value of adjusted R^2, i.e., R^2_{adj} is 0.9321 given in Cell B6 and it shows the variation in Y explained by the regressors X_1 and X_2, i.e., 93.21% of the variability in the monthly sales is explained by the regression model.

The standard error here refers to the estimated standard deviation of the error term e, i.e., $\hat{\sigma}$ and is given in Cell B7. From Fig. 12.11, $\hat{\sigma} = \sqrt{MSS_{Res}} = 1878.6585$.

12.6.2 Interpretation of the ANOVA Table

In the case of simple regression, the testing of overall regression can also be done by the t-statistic. But, in the case of multiple regression, ANOVA (analysis of variance) is a must for testing the significance of the overall regression model.

	A	B	C	D	E	F
10	ANOVA					
11		*df*	*SS*	*MS*	*F*	*Significance F*
12	Regression	2	1897523511.9554	948761755.9777	268.8199	0.0000
13	Residual	37	130586238.0446	3529357.7850		
14	Total	39	2028109750.0000			
15						

Fig. 12.12

The output given in Fig. 12.12 shows the ANOVA table for the monthly sales data. Let us consider the null hypothesis $H_0: B_1 = B_2 = ... = B_p = 0$ against the alternative hypothesis H_1: At least one of the regression coefficients is not zero.

From the ANOVA table output given in Fig. 12.12, the F-test statistic is 268.8199 given in Cell E12 and the associated p-value is 0.0000 given in Cell F12. Since the p-value for this test is less than 0.05 (i.e., 0.0000 < 0.05), we may reject the null hypothesis at 5% level of significance and conclude that the advertisement cost and price (regressors) are statistically significant in predicting the sales at 5% significance level. This implies that there is a linear relationship between sales and combination of advertisement cost and price.

12.6.3 Interpretation of the Regression Coefficients Table

The regression output shown in Fig. 12.13 gives the regression parameters and associated output as explained below:

- ✓ Cells A17, A18 and A19 represent the labels for B_0, B_1 and B_2, respectively, used for Columns B to I.
- ✓ Column "**Coefficient**" gives the least squares estimates of B_0, B_1 and B_2 in Cells B17, B18 and B19, respectively.
- ✓ Column "**Standard error**" gives the standard errors of the least square estimates of B_0, B_1 and B_2 in Cells C17, C18 and C19, respectively.
- ✓ Column "**t Stat**" gives the computed t-statistic for (i) $H_0: B_0 = 0$ against $H_1: B_0 \neq 0$, (ii) $H_0: B_1 = 0$ against $H_1: B_1 \neq 0$ and (iii) $H_0: B_2 = 0$ against $H_1: B_2 \neq 0$ in Cells D17, D18 and D19, respectively.
- ✓ Column "**P-value**" gives the p-values for the t-tests mentioned above.
- ✓ Columns "**Lower 95%**", "**Upper 95%**", "**Lower 98%**" and "**Upper 98%**" define 95% and 98% lower and upper confidence limits, respectively, for B_0, B_1 and B_2 in Cells F17:F19, G17:G19, H17:H19, and I17: I19, respectively.

	A	B	C	D	E	F	G	H	I
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 98.0%*	*Upper 98.0%*
17	Intercept	31277.9623	8572.2838	3.6487	0.0008	13908.8656	48647.0591	10434.9051	52121.0195
18	Advertisement Cost (₹ '00) X1	58.9278	5.1102	11.5314	0.0000	48.5735	69.2820	46.5026	71.3530
19	Price (₹ per litre) X2	-394.8564	88.3289	-4.4703	0.0001	-573.8278	-215.8849	-609.6236	-180.0892
20									

Fig. 12.13

We now use the output given in Fig. 12.13 to interpret the results as follows:

- **Fitting of the regression model**

  The values given in Cells B17, B18 and B19 correspond to the coefficients $\hat{B}_0$, $\hat{B}_1$ and $\hat{B}_2$, respectively.

  From Fig. 12.13, you can see that $\hat{B}_0 = 31277.9623, \hat{B}_1 = 58.9278$ and $\hat{B}_2 = -394.8564$.

  Hence, the fitted regression line from equation (5) is

  $$\hat{Y} = 31277.9623 + 58.9278 X_1 - 394.8564 X_2$$

  The ***Regression*** tool in ***Data Analysis ToolPak*** also produces the standard errors of $\hat{B}_0$, $\hat{B}_1$ and $\hat{B}_2$, which are

  $$SE(\hat{B}_0) = 8572.2838, \; SE(\hat{B}_1) = 5.1102 \text{ and } SE(\hat{B}_2) = 88.3289$$

- **Hypothesis testing for the regression coefficients**

  The ***Regression*** tool in ***Data Analysis ToolPak*** individually tests the hypothesis for each coefficient to be equal to zero. Here we use the p-value approach for decision making. We discuss them one at a time.

  i) $H_0: \hat{B}_0 = 0$ vs $H_1: \hat{B}_0 \neq 0$

     From Fig. 12.13, the t-statistic for the intercept $(\hat{B}_0)$ is 3.6487 (Cell D17) and its p-value is 0.0008 (Cell E17). Since the p-value is less than 0.05, we may reject our null hypothesis at 5% level of significance and conclude that the intercept is not equal to zero, i.e., the line of regression is not passing through the origin for the given data.

  ii) $H_0: \hat{B}_1 = 0$ vs $H_1: \hat{B}_1 \neq 0$

     From Fig. 12.13, the t-statistic for $\hat{B}_1$ is 11.5314 (Cell D18) and its p-value is 0.000 (Cell E18), which is less than 0.05. So we may reject our null hypothesis at 5% level of significance and conclude that the regression coefficient corresponding to X_1 is not equal to zero, i.e., the advertisement cost affects the monthly sales of the juice for the given data.

  iii) $H_0: \hat{B}_2 = 0$ vs $H_1: \hat{B}_2 \neq 0$

     From Fig. 12.13, the t-statistic for $\hat{B}_2$ is – 4.4703 (Cell D19) and its p-value is 0.0001 (Cell E18), which is less than 0.05. So we may reject our null hypothesis at 5% level of significance and conclude that the regression coefficient corresponding to X_2 is not equal to zero, i.e., the price also affects the monthly sales of the juice for the given data.

- **Confidence intervals for the regression coefficients**

We have discussed the (1–α)100% confidence intervals of the regression coefficients in Unit 11 of MSTE-002. Excel also provides us the confidence limits of regression coefficients.

From Excel output given in Fig. 12.13, 95% confidence interval for

- $\hat{B}_0$ is (13908.8656, 48647.0591),
- $\hat{B}_1$ is (48.5735, 69.2820), and
- $\hat{B}_2$ is (–573.8278, –215.8849).

If you want to find the confidence limits other than 95%, you can specify the desired confidence level as shown in Fig. 12.9. We have selected the ***Confidence Level Box*** and set the level to 98% in the ***Regression*** dialog box shown in Fig. 12.9 to find 98% confidence intervals.

From Fig. 12.13, 98% confidence interval for the regression coefficients $\hat{B}_0$, $\hat{B}_1$ and $\hat{B}_2$ are (10434.9051, 52121.0195), (46.5026, 71.3530) and (– 609.6236, – 180.0892), respectively.

12.6.4 Interpretation of the Residuals Output

You have learnt in Lab Session 11 that the **regression analysis** feature in Excel also includes other useful statistic(s) in the output such as residuals, standardised residuals and values needed for normal probability plot. We can check the adequacy of the fitted regression model using these outputs given in Fig. 12.10.

The predicted sales, residuals and the standard residuals are shown in Cells B25:B64, C25:C64 and D25:D64, respectively, in Fig. 12.14.

	A	B	C	D
23	RESIDUAL OUTPUT			
24				
25	*Observation*	*Predicted Sales (₹ '000) Y*	*Residuals*	*Standard Residuals*
26	1	16778.5047	-1378.5047	-0.7533
27	2	25629.6976	2170.3024	1.1861
28	3	21887.5826	-687.5826	-0.3758
29	4	33095.8862	-1695.8862	-0.9268
30	5	35847.8533	52.1467	0.0285
31	6	30738.7754	1061.2246	0.5800
32	7	19135.6154	2264.3846	1.2375
33	8	17562.2037	-2062.2037	-1.1270
34	9	12441.0983	-1241.0983	-0.6782
35	10	33879.5852	-1779.5852	-0.9725
36	11	21109.8974	990.1026	0.5411
37	12	16960.8985	839.1015	0.4586
38	13	24839.9848	1160.0152	0.6339
39	14	22871.7167	528.2833	0.2887
40	15	26012.5264	1587.4736	0.8675
41	16	28381.6647	-2281.6647	-1.2469
42	17	24633.5359	-433.5359	-0.2369
43	18	26024.5539	375.4461	0.2052
44	19	18734.7453	1265.2547	0.6915
45	20	23454.9806	1145.0194	0.6257
46	21	22288.4527	1611.5473	0.8807
47	22	21103.8836	-2103.8836	-1.1498
48	23	22276.4252	-2776.4252	-1.5173
49	24	24839.9848	-2739.9848	-1.4974
50	25	19724.8931	-2224.8931	-1.2159

Analysis | Multiple Regression

Fig. 12.14

12.7 RESIDUAL PLOT IN EXCEL 2007

To conduct the residual analysis, we plot the predicted sales values versus standard residuals shown in Fig. 12.14. For this purpose, we select the predicted sales data with label (Cells B25:B64) and standard residuals data with label (Cells D25:D64) by holding the ***Ctrl*** key.

We choose ***Scatter*** under the ***Insert*** tab to plot these residuals against the predicted sales values as discussed in Lab Session 11. We format the chart and change the ***Horizonal axis crosses*** at (–2.5) as discussed in Sec.11.6 of Lab Session 11. The resulting residual plot is shown in Fig. 12.15.

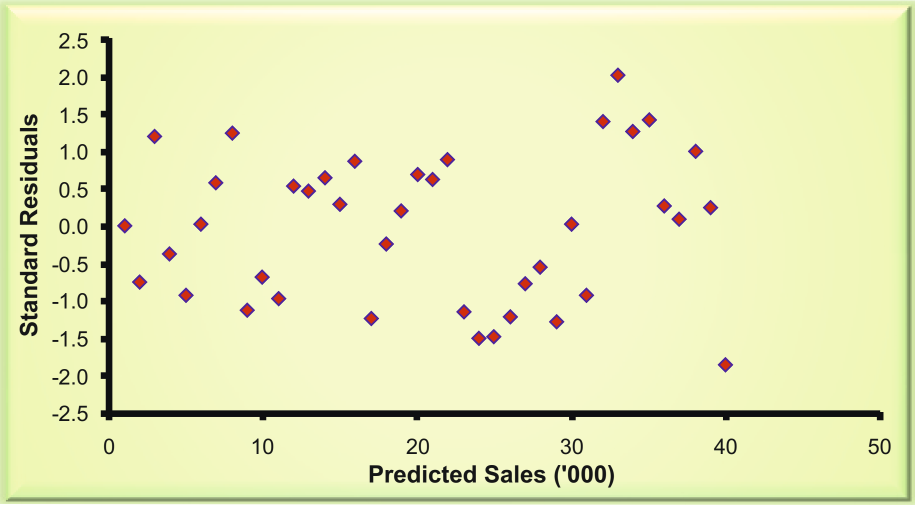


Fig. 12.15: Residual plot.

The standard residuals shown in Fig. 12.15 appear to be approximately evenly scattered (random) throughout a horizontal band around 0.0 (Y-axis). Hence, the assumptions of linear regression is valid or we can say that multiple linear regression model fitted well for the given data.

12.8 NORMAL PROBABILITY PLOT IN EXCEL 2007

Step 1: For plotting the normal probability plot, we follow the steps given in Sec. 11.7 of Lab Session 11 to sort the ordered residuals in ascending order. The output is shown in Fig. 12.16.

	E	F	G
23		PROBABILITY OUTPUT	
24			
25	*Ordered Residuals*	*Percentile*	*Sales (₹ '000) Y*
26	-3403.8836	1.25	10900
27	-2776.4252	3.75	11200
28	-2739.9848	6.25	11200
29	-2366.1380	8.75	11500
30	-2281.6647	11.25	12400
31	-2224.8931	13.75	15400
32	-2103.8836	16.25	15500
33	-2062.2037	18.75	17400
34	-1779.5852	21.25	17500
35	-1713.7357	23.75	17700
36	-1695.8862	26.25	17800
37	-1435.5196	28.75	18400
38	-1378.5047	31.25	19000
39	-1241.0983	33.75	19500

Analysis / Multiple Regression

Fig. 12.16

Step 2: We select Cells E25:F64 and plot a **scatter** diagram as discussed in Sec. 11.7 of Lab Session 11. The resulting normal probability plot is shown in Fig. 12.17.

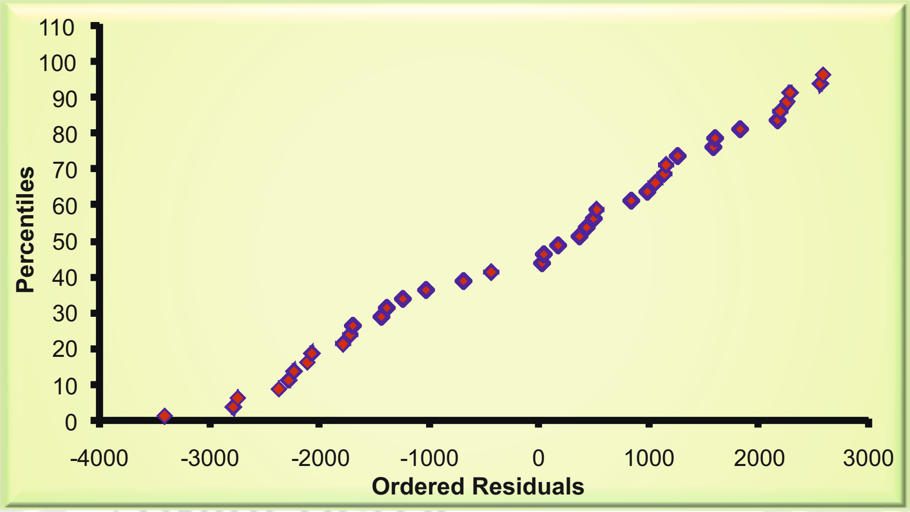


Fig. 12.17: Normal probability plot.

The normal probability plot shown in Fig. 12.17 reveals that the resulting points lie approximately along a straight line. It indicates that the distribution of error terms (residuals) is approximately normally distributed.

So far you have learnt how to fit and analyse the multiple linear regression in Excel 2007. We have already explained the matrix approach in Unit 11 of MSTE-02 when we have more than one predictor. In matrix approach, we write the normal equations in the matrix form and perform the regression analysis more conveniently. Although you can perform the multiple regression analysis using ***Regression*** available in ***Data Analysis ToolPak*** more quickly, we also discuss the matrix approach in Excel so that you can use it if need be.

But before proceeding further, you should do the following activity.

Activity 1

- Predict the sales for a particular month for which advertisement cost is ₹ 49000 and price is ₹ 78.
- Verify the residual property, i.e., $\sum_{i=1}^{n} e_i = 0$.

12.9 MATRIX APPROACH IN EXCEL 2007

In the case of two regressors, we find the estimated regression coefficients by solving normal equations as discussed in Unit 11 of MSTE-002. However, as the number of regressors increases, the complexity of the method also increases. In such situations, it is more convenient to use the matrix approach to do the regression analysis. The main steps involved in using the matrix approach are as follows:

Step 1: For p regressors, the regression model is given as follows:

$$Y = B_0X_0 + B_1X_1 + B_2X_2 + ... + B_pX_p + e \qquad ...(6)$$

where

B_0 – intercept,

B_i – i^{th} regression coefficient $(i = 1, 2, ..., p)$, and

e – error term with mean zero and variance σ^2.

Step 2: The equation of regression line in the matrix notation can be written as:

$$\mathbf{Y = X\,B + E} \qquad ...(7)$$

where

$$\mathbf{Y}_{\mathbf{n\times1}} = \begin{pmatrix} Y_1 \\ . \\ . \\ Y_n \end{pmatrix}, \qquad \mathbf{X}_{\mathbf{n\times(p+1)}} = \begin{pmatrix} 1 & X_{11} & . & X_{p1} \\ . & . & . & . \\ . & . & . & . \\ 1 & X_{1n} & . & X_{pn} \end{pmatrix},$$

$$\mathbf{B}_{(\mathbf{p}+1)\times1} = \begin{pmatrix} B_0 \\ B_1 \\ . \\ B_p \end{pmatrix}, \quad \text{and} \quad \mathbf{E}_{\mathbf{n\times1}} = \begin{pmatrix} e_1 \\ . \\ . \\ e_n \end{pmatrix}.$$

Step 3: Then (p+1) normal equations can be written as

$$\mathbf{X'Y = X'XB} \qquad ...(8)$$

where $\mathbf{X'}$– transpose matrix of $\mathbf{X}$

Step 4: If $\mathbf{X'X}$ is non-singular, i.e., ($\mathbf{X'X}$) is of rank (p+1), the least square estimators of $\mathbf{B}$, denoted by $\hat{\mathbf{B}}$, can be written as

$$\hat{\mathbf{B}} = \mathbf{(X'X)^{-1}X'Y} \qquad ...(9)$$

Step 5: The variance covariance matrix of regression coefficients is a $(p+1)\times(p+1)$ matrix and its diagonal elements give the variances of coefficients and the off diagonal elements give the covariances. If we denote

$$\left(\hat{\mathbf{B}}\right) = \hat{\sigma}^2\,(\mathbf{X'X})^{-1} = (\sigma_{jk}), \qquad j, k = 0, 1, ..., p \qquad ...(10)$$

then $V(\hat{B}_j) = \sigma_{jj}$, and $Cov(\hat{B}_j, \hat{B}_k) = \sigma_{jk}$, $(j \neq k)$ …(11)

Step 6: Standard error (SE) of $\hat{B}_j$ is given by

$$SE\left(\hat{B}_j\right) = \sqrt{\sigma_{jj}} \qquad …(12)$$

Step 7: Using the matrix notation, we can also create ANOVA table. You know that it classifies the total variation into the variation due to regression model and residuals (or error), which can be computed as follows:

i) Total sum of squares (with d.f. $= n-1$)

$$SS_T = \mathbf{Y}'\mathbf{Y} - n\bar{Y}^2 \qquad ...(13)$$

ii) Residual sum of squares (with d.f.$=n-p-1$)

$$SS_{Res} = \mathbf{Y}'\mathbf{Y} - \hat{B}'(X'Y) \qquad ...(14)$$

iii) Regression sum of squares (with d.f. = p)

$$SS_{Reg} = \hat{\mathbf{B}}'\mathbf{X}'\mathbf{Y} - n\bar{Y}^2 \qquad ...(15)$$

Step 8: We can calculate regression and residual mean sum of squares as

$$MSS_{Reg} = \frac{SS_{Reg}}{p} \qquad ...(16)$$

$$MSS_{Res} = \frac{SS_{Res}}{n-p-1} \qquad ...(17)$$

Step 9: F-statistic is given by

$$F_{Cal} = \frac{MSS_{Reg}}{MSS_{Res}} \qquad ...(18)$$

Step 10: The coefficient of determination (R^2) and adjusted coefficient of determination (R^2_{adj}) are given as

$$R^2 = 1 - \frac{SS_{Res}}{SS_T} \qquad ...(19)$$

$$R^2_{adj} = 1 - \frac{SS_{Res}/(n-p-1)}{SS_T/(n-1)} \qquad ...(20)$$

Step 11: We can also use the matrix approach for testing the null hypothesis for regression coefficient individually, i.e.,

$H_0: B_j = 0$ vs $H_1: B_j \neq 0$ for $j = 0, 1, 2, ..., p$. The t-statistic can be calculated as

$$t_j = \frac{\hat{B}_j}{SE(\hat{B}_j)}; \quad j = 0, 1, ..., p \qquad ...(21)$$

Step 12: $(1-\alpha)100\%$ lower and upper confidence limits for j^{th} $(j = 0, 1, ..., p)$ regression coefficient are

$$(B_j)_L = \hat{B}_j - t_{\alpha/2}\, SE(\hat{B}_j) \qquad ...(22)$$

$$(B_j)_U = \hat{B}_j + t_{\alpha/2}\, SE(\hat{B}_j) \qquad ...(23)$$

Steps in Excel

The main steps for carrying out multiple regression analysis using matrix approach in Excel 2007 are described below:

Step 1: We copy and paste the data typed in Excel sheet named "**Multiple Regression**" on a new worksheet and name it "**Matrix Approach**" (see Fig. 12.18).

	A	B	C	D
1	S.No.	Sales (₹ '000) Y	Advertisement Cost (₹ '00) X1	Price (₹ per litre) X2
2	1	15400	290	80
3	2	27800	400	74
4	3	21200	370	79
5	4	31400	520	73
6	5	35900	560	72
7	6	31800	480	73
8	7	21400	330	80
9	8	15500	310	81
10	9	11200	270	88
11	10	32100	540	74

Analysis | Multiple Regression | Matrix Approach

Fig. 12.18

Step 2: For using matrix approach in Excel, we need to insert a column before the regressors X_1 and X_2. To insert a column before Column C as shown in Fig. 12.18, we

1. select Column C,
2. click on ***Home*** tab, and
3. click on ***Insert*** as shown in Fig. 12.19a.

We get a new empty Column C as shown in Fig. 12.19b.

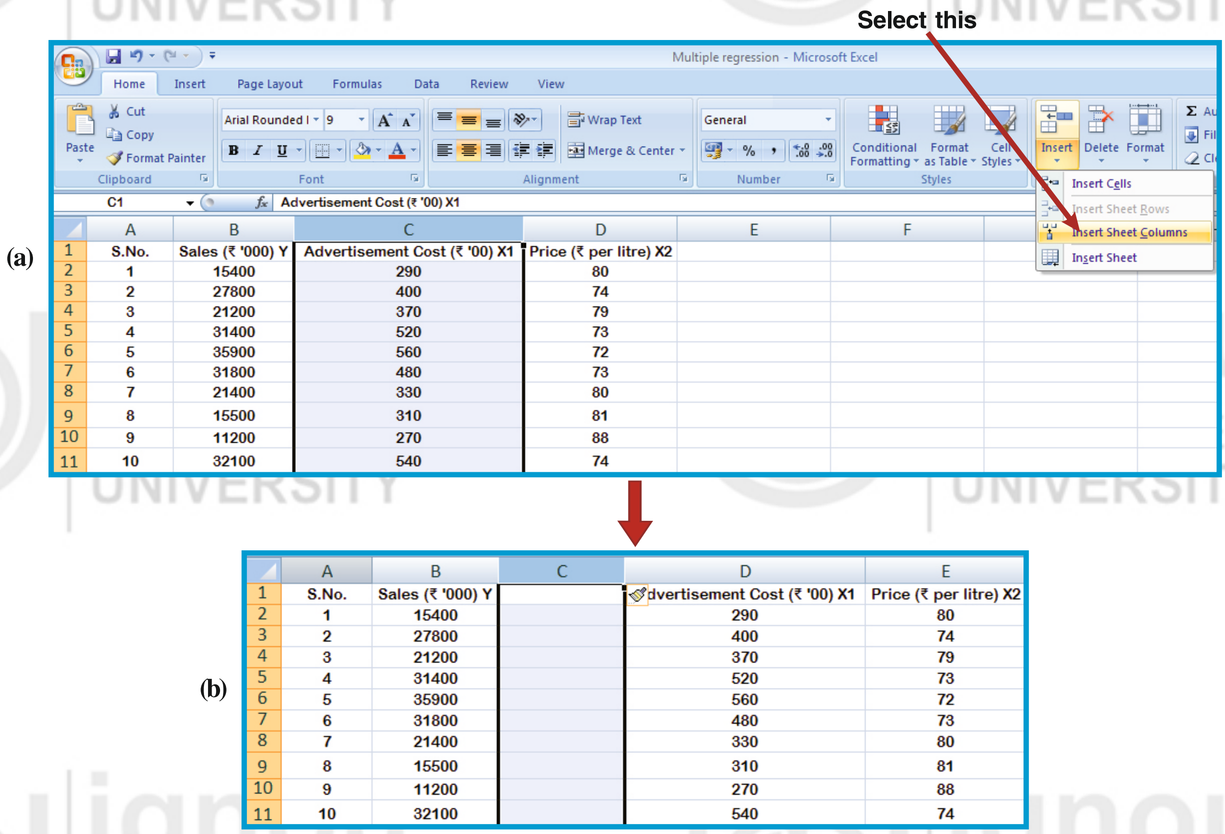
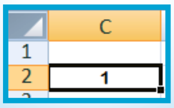
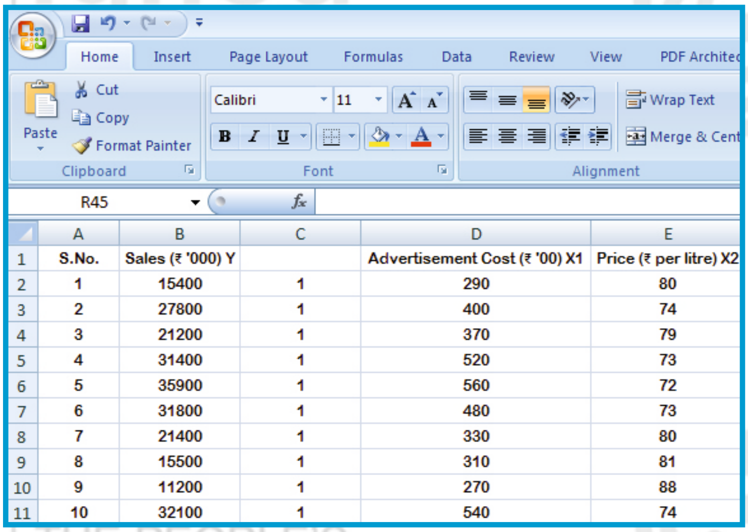


Fig. 12.19

Step 3: We type "1" in Cells C2:C41 as shown in Fig. 12.20.

Tips: Type 1 in Cell C2 only and drag it down up to Cell C41 to fill 1 in Cells C2:C41.

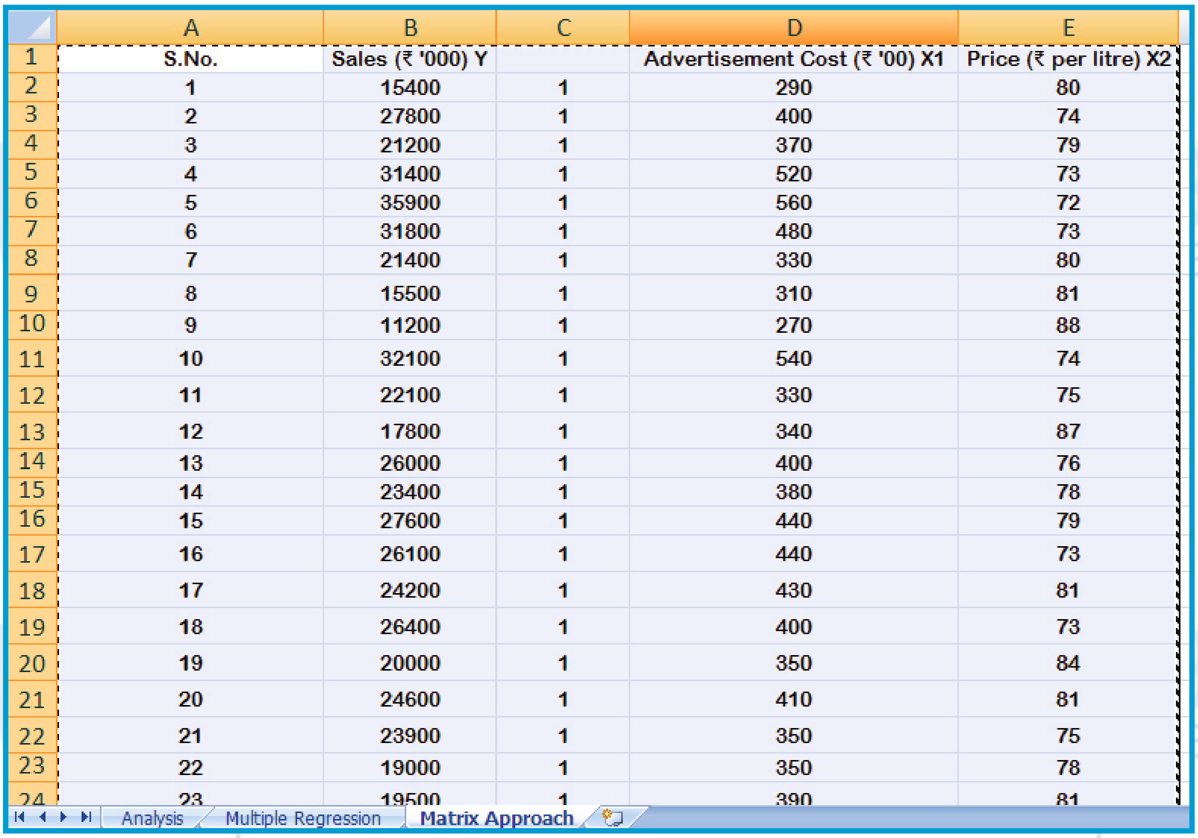




	A	B	C	D	E
1	S.No.	Sales (₹ '000) Y		Advertisement Cost (₹ '00) X1	Price (₹ per litre) X2
2	1	15400	1	290	80
3	2	27800	1	400	74
4	3	21200	1	370	79
5	4	31400	1	520	73
6	5	35900	1	560	72
7	6	31800	1	480	73
8	7	21400	1	330	80
9	8	15500	1	310	81
10	9	11200	1	270	88
11	10	32100	1	540	74

Fig. 12.20

Step 4: We have explained in Unit 11 of MSTE-002 that the matrix approach needs the **transpose** of Y values and X values. So we select and copy the data given in Cells A1:E41 as shown in Fig. 12.21.



	A	B	C	D	E
1	S.No.	Sales (₹ '000) Y		Advertisement Cost (₹ '00) X1	Price (₹ per litre) X2
2	1	15400	1	290	80
3	2	27800	1	400	74
4	3	21200	1	370	79
5	4	31400	1	520	73
6	5	35900	1	560	72
7	6	31800	1	480	73
8	7	21400	1	330	80
9	8	15500	1	310	81
10	9	11200	1	270	88
11	10	32100	1	540	74
12	11	22100	1	330	75
13	12	17800	1	340	87
14	13	26000	1	400	76
15	14	23400	1	380	78
16	15	27600	1	440	79
17	16	26100	1	440	73
18	17	24200	1	430	81
19	18	26400	1	400	73
20	19	20000	1	350	84
21	20	24600	1	410	81
22	21	23900	1	350	75
23	22	19000	1	350	78
24	23	19500	1	390	81

Fig. 12.21

Step 5: Refer to Fig. 12.22. We

1. select the Cell A45,
2. click on the ***Home*** tab,
3. click on the arrow below ***Paste*** option, and
4. click on the ***Transpose*** option.

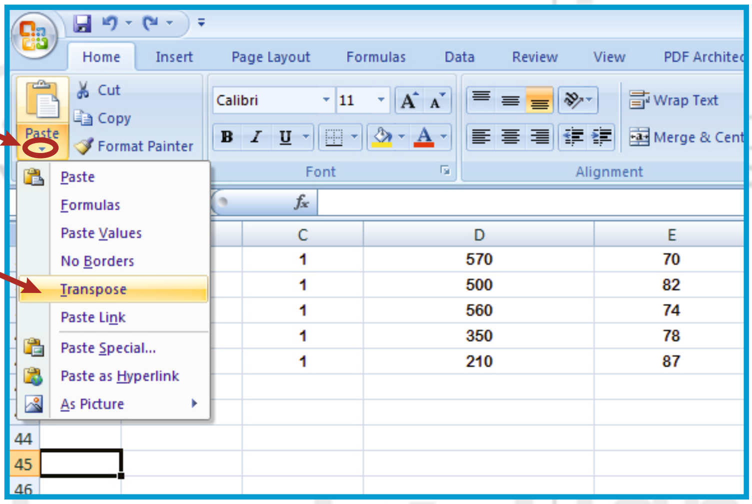


Fig. 12.22

Step 6: We get the transpose matrix of **Y** and **X** values, i.e., **Y′** and **X′**, respectively, given in Cells B46:AO46 and B47:AO49 (Fig. 12.23).

	A	B	C	D	...	AM	AN	AO
44								
45	S.No.	1	2	3		38	39	40
46	Sales (₹ '000) Y	15400	27800	21200		35500	17700	11500
47		1	1	1		1	1	1
48	Advertisement Cost (₹ '00) X1	290	400	370		560	350	210
49	Price (₹ per litre) X2	80	74	79		74	78	87
50								

Fig. 12.23

Step 7: For determining the regression coefficients B_0, B_1 and B_2, we need to obtain the matrices **X′Y, X′X** and inverse of **X′X**, i.e., $(\mathbf{X'X})^{-1}$. We first determine **X′Y**. We know that the matrices **X′** and **Y** are of order (3×40) and (40×1), respectively. So **X′Y** will be a (3×1) matrix. We now select 3 cells of our choice. Here we have selected three Cells G3:G5 (vertical cells).

We choose ***Mmult*** function from ***Math & Trig*** menu under ***Formulas*** tab. It gives us a new dialog box shown in Fig. 12.24.

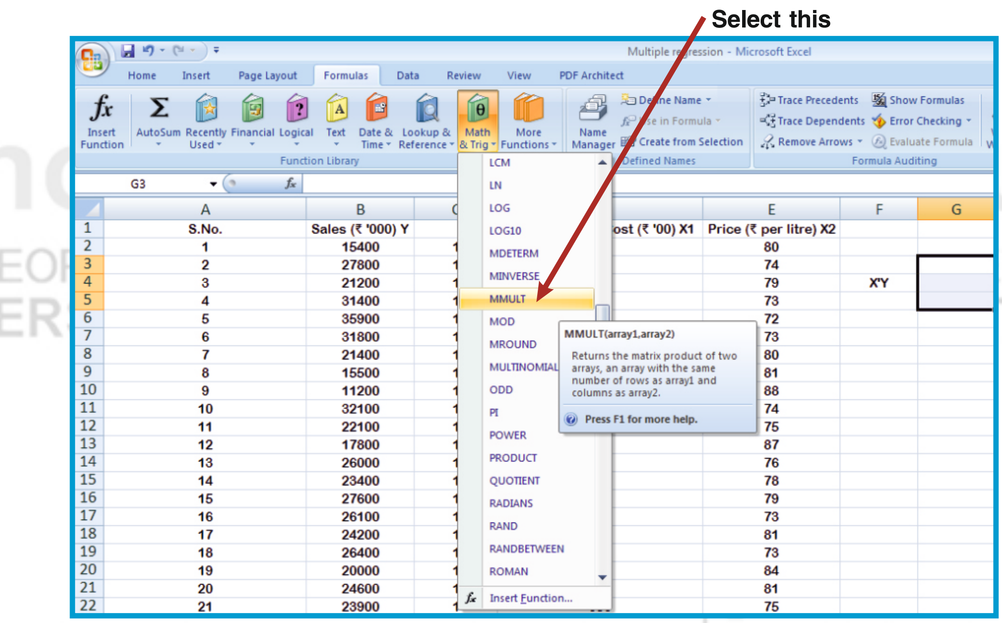


Fig. 12.24

Step 8: We select the range of $\mathbf{X}'$, i.e., Cells B47:AO49 as ***Array 1*** and range of **Y**, i.e., Cells B2:B41 as ***Array 2*** (Fig. 12.25). Note that we do not alter the order when dealing with matrices as it may change the results.

> To get the resulting matrix, press "***Ctrl*** + ***Shift*** + ***Enter***" together.
>
> Or, you can also press "***Ctrl*** + ***Shift*** + ***OK***" together.

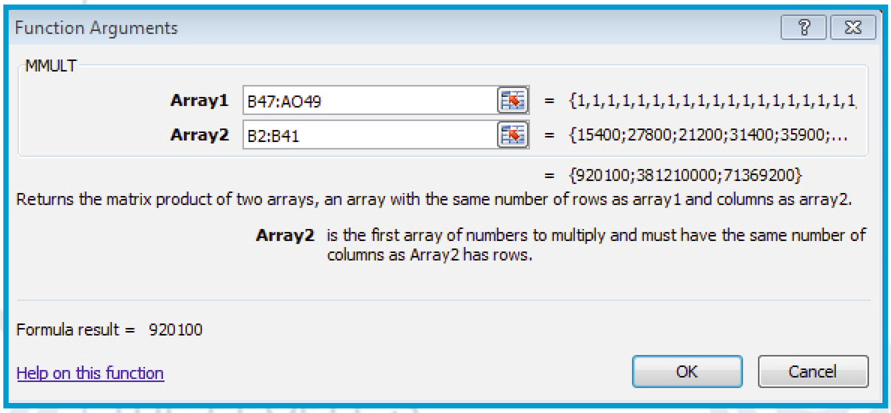


Fig. 12.25

Step 9: We now press (***Ctrl*** + ***Shift***) keys along with ***Enter*** key on the keyboard without leaving (***Ctrl*** + ***Shift***) keys. We obtain the $\mathbf{X}'\mathbf{Y}$ matrix in the form of an array (Fig. 12.26).

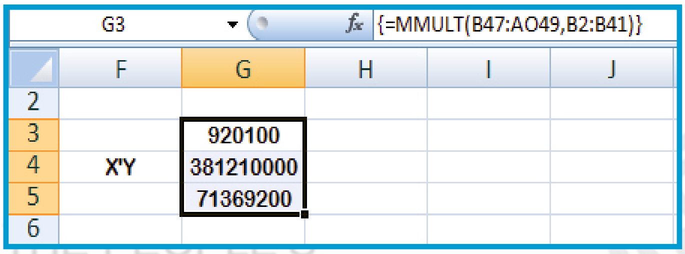


Fig. 12.26

Step 10: $\mathbf{X}'$ and $\mathbf{X}$ are matrices of orders (3×40) and (40×3), respectively. So $\mathbf{X}'\mathbf{X}$ is a (3×3) matrix. We select (3×3) cells (3 rows and 3 columns). Here we have selected Cells J3:L5 (Fig. 12.27). We use ***Mmult*** function and select Cells B47:AO49 and Cells C2:E41 as ***Array 1*** and ***2***, respectively, and press ***Enter*** along with (***Ctrl*** + ***Shift***) keys to get the $\mathbf{X}'\mathbf{X}$ matrix.

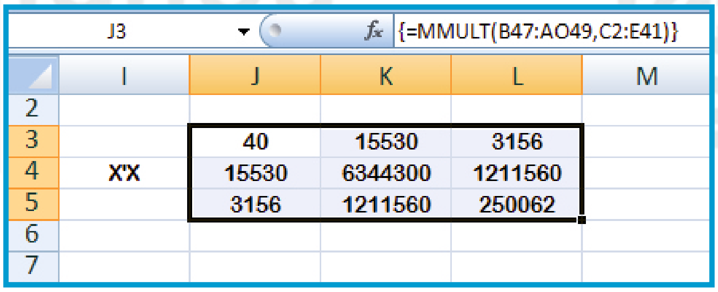


Fig. 12.27

Step 11: For finding the inverse of $\mathbf{X}'\mathbf{X}$, we use ***Minverse*** function as shown in Fig. 12.28 and add Cells J3:L5 in the dialog box that opens (see Fig. 12.29).

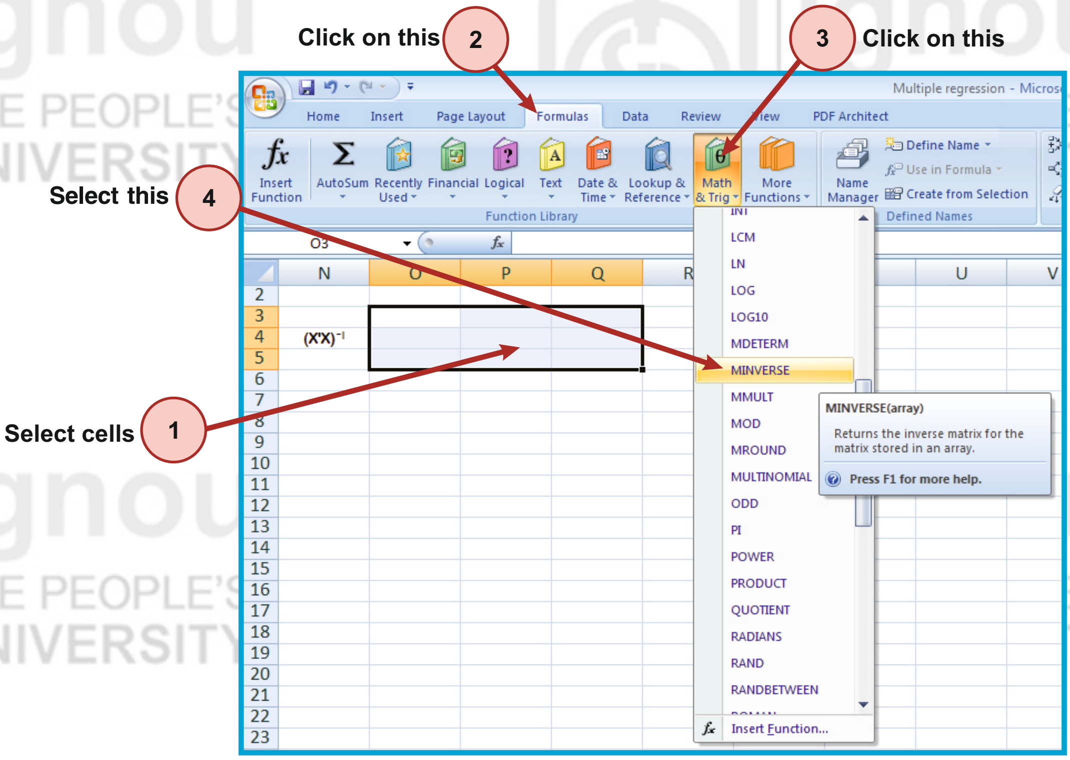


Fig. 12.28

Step 12: As discussed in previous steps, after pressing ***Enter*** along with (***Ctrl*** + ***Shift***) keys, we get the inverse matrix of (**X'X**), i.e., $(\mathbf{X'X})^{-1}$ as shown in Fig. 12.29.

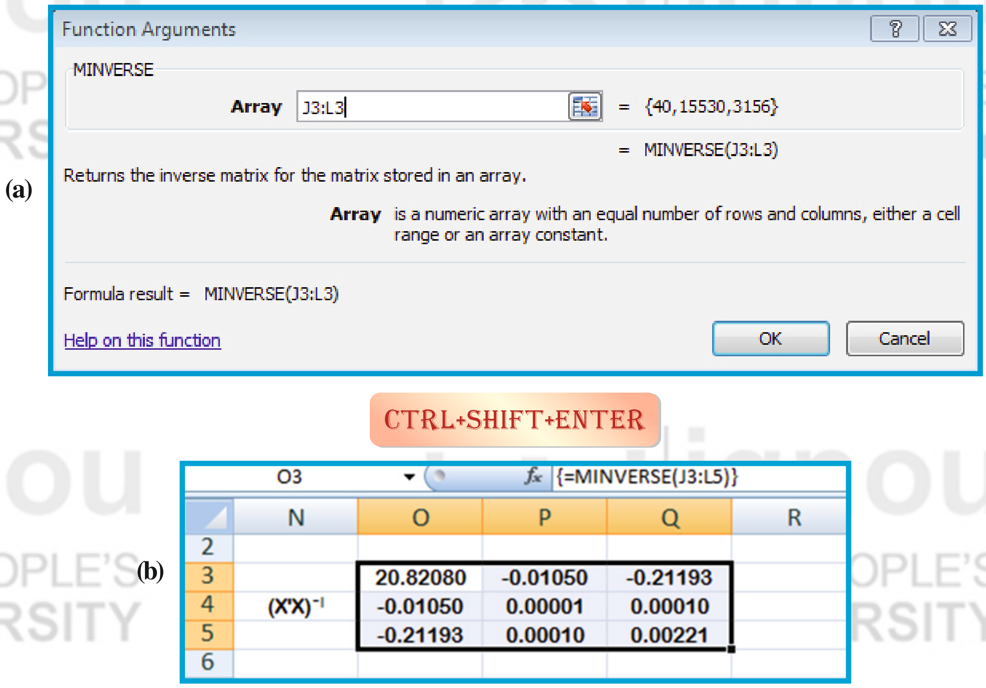


Fig. 12.29

Step 13: We recall equation (8) to determine the **B** matrix. We find the matrix after multiplying $(\mathbf{X'X})^{-1}$ and $\mathbf{X'Y}$ matrices by taking the values of $(\mathbf{X'X})^{-1}$, i.e., Cells O3:Q5 in ***Array 1*** and values of $\mathbf{X'Y}$ i.e., Cells G3:G5 in ***Array 2***. Then we press (***Ctrl*** + ***Shift*** + ***Enter***) together. It gives us the required values of regression coefficients shown in Fig. 12.30.

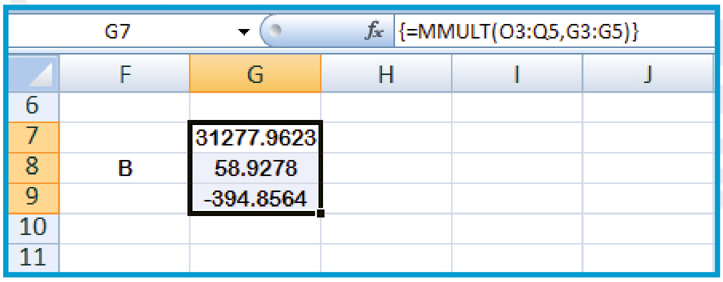


Fig. 12.30

From Fig. 12.30, you can see that $B_0 = 31277.9623$, $B_1 = 58.9278$ and $B_2 = -394.8564$. These are the same as the values given in Fig. 12.13.

Step 14: To find B' (transpose of B), we can use ***Transpose*** function from ***Lookup & Reference*** under ***Formulas*** tab as shown in Fig. 12.31 in the same manner as we used ***Mmult*** or ***Minverse*** functions.

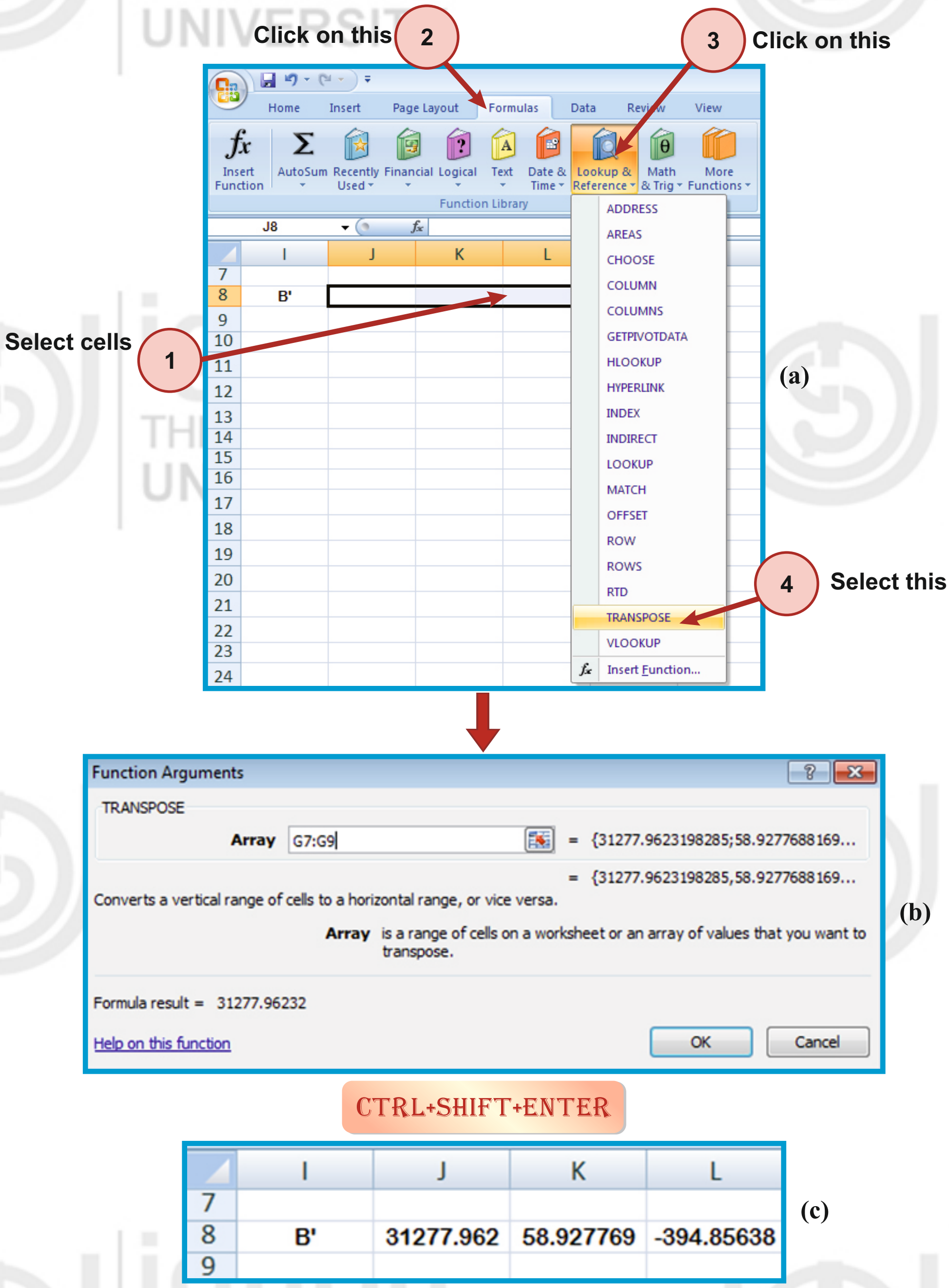


Fig. 12.31

Step 15: Our next step is to do the regression analysis through matrix approach for which we prepare Cells F10:G24 shown in Fig. 12.32 in the following manner:

- ✓ Type the values of n, (n–1), p and (n–p–1) in Cells G10, G11, G12 and G13, respectively.
- ✓ Use the formula "=Average(B2:B41)" in Cell G14 to determine $\bar{Y}$.
- ✓ To compute $\mathbf{Y'Y}$, we apply ***Mmult*** function in Cell G15 considering Cells B46:AO46 and B2:B41 as ***Array 1*** and ***2***, respectively.
- ✓ Similarly, we use Cells J8:L8 and G3:G5 as ***Array 1*** and ***2***, respectively, to find out $\mathbf{B'(X'Y)}$ in Cell G16.
- ✓ To find the total sum of squares (SS_T), we type "=G15-G11*G14*G14" in Cell G17.
- ✓ To compute regression sum of squares (SS_{Reg}), we type "=G16-G11*G14*G14" in Cell G18.
- ✓ To calculate the residual sum of squares (SS_{Res}), we type "=G17-G18" in Cell G19.
- ✓ To find the regression mean sum of squares (MSS_{Res}), we type "=G18/G12" in Cell G20.
- ✓ To determine the residual mean sum of squares (MSS_{Res}), we type "=G19/G13" in Cell G21.
- ✓ To find the F calculated value (F_{Cal}) for testing the overall significance of regression model, we type "=G20/G21" in Cell G22.
- ✓ To compute the p-value for the F test, we type "=Fdist(G22,G12,G13)" in Cell G23.

	F	G
10	n	40
11	n-1	39
12	p	2
13	n-p-1	37
14	$\bar{Y}$	23002.5000
15	Y'Y	23192710000.0000
16	B'(X'Y)	23062123761.9535
17	SS_T	2028109750.0000
18	SS_{Reg}	1897523511.9535
19	SS_{Res}	130586238.0465
20	MSS_{Reg}	948761755.9767
21	MSS_{Res}	3529357.7850
22	F_{Cal}	268.8199
23	p-value	0.0000
24	R^2	0.9356
25	R^2_{adj}	0.9321

Fig. 12.32

- ✓ To find the coefficient of determination (R^2), we type "=1-(G19/G17)" in Cell G24.
- ✓ To calculate the adjusted coefficient of determination (R^2_{adj}), we type "=1-((G19/G13)/(G17/G11))" in Cell G25.

The results are shown in Fig. 12.32.

Step 16: We know that the estimated variance of error is $\hat{\sigma}^2 = MSS_{Res}$ as given in Cell G21. The variance-covariance matrix of the regression coefficients using equations (10) and (11) can be calculated in Cells O7:Q9 by multiplying $\hat{\sigma}^2$ given in Cell G21 with the respective values of $(\mathbf{X'X})^{-1}$ given in Cells O3:Q5. These are shown in Fig. 12.33.

	N	O	P	Q
6				
7		73484050.0644	-37042.0218	-747962.6258
8	$\hat{\sigma}^2(X'X)^{-1}$	-37042.0218	26.1143	340.9778
9		-747962.6258	340.9778	7802.0024

Fig. 12.33

The standard error of the regression coefficients using equation (12) and the value of t-statistic (t_j) using equation (21) for j = 0, 1 and 2 can be computed as shown in Figs. 12.34a and b, respectively.

We can also determine the corresponding p-values for j = 0, 1 and 2 using ***Tdist*** function as shown in Fig. 12.34c.

	N	O	
11	$SE(\hat{\beta}_0) = \sqrt{\sigma_{00}}$	8572.2838	→ = Sqrt(O7)
12	$SE(\hat{\beta}_1) = \sqrt{\sigma_{11}}$	5.1102	→ = Sqrt(P8)
13	$SE(\hat{\beta}_2) = \sqrt{\sigma_{22}}$	88.3289	→ = Sqrt(Q9)

(a)

	N	O	
15	t_0	3.6487	→ = G7/O11
16	t_1	11.5314	→ = G8/O12
17	t_2	-4.4703	→ = G9/O13

(b)

	N	O	
15	t_0	3.6487	→ = Tdist(O15,37,2)
16	t_1	11.5314	→ = Tdist(O16,37,2)
17	t_2	-4.4703	→ = Tdist(Abs(O17),37,2)

(c)

Fig. 12.34

The results for fitting of the multiple regression, testing of overall regression models and the individuals regression coefficients obtained through matrix approach given in Figs. 12.30, 12.32 and 12.34 are the same as the output given in Fig. 12.10 obtained from ***Data Analysis ToolPak***. The interpretation of these results will be the same as discussed in Sec. 12.6.

Now, it is time for you to solve some problems.

Activity 2

You may record the total electricity consumption in kWh in your house for 20 days. You may also record the numbers of hours your television (TV) or air conditioner (AC) is turned on for each day during this period. You can also note the temperature each day. In this way, you will collect the data of electricity consumption, total TV or AC hours per day and temperature every day for 20 days.

Perform the multiple regression analysis using Excel 2007 for this data and interpret the results.

Activity 3

Work out the following exercises with the help of MS Excel 2007 and interpret the results:

A1) Examples 1, 2, 3, 4 and 6 given in Unit 11 of MSTE-002.

A2) Exercises E1, E2 and E4 given in Unit 11 of MSTE-002.

Match the results with the manual calculations done in Unit 11 of MSTE-002.

Continuous Assessment 12

Suppose we are interested in developing a linear model for the electricity consumption of a household having one AC so that we can predict the electricity consumption during summers. For this purpose, the sample of 40 houses was selected. We have recorded the electricity consumption (in kWh), size of house (in square feet) and AC hours for one month during summers in Table 2.

Table 2: Electricity consumption data

S.No.	Unit (in kWh)	Area (in sq ft)	Ac (in hour)	S.No.	Unit (in kWh)	Area (in sq ft)	Ac (in hour)
1	1060	1316	5	21	1565	1696	15
2	1150	1420	7	22	1215	1464	9
3	1365	1556	12	23	1275	1488	10
4	1275	1488	9	24	1465	1632	13
5	1425	1612	13	25	1080	1356	7
6	1310	1516	10	26	975	1196	4
7	1365	1556	12	27	1040	1256	5
8	1075	1352	6	28	1340	1540	11
9	925	1168	4	29	865	1144	4
10	1340	1540	11	30	1175	1440	8
11	1425	1612	13	31	1080	1356	7

S.No.	Unit (in kWh)	Area (in sq ft)	Ac (in hour)	S.No.	Unit (in kWh)	Area (in sq ft)	Ac (in hour)
12	1150	1420	8	32	1500	1652	15
13	1060	1316	5	33	1175	1440	9
14	1545	1680	15	34	1050	1296	5
15	1140	1388	7	35	1365	1580	12
16	1075	1352	6	36	1465	1632	15
17	1620	1736	16	37	1215	1464	9
18	1050	1296	5	38	1365	1580	12
19	1310	1516	10	39	1140	1388	7
20	1645	1760	16	40	1005	1224	4

- Prepare a scatter matrix to get a rough idea about the relationship among the variables. Develop a multiple linear regression model for this data using the matrix approach.
- Test the significance at 1% level of significance and find the 99% confidence interval of the regression parameters using the matrix approach.
- Also check the linearity and normality assumptions for the regression analysis.
- Compare the results obtained from the matrix approach and Data Analysis ToolPak.

Home Work: Do It Yourself

1) Follow the steps explained in Secs. 12.4 to 12.9 to comprehend the multiple regression analysis for the data of Table 1. Use a different format for the scatter plot matrix, residual and normal probability plots. Take their screenshots and keep them in your record book.
2) Develop the spreadsheets for the exercise "Continuous Assessment 12" as explained in this lab session. Take screenshots of the final spreadsheets and the plots.
3) **Do not forget** to keep the screenshots in your record book as these will contribute to your continuous assessment in the Laboratory.