PART C

Regression Modelling

REGRESSION MODELLING

The third part of this lab course comprises Sessions 11 to 14, which will enable you to develop the practical skills for performing regression analysis using MS Excel 2007. In lab sessions of Part A, you have learnt how to construct the control charts for variables, which are used to control the process when the quality characteristic is **measured** numerically. You have also learnt how to construct the control charts for attributes in Part B. As you know, these charts are used to control the process when the number of defectives or defects are **counted**.

In Block 3 of MSTE-002 (Industrial Statistics-II), you have learnt that regression analysis is one of the most important and widely used tool for analysing multivariate data in statistical data analysis. Using regression analysis, you can fit a mathematical relationship between the variable under consideration and other associated variables. You can also use this relationship to predict the values of the variable under consideration called response variable on the basis of associated variables called predictors or regressors. In the lab sessions of Part C, you will learn the technique of regression analysis through illustrative examples using MS Excel 2007.

In Session 11, you will learn how to **fit a simple linear regression model** to **predict** a response variable based on the predictor. We also elaborate on the **inferential analysis** of the regression coefficient, **normal probability plot**, **residual analysis**, etc., in this session.

It is always beneficial to use more than one predictor, which contributes to the model to build a regression model. Session 12 deals with the situation where you have **more than one predictor**.

In many studies, you may also come across situations when one or more predictor variables are **qualitative**. For example, sex (male/female), promotional activity (Yes/No), etc. Such a situation is handled in Session 13 through an example. In Session 14, we explain three variable selection methods, namely, **forward selection**, **backward elimination** and **stepwise methods,** which enable us to choose the most appropriate variables for the regression model.

All sessions of this part have been designed in such a way that after successfully completing them, you will be familiar with a mathematical relationship between the response variable and one or more predictor variables and its related analysis using MS Excel 2007. You will also be able to apply the regression technique for any given data.

You should study Block 3 of MSTE-002 thoroughly before performing the activities of the lab sessions of Part C. In this part, we shall apply all concepts and formulae explained in Block 3 of MSTE-002.

Simple Linear Regression

SESSION 11

Structure

11.1 Introduction
Objectives
11.2 Problem Description
11.3 Procedure for Simple Regression Analysis
11.4 Fitting and Analysis of Simple Regression in Excel 2007
11.5 Interpretation of the Results of Regression Analysis
Interpretation of the Regression Statistics Table
Interpretation of the ANOVA Table
Interpretation of the Regression Coefficients Table
Interpretation of the Residual Output
11.6 Residual Plot in Excel 2007
11.7 Normal Probability Plot in Excel 2007
11.8 Fitted Line of Regression on the Scatter Plot in Excel 2007

11.1 INTRODUCTION

In regression analysis, one variable is referred to as dependent variable or response variable, whereas the other variables are referred to as independent variables, predictor variables or regressor variables. Regression analysis with only one predictor variable is known as **simple regression analysis** and with two or more predictor variables is known as **multiple regression analysis**. The purpose of regression analysis is to establish an average relationship between a response variable and one or more predictor variables to predict the response variable using this relationship.

Prerequisite

- Lab Sessions 1, 3 and 6 of MSTL-001 (Basic Statistics Lab).
- Units 9, 10 and 12 of MSTE-002 (Industrial Statistics-II).

In this lab session, we shall explain simple regression analysis. You have already learnt about the manual calculation for least square estimation of the parameters, fitting of the regression line, residual analysis, test of significance of regression coefficients, confidence interval and coefficient of determination in Units 9 and 10 of MSTE-002 (Industrial Statistics-II). In this lab session, you will use MS Excel 2007 to obtain the solution of the problems using simple regression analysis. You will learn multiple regression analysis in the next lab session.

Simple linear regression analysis can easily be completed using the *Data Analysis* option under the *Data* tab in Excel 2007.

Objectives

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

- prepare the spreadsheet for regression analysis in MS Excel 2007;
- estimate the model parameters using the principle of least squares;
- fit the simple regression line;
- test the significance of regression parameters;
- determine the confidence interval of the regression parameters;
- construct the residual and normal probability plots; and
- interpret the results of simple regression analysis.

11.2 PROBLEM DESCRIPTION

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

Table 1: Monthly sales and advertisement cost of a company

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

Using this data:

- Draw a scatter diagram. Does this data fit a straight line?
- Perform the regression analysis through Excel 2007 and interpret the results.
- Draw the fitted regression line on the scatter plot. Does a regression line appear to give a good fit here?

- Predict the sales for a particular month for which the advertisement cost is ₹ 49000.
- Obtain the residuals and verify the property $\sum_{i=1}^{n} e_i = 0$.
- Draw the residual curve and normal probability plot.

11.3 PROCEDURE FOR SIMPLE REGRESSION ANALYSIS

You have already learnt the procedure of simple regression analysis in Units 9 and 10 of MSTE-002. So here we briefly list the relevant formulae as follows:

Step 1: Let Y and X be the dependent (response) and independent (predictor) variables, respectively. The equation of regression line (Y on X) is given as

$$Y = a + bX + e \qquad \ldots(1)$$

where a is the intercept, b, the slope and e, a normally distributed random error component with mean zero and variance σ^2.

Step 2: If $\hat{a}$ and $\hat{b}$ are the least squares estimators of a and b, respectively, we can estimate the regression parameters using the method of least squares as follows:

$$\hat{a} = \overline{Y} - \hat{b}\overline{X} \quad \text{and} \qquad \ldots (2)$$

$$\hat{b} = \frac{S_{xy}}{S_{xx}} \qquad \ldots (3)$$

where $\overline{Y} = \dfrac{\sum_{i=1}^{n} Y_i}{n}$

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$S_{xx} = \sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n} = \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$$

$$S_{xy} = \sum_{i=1}^{n} Y_i X_i - \frac{\left(\sum_{i=1}^{n} Y_i\right)\left(\sum_{i=1}^{n} X_i\right)}{n} = \sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$$

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

$$\hat{Y} = \hat{a} + \hat{b}X \qquad \ldots(4)$$

Step 4: The difference between the observed value Y_i and the corresponding fitted value $\hat{Y}_i$ is called the ith residual and is given by

$$r_i = Y_i - \hat{Y}_i, \qquad i = 1, 2, \ldots, n \qquad \ldots(5)$$

Step 5: The ith standardised residual is given by

$$d_i = \frac{r_i}{\sqrt{SS_{Res}}}, \qquad i = 1, 2, \ldots, n \qquad \ldots(6)$$

Step 6: Using ANOVA (analysis of variance) table, the total sum of squares (SS_T) is split into two components: (i) Regression or explained sum of squares (SS_{Reg}) and (ii) Residual or error sum of squares (SS_{Res}), i.e.,

$$SS_T = SS_{Reg} + SS_{Res} \qquad \ldots(7)$$

Step 7: The ratio of variability in Y accounted for by X to the total variability in Y around mean is called the **coefficient of determination** and denoted by $\mathbf{R^2}$. It can be computed as

$$R^2 = \frac{SS_{Reg}}{SS_T} = 1 - \frac{SS_{Res}}{SS_T} \qquad \ldots(8)$$

R^2 represents the proportion of variation in response variable Y explained by the predictor X. It gives the overall goodness-of-fit measure. The value of R^2 lies between 0 and 1.

Step 8: The estimate of σ^2 is called the residual mean squares (MSS_{Res}) and can be computed as follows:

$$\hat{\sigma}^2 = MSS_{Res} = \frac{SS_{Res}}{n-2} \qquad \ldots(9)$$

Step 9: $\hat{a}$ and $\hat{b}$ are the unbiased estimators of the regression parameters a and b, respectively, and their variances are given by

$$\hat{V}(\hat{a}) = \hat{\sigma}^2\left(\frac{1}{n} + \frac{\overline{X}^2}{S_{xx}}\right) \quad \text{and} \quad \hat{V}(\hat{b}) = \frac{\hat{\sigma}^2}{S_{xx}} \qquad \ldots(10)$$

Step 10: For testing H_0: $a = a_0$ against H_1: $a \neq a_0$, the t-statistic can be calculated as follows if the null hypothesis is correct:

$$t = \frac{(\hat{a} - a_0)}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\overline{X}^2}{S_{xx}}\right)}} \qquad \ldots(11)$$

This is distributed as a Student's 't'-statistic with $(n-2)$ degrees of freedom (d.f.). We may reject H_0, if $|t| \geq t_{(n-2),\alpha/2}$ at $\alpha\%$ level of significance.

Step 11: For testing H_0: $b = b_0$ against H_1: $b \neq b_0$, the t-statistic can be calculated as

$$t = \frac{(\hat{b} - b_0)}{\sqrt{\hat{\sigma}^2 / S_{xx}}} \qquad \ldots(12)$$

which follows a Student's t-distribution with $(n-2)$ d.f. under H_0. We may reject H_0, if $|t| \geq t_{(n-2),\alpha/2}$ at $\alpha\%$ level of significance.

Step 12: $(1-\alpha)100\%$ lower and upper confidence limits for intercept (a) are given as

$$a_L = \hat{a} - t_{\alpha/2}\sqrt{\hat{V}(\hat{a})} \text{ and } a_U = \hat{a} + t_{\alpha/2}\sqrt{\hat{V}(\hat{a})} \qquad \ldots(13)$$

$(1-\alpha)100\%$ lower and upper confidence limits for slope (b) are given as

$$b_L = \hat{b} - t_{\alpha/2}\sqrt{\hat{V}(\hat{b})} \text{ and } b_U = \hat{b} + t_{\alpha/2}\sqrt{\hat{V}(\hat{b})} \qquad \ldots(14)$$

Step 14: The percentiles for the normal probability plot can be computed as follows:

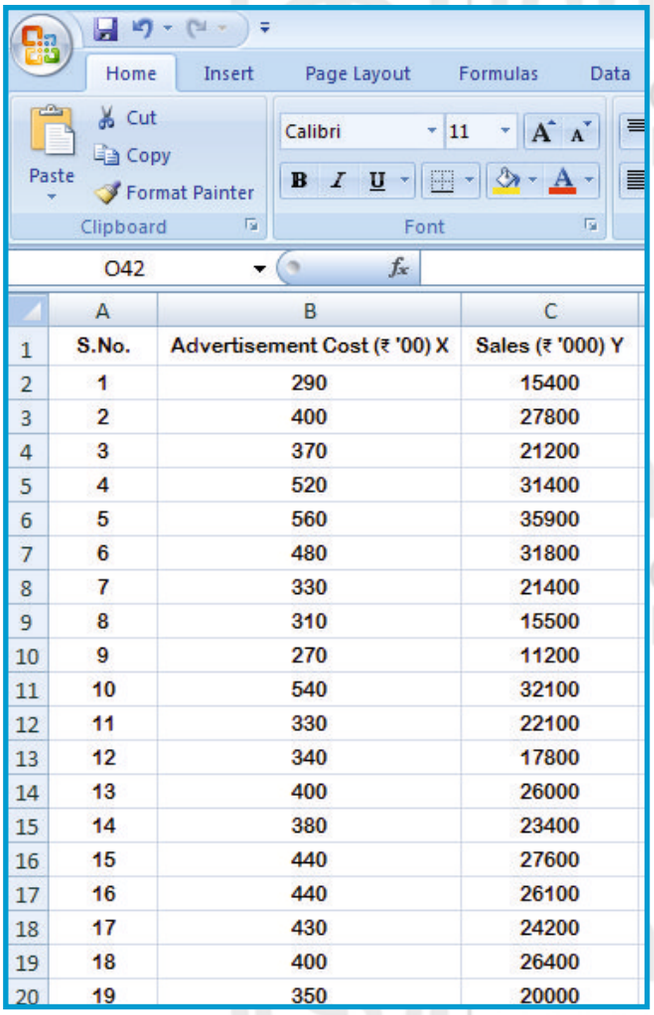
$$P_i = \frac{(i - 1/2)}{n} \times 100, \; i = 1, 2, \ldots, n \qquad \ldots(15)$$

11.4 FITTING AND ANALYSIS OF SIMPLE REGRESSION IN EXCEL 2007

The problem we are considering here will help you understand the method and concepts you have learnt in Units 9 and 10 of MSTE-002. In order to apply simple linear regression analysis to the given data, we follow the steps listed below:

Step 1: We enter the given 40 observations on sales and advertisement cost in an Excel spreadsheet. In the given data, the dependent variable (Y) is the monthly sales, whereas the independent variable (X) is advertisement cost since the values of sales depend upon the values of advertisement cost. A portion of the data as it appears in an Excel spreadsheet is shown in Fig.11.1.

You may also use different **Statistical Functions** for manual calculation provided in Excel to get the same results as ***Data Analysis ToolPak*** of Excel to perform the regression analysis like ***Intercept***, ***Slope***, ***Rsq***, ***Tdist***, ***Tinv***, etc.



	A	B	C
1	S.No.	Advertisement Cost (₹ '00) X	Sales (₹ '000) Y
2	1	290	15400
3	2	400	27800
4	3	370	21200
5	4	520	31400
6	5	560	35900
7	6	480	31800
8	7	330	21400
9	8	310	15500
10	9	270	11200
11	10	540	32100
12	11	330	22100
13	12	340	17800
14	13	400	26000
15	14	380	23400
16	15	440	27600
17	16	440	26100
18	17	430	24200
19	18	400	26400
20	19	350	20000

Fig. 11.1: Partial screenshot of the spreadsheet for the given data.

Step 2: For scatter plot, we place the response variable (monthly sales) on the vertical (or Y) axis and the predictor (advertisement cost) on the horizontal (or X) axis, which displays the relationship between two variables. To plot sales versus advertisement cost in Excel, we select the data (Cells B2:C41) and choose ***Scatter*** under the ***Insert*** tab as shown in Fig. 11.2.

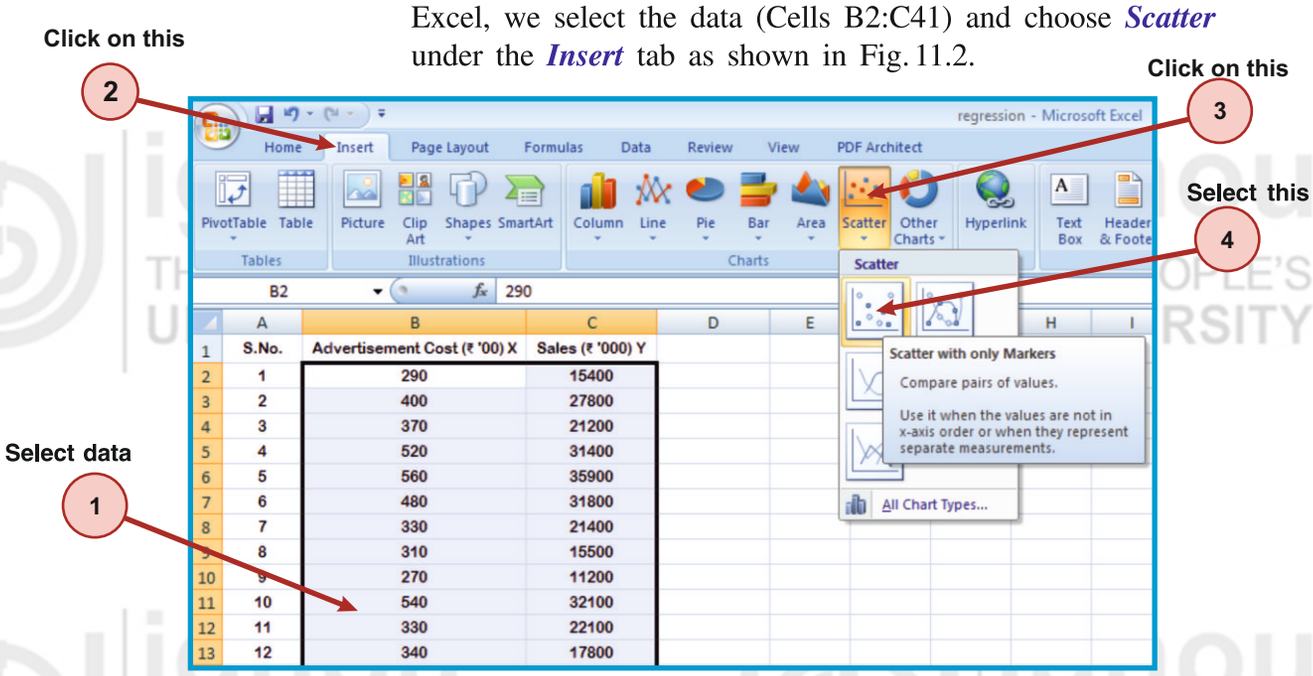


Fig. 11.2

Step 3: We format the chart as explained in Sec. 3.3 of Lab Session 3 of MSTL-001 (Basic Statistics Lab). Thus, we obtain the scatter plot shown in Fig. 11.3.

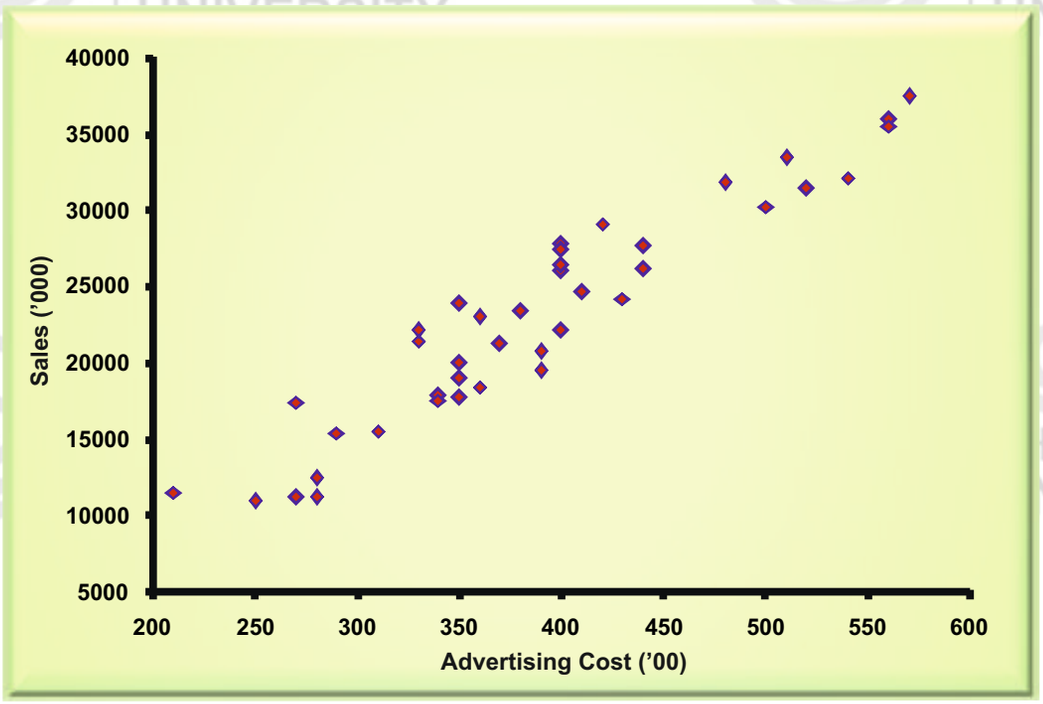


Fig. 11.3: Scatter diagram.

Notice from Fig. 11.3 that the relationship between sales and advertisement cost can be approximately described by a straight line. There may be many straight lines that could be drawn through these points. So the question is: How do we choose the best regression line? We fit the regression line using the method of least squares. The equation of regression line is given by equation (1).

Here we use ***Data Analysis ToolPak*** for regression analysis in Excel. You may also do the regression analysis in Excel manually as you have already learnt in Units 9 and 10 of MSTE-002.

The ***Data Analysis ToolPak*** includes a number of advanced data analysis tools including Regression. The Regression procedure provides regression statistic(s), ANOVA, regression coefficients, their standard errors, t statistic(s), p-values and upper and lower confidence limits and residual output.

Step 4: In Excel 2007, we can fit a regression line on the given data by choosing ***Data Analysis*** under the ***Data*** tab and subsequently selecting ***Regression***. We come across the dialog boxes shown in Fig. 11.4.

For regression analysis in Excel 2007, we

1. click on the ***Data*** tab,
2. select ***Data Analysis*** as shown in Fig. 11.4a,
3. select the ***Regression*** option,
4. click on ***OK*** as shown in Fig. 11.4b, and
5. get a new dialog box as shown in Fig. 11.4c.

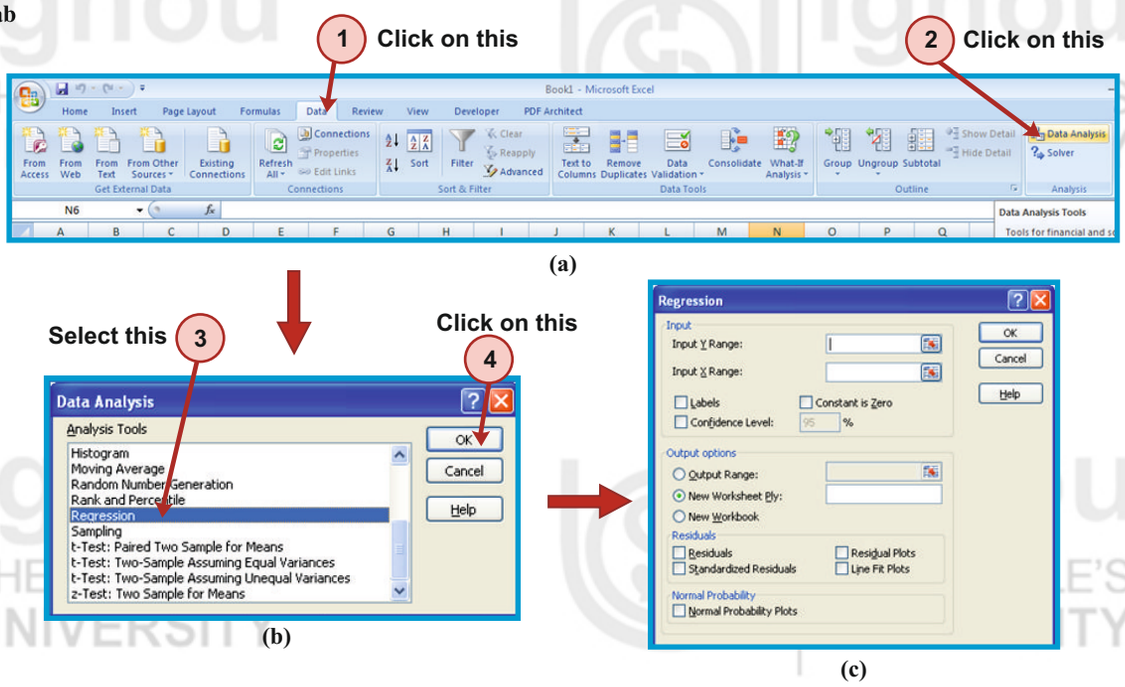


Fig. 11.4

Step 5: Refer to Fig. 11.5. We

1. specify the data with label for monthly sales in ***Input Y Range***, i.e., Cells C1:C41,
2. specify the data with label for the monthly advertisement cost in ***Input X Range***, i.e., Cells B1:B41,
3. check the ***Labels*** box since we have included data labels, i.e., advertisement cost given in Cell B1 and sales given in Cell C1 in the input ranges,
4. specify the 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%,
5. provide a new worksheet name under ***Output Options***. Here we use the name "**Regression Analysis**" for the output sheet, and
6. check the ***Residuals***, ***Standardised Residuals*** and ***Normal Probability Plot*** boxes and then click on ***OK***.

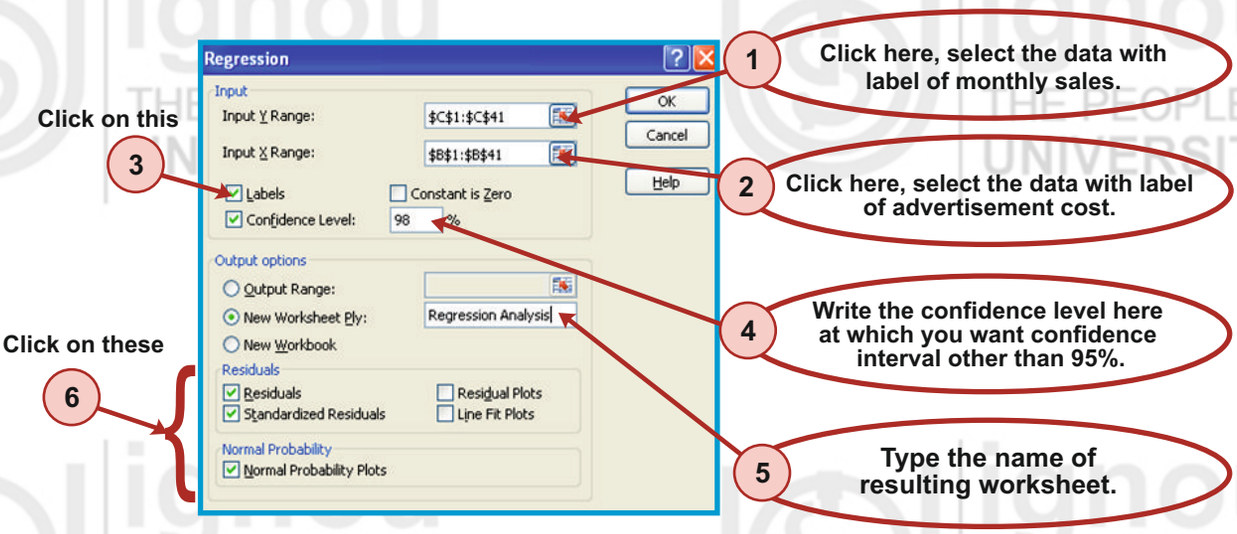
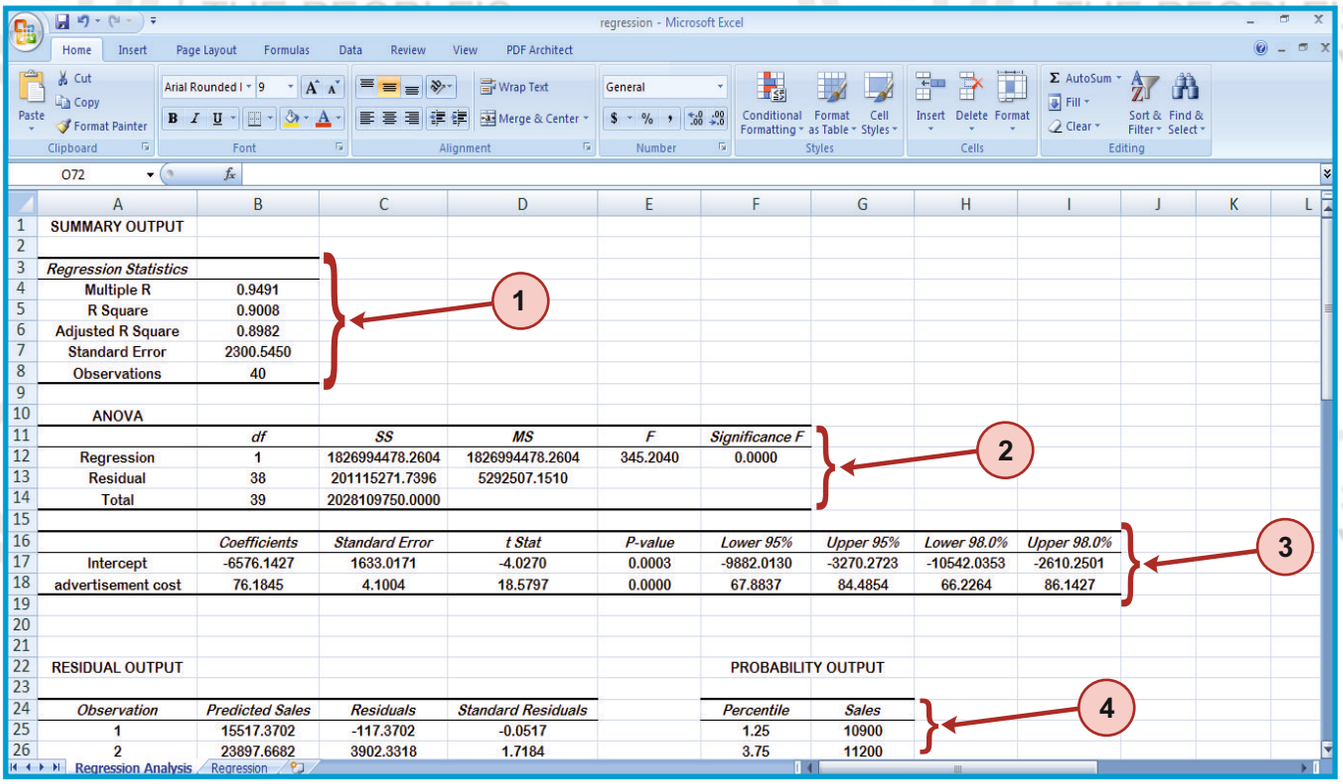


Fig. 11.5

Step 6: After completing Step 5, we obtain the resulting worksheet (subset shown in Fig. 11.6).



	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.9491							
5	R Square	0.9008							
6	Adjusted R Square	0.8982							
7	Standard Error	2300.5450							
8	Observations	40							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	1	1826994478.2604	1826994478.2604	345.2040	0.0000			
13	Residual	38	201115271.7396	5292507.1510					
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	-6576.1427	1633.0171	-4.0270	0.0003	-9882.0130	-3270.2723	-10542.0353	-2610.2501
18	advertisement cost	76.1845	4.1004	18.5797	0.0000	67.8837	84.4854	66.2264	86.1427
19									
20									
21									
22	RESIDUAL OUTPUT					PROBABILITY OUTPUT			
23									
24	*Observation*	*Predicted Sales*	*Residuals*	*Standard Residuals*		*Percentile*	*Sales*		
25	1	15517.3702	-117.3702	-0.0517		1.25	10900		
26	2	23897.6682	3902.3318	1.7184		3.75	11200		

Fig. 11.6: Partial screenshot of the output sheet "Regression Analysis".

So far you have learnt how to construct the scatter diagram and obtain the results of regression analysis in Excel 2007. It is also very important to learn how to interpret the results obtained. This is discussed in Sec. 11.5.

11.5 INTERPRETATION OF THE RESULTS OF REGRESSION ANALYSIS

Note that the regression output given in Fig. 11.6 has four components:

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

We now discuss how to interpret these components.

11.5.1 Interpretation of the Regression Statistics Table

The first part of the output (Fig. 11.7) shows Regression Statistic(s) in which R-Square is of the greatest interest.

	A	B
1	SUMMARY OUTPUT	
2		
3	*Regression Statistics*	
4	Multiple R	0.9491
5	R Square	0.9008
6	Adjusted R Square	0.8982
7	Standard Error	2300.5450
8	Observations	40
9		

→ It is used when the model has more than one predictors.
→ Sample estimator of the standard deviation of the error.
→ Number of observations used in the regression analysis (n).

Fig. 11.7

We can also calculate R^2 from ANOVA table as follows:

The general formula for R^2 given in equation (8) can be written as

$$R^2 = 1 - \frac{\text{Residual SS}}{\text{TotalSS}}$$

$$= 1 - \frac{201115271.7396}{2028109750}$$

(from ANOVA table given in Fig 11.8).

$\therefore\ R^2 = 0.9008$

which equals R^2 given in Fig. 11.7.

- Multiple R represents the correlation coefficient between dependent (sales) and independent (advertisement cost) variables. Its value is 0.9491 given in Cell B4 for this problem. We can say that sales and advertisement cost are highly positively correlated with each other.

- $R^2 = 0.9008$ given in Cell B5 means that 90.08% of the variation in Y is explained by the predictor X, i.e., 90.08% of the variability in monthly sales is accounted for by the regression model.

- The value of standard error given in Cell B7 here refers to the estimated standard deviation of the error term e, i.e., $\hat{\sigma}$. It is sometimes called the standard error of the regression. From Fig. 11.7, $\hat{\sigma} = 2300.5450$.

11.5.2 Interpretation of the ANOVA Table

The ANOVA (analysis of variance) table given in Fig. 11.8 splits the total sum of squares into two components:

i) Regression (or explained) sum of squares, and

ii) Residual (or error) sum of squares.

Excel computes the value of F-statistic given in Cell E12 as:

$$F = \frac{\text{Regression MSS}}{\text{Residual MSS}}$$

$$= \frac{1826994478.2604}{201115271.7396}$$

$= 345.2040$ (Fig. 11.8.)

Note that the square root of the residual mean square given in Cell D13 shown in Fig. 11.8 is the standard error of estimate, i.e., $\hat{\sigma} = \sqrt{5292507.1510} = 2300.5450$, which is equivalent to the standard error shown in Cell B7 of Fig. 11.7.

	A	B	C	D	E	F
10	ANOVA					
11		*df*	*SS*	*MS*	*F*	*Significance F*
12	Regression	1	1826994478.2604	1826994478.2604	345.2040	0.0000
13	Residual	38	201115271.7396	5292507.1510		
14	Total	39	2028109750.0000			

Fig. 11.8

- **Testing the significance of the regression model**

To compute the p-value for F test in Excel, we type " =Fdist (345.2040, 1, 38)" in any cell and press ***Enter***.

The p-value will be 0.000, which is the same as the value given in Cell F12.

The output given in Fig. 11.8 shows the ANOVA table for the monthly sales data. If we consider $H_0: b = 0$ vs $H_1: b \neq 0$, this hypothesis relates to the significance of the regression model.

If $H_0: b = 0$ is not rejected, it implies that there is no linear relationship between X and Y, i.e., X is not contributing in the prediction of Y. The residual mean square (Cell D13) is a measure of how poorly or how well the regression line fits the actual data points. A large residual mean square indicates a poor fit.

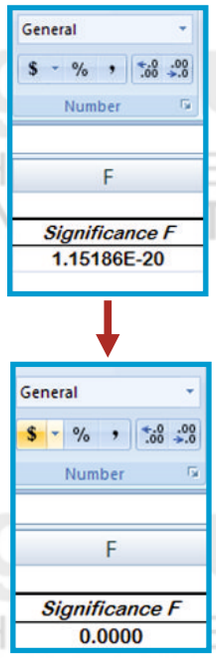
We use an analysis of variance approach to test the significance of the regression model. The analysis of variance is based on the partitioning of the total variability in the response variable Y. Here we use the usual analysis of variance (F-test) to test the null hypothesis $H_0: b = 0$.

F-statistic is given as

$$F = \frac{MSS_{Reg}}{MSS_{Res}}$$

From the ANOVA table given in Fig. 11.8, the F-test statistic (Cell E12) is 345.2040. The Column F labelled significance in Fig. 11.8 has the associated p-value 0.0000 (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. Thus, we may conclude that there is a significant relationship between sales and advertisement cost.

Sometimes the p-value comes in Excel (Cell F12) in the format of E as shown below. So we need to change the format of that cell to a number as you have learnt in Lab Session 1 of MSTL-001.



11.5.3 Interpretation of the Regression Coefficients Table

The regression output given in Fig. 11.9 depicts the regression parameters and the associated output. In the regression coefficients table:

- ✓ Cells A17 and A18 named **Intercept** and **advertisement cost** represent the labels for the intercept (a) and slope (b), respectively.
- ✓ The column **Coefficients** gives the least squares estimates of a and b, i.e., $\hat{a}$ and $\hat{b}$ in Cells B17 and B18, respectively.
- ✓ The column **Standard error** gives the standard errors of the least squares estimates of a and b in Cells C17 and C18, respectively.
- ✓ The column **t Stat** gives the computed t-statistic for H_0: a = 0 against H_1: a ≠ 0 and H_0: b = 0 against H_1: b ≠ 0 in Cells D17 and D18, respectively.

  These values are the coefficients divided by the respective standard errors [equations (11) and (12)].
- ✓ The column **P-value** gives the p-value for testing H_0: a = 0 against H_1: a ≠ 0 and H_0: b = 0 against H_1: b ≠ 0 in Cells E17 and E18, respectively.
- ✓ The columns **Lower 95%**, **Upper 95%**, **Lower 98%** and **Upper 98%** define a 95% and 98% confidence interval, respectively, for the values of a and b in Cells F17:F18, G17:G18, H17:H18 and I17:I18, 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	-6576.1427	1633.0171	-4.0270	0.0003	-9882.0130	-3270.2723	-10542.0353	-2610.2501
18	advertisement cost	76.1845	4.1004	18.5797	0.0000	67.8837	84.4854	66.2264	86.1427
19									
20									
21									

Fig. 11.9

We now interpret the regression coefficients table given in Fig. 11.9 as follows:

➢ **Fitting of the regression model**

From Fig. 11.9, you can see that $\hat{a}$ = – 6576.1427 and $\hat{b}$ = 76.1845.

Hence, the fitted regression equation (Fig. 11.10) is

$$\hat{Y} = -6576.1427 + 76.1845X$$

or, Monthly Sales = – 6576.1427 + 76.1845 × Advertisement Cost

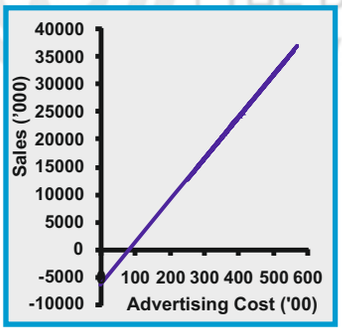


Fig. 11.10

If Y and X variables are multiplied by p and q, respectively, as

$\hat{Y}_c = p \times \hat{Y}$ and

$\hat{X}_c = q \times \hat{Y}$

The corrected fitted regression line after changing the scale will be

$$\hat{Y}_c = p \times \hat{a} + \left[\frac{p \times \hat{b}}{q}\right](q \times \hat{X})$$

$$\hat{Y}_c = p \times (\hat{a} + \hat{b}\hat{X})$$

We can say that the sales can be predicted by multiplying the advertisement cost by 76.1845 and subtracting 6576.1427 for a particular month. We can also say that the sales of two months are expected to differ by the product of 76.1845 with the difference between the respective advertisement costs.

➢ **Predicted value of Y for given X**

We compute the predicted monthly sales for ₹ 490 (in'00) advertisement cost by typing "= – 6576.1427 + 76.1845*490" in Cell B44 as shown in Fig. 11.11 (you may also choose any cell to calculate this value).

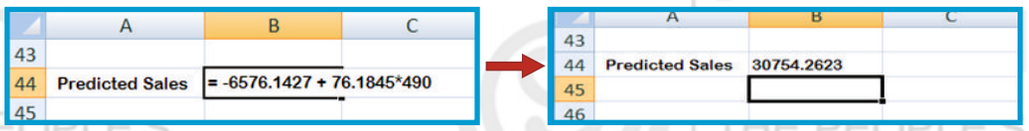


Fig. 11.11

If advertisement cost for a particular month is ₹ 490 (in'00), the predicted value of sales is given by $\hat{Y}$ = ₹ 30754.2623 (in'000), i.e., ₹ 30754262.30.

➢ **Standard error of the regression coefficients**

The standard errors of $\hat{a}$ and $\hat{b}$ given in Cells C17 and C18, respectively, are

$$SE(\hat{a}) = 1633.0171 \text{ and } SE(\hat{b}) = 4.1004$$

➢ **Hypothesis testing for the slope**

The ***Regression*** tool in ***Data Analysis ToolPak*** tests the hypothesis that the intercept is equal to zero, i.e.,

$$H_0: a = 0 \text{ vs } H_1: a \neq 0$$

To compute the p-value for t-test in Excel, we type "= Tdist(4.027, 38, 2)" in any cell. It will give 0.0003, which is the same as the value given in Cell E17.

We use the p-value approach for decision making. From Fig. 11.9, the t-statistic for the intercept is – 4.0270 (Cell D17) and its p-value is 0.0003 (Cell E17). Since 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.

➢ **Hypothesis testing for the slope**

The ***Regression*** tool in ***Data Analysis ToolPak*** also tests the hypothesis that the slope is equal to zero, i.e.,

$$H_0: b = 0 \text{ vs } H_1: b \neq 0$$

We use the p-value approach for decision making. From Fig. 11.9, the t-statistic for the slope is 18.5797 (Cell D18) and its p-value is 0.0000 (Cell E18), which is less than 0.05. Therefore, we may reject our null hypothesis at 5% level of significance and conclude that the slope is

not equal to zero, i.e., advertisement cost affects the monthly sales of the juice for the given data.

Note: When we have one predictor, we can also use t test instead of F test because the testing of slope using t test and the testing the significance of regression model using F test give the same result in the case of one predictor.

Note that the F ratio 345.2040 (Fig. 11.8) is just the square of t value, i.e., 18.5797 given in Fig. 11.9 for one predictor.

➢ **Confidence intervals for the regression coefficients**

Recall equations (13) and (14) for $(1 - \alpha)100\%$ confidence limits of the intercept and slope given in Sec. 11.3. MS Excel also provides us the confidence limits of intercept and slope as shown in Fig. 11.12.

Excel computes the 95% lower and upper confidence limits for the slope by typing

"=76.1845 – Tinv(0.05, 38) × 4.1004" and "=76.1845 + Tinv(0.05, 38) ×4.1004" in any two cells, respectively.

It will give 67.8837 and 84.4854 which correspond to lower and upper control limits, respectively.

	F	G	H	I
15				
16	*Lower 95%*	*Upper 95%*	*Lower 98.0%*	*Upper 98.0%*
17	-9882.0130	-3270.2723	-10542.0353	-2610.2501
18	67.8837	84.4854	66.2264	86.1427
19				

Fig. 11.12

We obtain the 95% confidence interval for intercept (a) given in Cells F17:G17 of Fig. 11.12. It is (–9882.0130, –3270.2723).

95% confidence interval for the slope (regression coefficient, b) given in Cells F18:G18 of Fig. 11.12 is (67.8837, 84.4854).

We can say that the rate of change in monthly sales will lie in the interval (67.8837, 84.4854), 95% of times when there is a unit change in the advertisement cost.

By default, Excel provides 95% confidence limits. If we want to find the confidence limits other than 95%, we specify the desired confidence level as shown in Fig. 11.5. For this problem, we have selected the ***Confidence Level*** box and set the level to 98% in the regression dialog box to find 98% confidence intervals.

From Fig. 11.12, the 98% confidence intervals for the intercept and slope are (–10542.0353, –2610.2501) and (66.2264, 86.1427), respectively.

11.5.4 Interpretation of the Residual Output

The **Regression** feature of Excel 2007 also includes other useful statistic(s) in the output such as **residuals, standardised residuals** and values needed for the **normal probability plot**. As you have studied in Units 9 and 12 of MSTE-002, we can check the adequacy of the fitted regression model by making use of residuals. We now discuss each result.

We can also manually calculate predicted sales, residuals and standard residuals as discussed in Block 3 of MSTE-002.

➢ **Predicted sales:** Fig. 11.13 shows the predicted sales in Cells B25:B64 for the given monthly expenditure on advertisement.

- **Residuals:** The residuals are shown in Cells C25:C64 of Fig. 11.13.
- **Standard residuals:** The standard residuals are shown in Cells D25:D64 of Fig. 11.13.

	A	B	C	D
22	RESIDUAL OUTPUT			
23				
24	*Observation*	*Predicted Sales*	*Residuals*	*Standard Residuals*
25	1	15517.3702	-117.3702	-0.0517
26	2	23897.6682	3902.3318	1.7184
27	3	21612.1324	-412.1324	-0.1815
28	4	33039.8115	-1639.8115	-0.7221
29	5	36087.1925	-187.1925	-0.0824
30	6	29992.4304	1807.5696	0.7960
31	7	18564.7513	2835.2487	1.2485
32	8	17041.0607	-1541.0607	-0.6786
33	9	13993.6797	-2793.6797	-1.2302
34	10	34563.5020	-2463.5020	-1.0848
35	11	18564.7513	3535.2487	1.5568
36	12	19326.5966	-1526.5966	-0.6723
37	13	23897.6682	2102.3318	0.9258
38	14	22373.9777	1026.0223	0.4518
39	15	26945.0493	654.9507	0.2884
40	16	26945.0493	-845.0493	-0.3721
41	17	26183.2040	-1983.2040	-0.8733
42	18	23897.6682	2502.3318	1.1019
43	19	20088.4418	-88.4418	-0.0389
44	20	24659.5135	-59.5135	-0.0262
45	21	20088.4418	3811.5582	1.6785
46	22	20088.4418	-1088.4418	-0.4793
47	23	23135.8229	-3635.8229	-1.6011

Regression Analysis | Regression

Fig. 11.13

We can use the residual output to verify the residual property as follows:

- **Verification of the residual property**

To verify the residual property that the sum of residuals is equal to zero, i.e., $\sum_{i=1}^{n} e_i = 0$, we type "=Sum(C25:C64)" in Cell C66 of Excel sheet named "**Regression Analysis**" and then press ***Enter*** as shown in Fig. 11.14.

It gives us the value of $\sum_{i=1}^{n} e_i$, which is equal to zero in this problem.

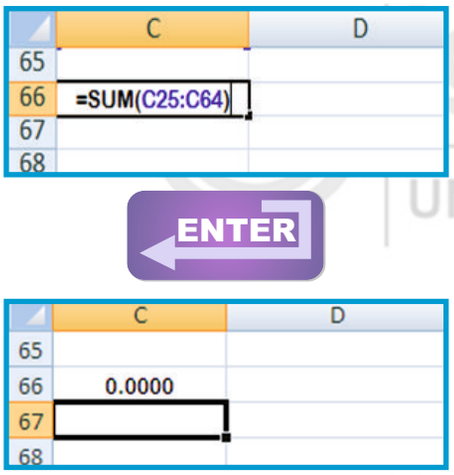


Fig. 11.14

Hence, the property $\sum_{i=1}^{n} e_i = 0$, is verified.

11.6 RESIDUAL PLOT IN EXCEL 2007

To conduct residual analysis, we plot the predicted sales versus standardised residuals given in the residual output shown in Fig. 11.15. For this purpose, we select the **Predicted Sales** data with labels (Cells B24:B64) and the **Standard Residuals** data with labels (Cells D24:D64) by holding the *Ctrl* key .

	A	B	C	D
22	RESIDUAL OUTPUT			
23				
24	*Observation*	*Predicted Sales*	*Residuals*	*Standard Residuals*
25	1	15517.3702	-117.3702	-0.0517
26	2	23897.6682	3902.3318	1.7184
27	3	21612.1324	-412.1324	-0.1815
28	4	33039.8115	-1639.8115	-0.7221
29	5	36087.1925	-187.1925	-0.0824
30	6	29992.4304	1807.5696	0.7960
31	7	18564.7513	2835.2487	1.2485
32	8	17041.0607	-1541.0607	-0.6786
33	9	13993.6797	-2793.6797	-1.2302
34	10	34563.5020	-2463.5020	-1.0848
35	11	18564.7513	3535.2487	1.5568
36	12	19326.5966	-1526.5966	-0.6723
37	13	23897.6682	2102.3318	0.9258
38	14	22373.9777	1026.0223	0.4518
39	15	26945.0493	654.9507	0.2884
40	16	26945.0493	-845.0493	-0.3721
41	17	26183.2040	-1983.2040	-0.8733
42	18	23897.6682	2502.3318	1.1019
43	19	20088.4418	-88.4418	-0.0389
44	20	24659.5135	-59.5135	-0.0262
45	21	20088.4418	3811.5582	1.6785
46	22	20088.4418	-1088.4418	-0.4793
47	23	23135.8229	-3635.8229	-1.6011

Regression Analysis | Regression

Select data

Fig. 11.15

We choose *Scatter* under the *Insert* tab to plot these residuals against predicted sales values as discussed in Step 2 of Sec. 11.4 (Fig. 11.16).

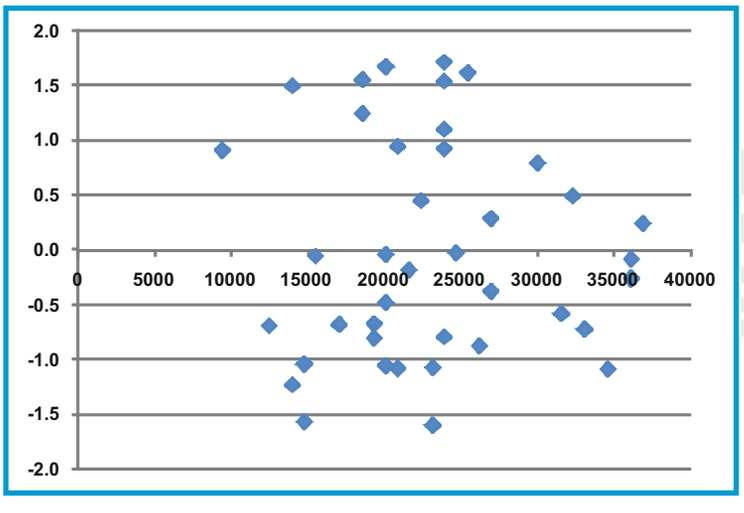


Fig. 11.16

We select various formatting options for the chart as explained in Sec. 3.3 of Lab Session 3 of MSTL-001. Here we need to carry out one more step. By default, the X-axis labels appear in the middle of a residual graph (across a Y value of 0) as shown in Fig. 11.16.

To move the labels of the Y-axis to the bottom of the chart, we

1. click the ***Y-axis*** of the chart shown in Fig. 11.16,
2. choose ***Format Selection*** under the ***Format*** tab as shown in Fig. 11.17a, and
3. specify the ***Horizontal axis crosses*** at the lower axis value of Y in the resulting dialog box as shown in Fig. 11.17b. We have used the value −2.0 here.

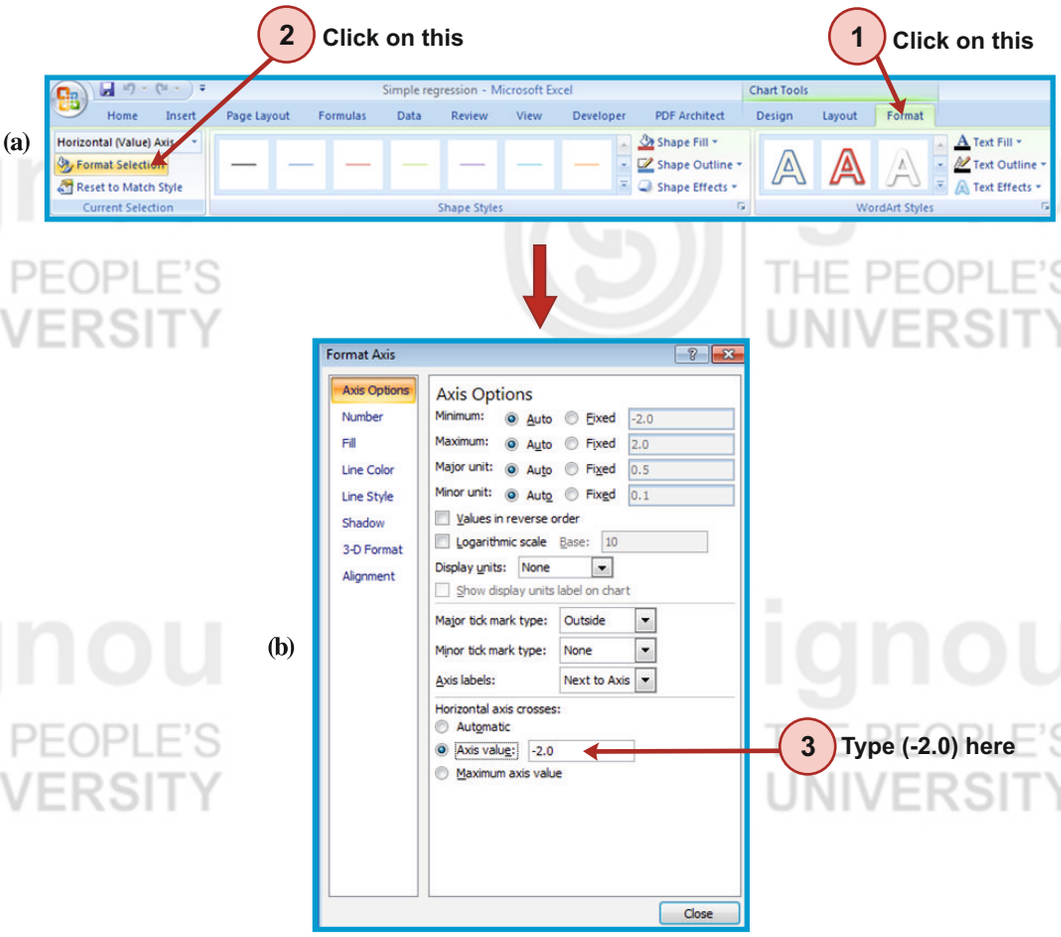


Fig. 11.17

We format the chart as explained in Sec. 3.3 of Lab Session 3 of MSTL-001 and obtain the residual plot shown in Fig. 11.18.

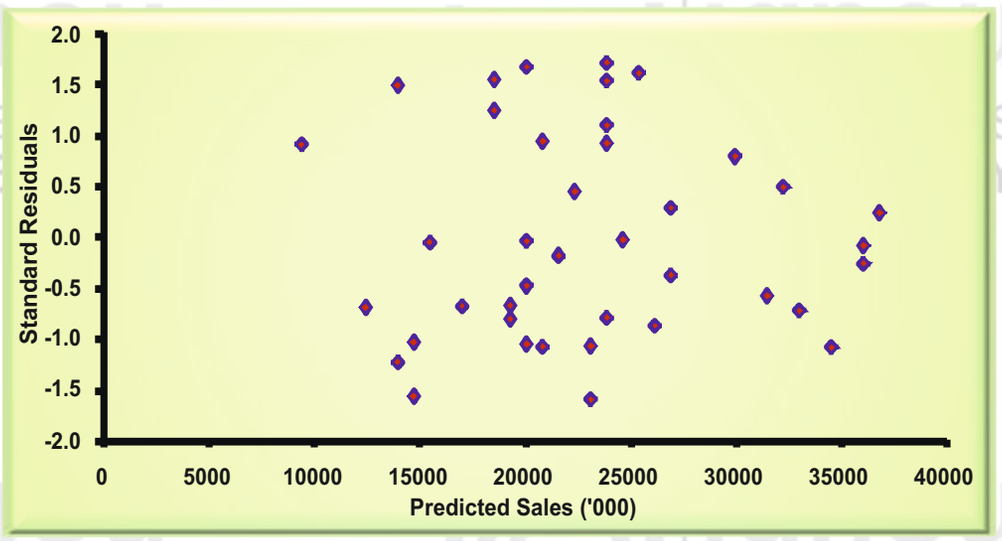


Fig. 11.18: Residual Plot.

The standard residuals shown in Fig. 11.18 appear to be fairly randomly scattered throughout a horizontal band around 0.0 (Y-axis). Hence, the assumption of linear regression is valid or we can say that the simple linear regression model fitted well for the given data.

We can also manually calculate these percentiles from equation (15).

11.7 NORMAL PROBABILITY PLOT IN EXCEL 2007

By constructing the normal probability plot for the residuals, we check the normality assumption as the t-test, F-test and confidence interval depend on the normality assumption of the residuals.

To plot a normal probability plot, we shall use the ***Regression*** tool of ***Data Analysis ToolPak***, which calculates the percentiles given in Cells F25:F64 of Fig. 11.19.

	E	F	G
22		PROBABILITY OUTPUT	
23			
24		*Percentile*	*Sales*
25		1.25	10900
26		3.75	11200
27		6.25	11200
28		8.75	11500
29		11.25	12400
30		13.75	15400
31		16.25	15500
32		18.75	17400
33		21.25	17500
34		23.75	17700
35		26.25	17800
36		28.75	18400
37		31.25	19000
38		33.75	19500
39		36.25	20000
40		38.75	20700
41		41.25	21200
42		43.75	21400
43		46.25	22100
44		48.75	22100
45		51.25	23000
46		53.75	23400
47		56.25	23900

Regression Analysis | Regression

Fig. 11.19

We follow the steps given below to obtain the normal probability plot.

Step 1: Refer to Fig. 11.20. We

- ✓ type "**Ordered Residuals**" in Cell E24,
- ✓ select and ***Copy*** Cells C25:C64,
- ✓ click on Cell E25 and choose ***Paste*** option in ***Home*** tab and paste the data in Cells E25:E64.
- ✓ select Cells E25:E64 (Fig. 11.20a),
- ✓ choose ***Sort Smallest to Largest*** from ***Sort & Filter*** option in ***Home*** tab to arrange these values in increasing order as shown in Fig. 11.20b, and
- ✓ check on the ***Continue with the current selection*** and click on ***Sort*** as shown in Fig. 11.20c.

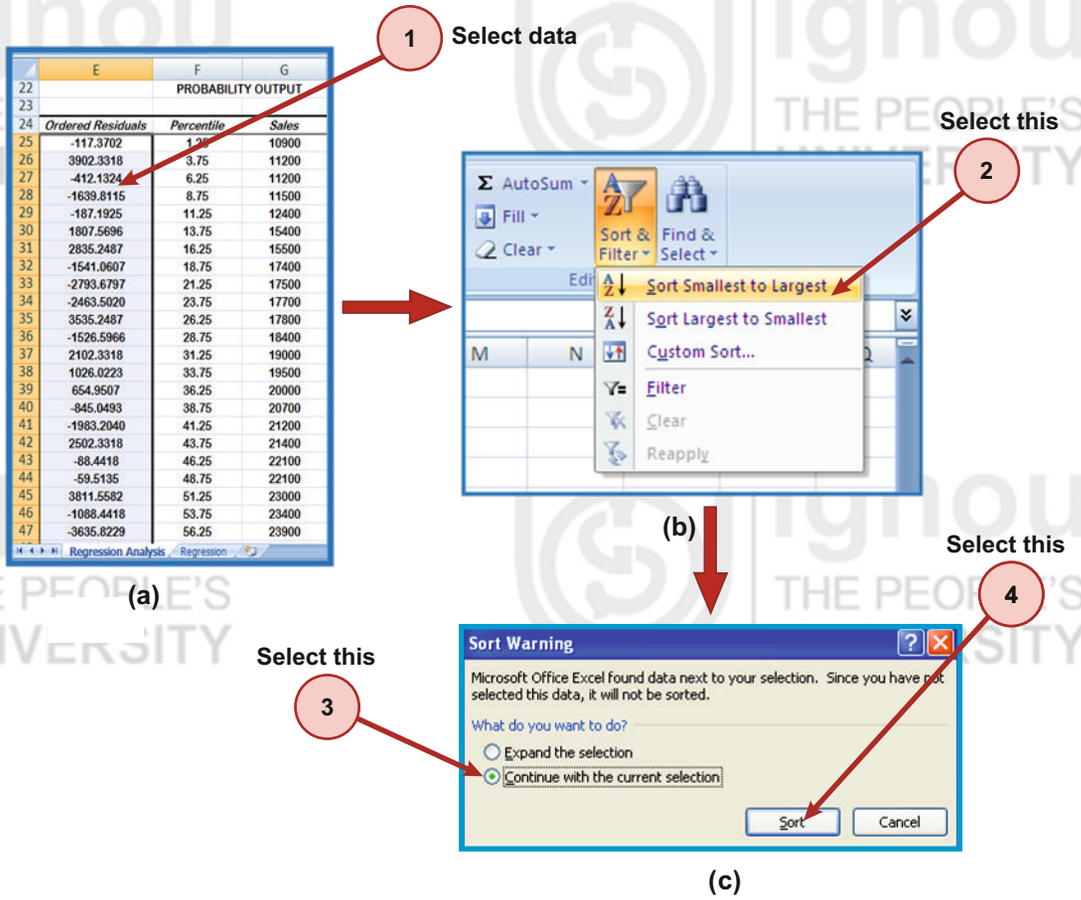


Fig. 11.20

Step 2: In this way, we arrange the selected range of data in ascending order as shown in Fig. 11.21.

	E	F	G
22		PROBABILITY OUTPUT	
23			
24	*Ordered Residuals*	*Percentile*	*Sales*
25	-3635.8229	1.25	10900
26	-3555.5249	3.75	11200
27	-2793.6797	6.25	11200
28	-2463.5020	8.75	11500
29	-2450.2871	11.25	12400
30	-2435.8229	13.75	15400
31	-2388.4418	16.25	15500
32	-2355.5249	18.75	17400
33	-1983.2040	21.25	17500
34	-1826.5966	23.75	17700
35	-1797.6682	26.25	17800
36	-1639.8115	28.75	18400
37	-1569.9891	31.25	19000
38	-1541.0607	33.75	19500
39	-1526.5966	36.25	20000
40	-1316.1209	38.75	20700
41	-1088.4418	41.25	21200
42	-845.0493	43.75	21400
43	-587.1925	46.25	22100
44	-412.1324	48.75	22100
45	-187.1925	51.25	23000
46	-117.3702	53.75	23400
47	-88.4418	56.25	23900

Regression Analysis | Regression

Fig. 11.21

Step 3: We now select Cells E24:F64 and obtain the scatter plot for this data as explained in Sec. 11.4. We format the chart as explained in Sec. 3.3 of Lab Session 3 of MSTL-001. Fig. 11.22 shows the normal probability plot.

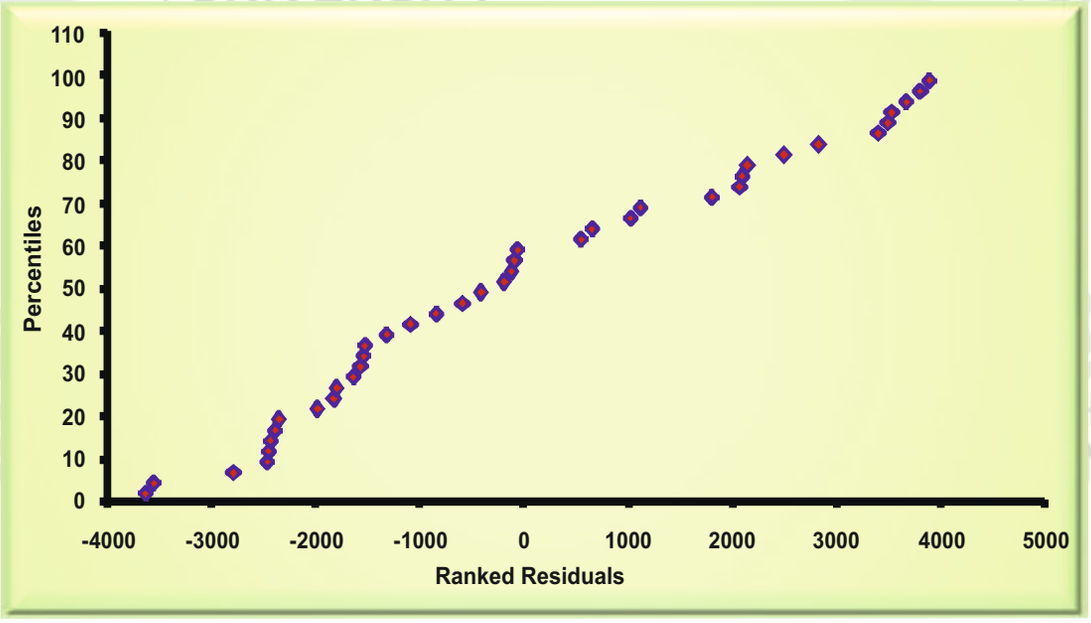


Fig. 11.22: Normal probability plot.

The normal probability plot shown in Fig. 11.22 reveals that the resulting points lie approximately along a straight line. It indicates that the distribution of error terms (residuals) is approximately normally distributed.

11.8 FITTED LINE OF REGRESSION ON THE SCATTER PLOT IN EXCEL 2007

You have studied the manual plotting of regression line on the scatter diagram in Unit 9 of MSTE-002. In the following steps, we explain how to draw the regression line on the scatter diagram directly in Excel 2007:

Step 1: To plot the fitted regression line on the scatter diagram (Fig. 11.3), we

1. click on the scatter diagram shown in Fig. 11.3,
2. click on ***Layout*** tab under ***Chart Tools***, and
3. choose ***More Trendline Options*** from the ***Trendline*** menu as shown in Fig. 11.23a.

Step 2: A new dialog box appears as shown in Fig. 11.23b. In this dialog box, we

- ✓ check on the ***Linear*** under ***Trend/Regression Type***,
- ✓ write the trendline name under ***Custom*** option. In this example, we write "**Fitted Regression Line**",
- ✓ tick on the ***Display Equation on chart*** if we wish to display the equation of regression line on the chart, and
- ✓ choose the option ***Close***.

You can also do the calculations described in Secs. 11.4, 11.6, 11.7, 11.8 and 11.10 manually as you have learnt in Units 9 and 10 of MSTE-002.

It will give you the same results as you have obtained in this lab session.

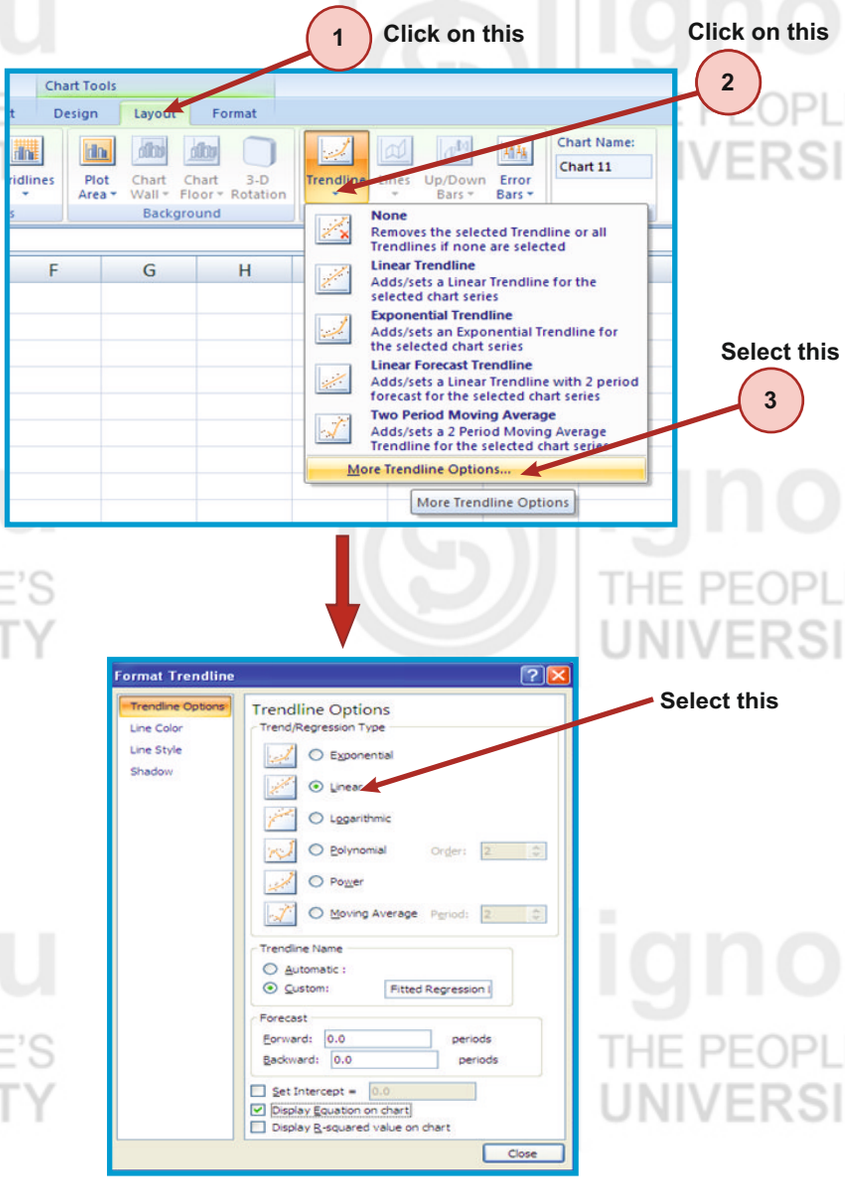


Fig. 11.23

Thus, we get the scatter diagram with fitted regression line as shown in Fig. 11.24.

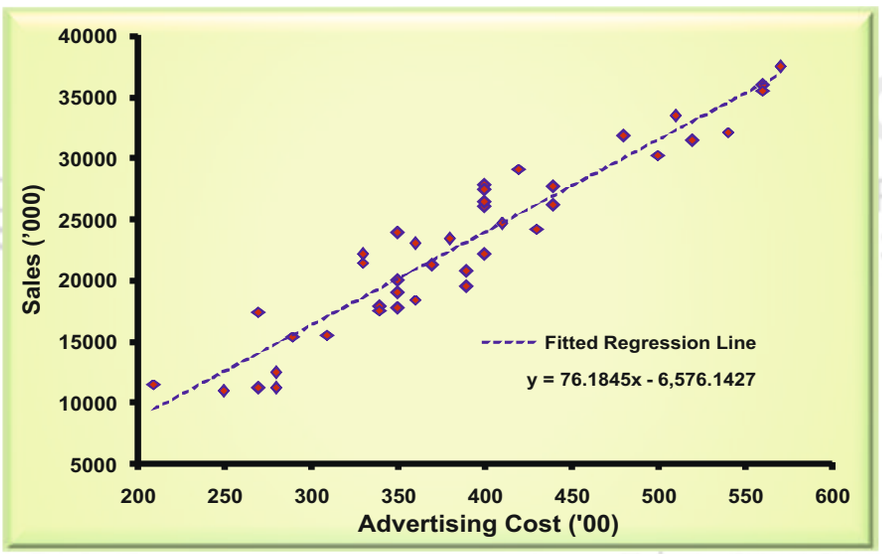


Fig. 11.24

Before studying further, do the following activities for practice.

Activity 1

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

Now, you can carry out regression analysis for your own data and interpret the results as well.

Activity 2

Use MS Excel 2007 to work out the following exercises and interpret the results:

A1) Examples 1 and 2 given in Unit 9 of MSTE-002.

A2) Exercises E7 and E9 given in Unit 9 of MSTE-002.

A3) Examples 1, 2, 3, 4 and 5 given in Unit 10 of MSTE-002.

A4) Exercises E1, E3, E4 and E5 given in Unit 10 of MSTE-002.

You may also like to match the results with the manual calculations done in Units 9 and 10 of MSTE-002.

You should now apply the method to another problem.

Continuous Assessment 11

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

Table 2: Electricity consumption data

S. No.	Electricity Consumption (in kWh)	Size of House (in sq ft)	S. No.	Electricity Consumption (in kWh)	Size of House (in sq ft)
1	1060	1316	21	1565	1696
2	1150	1420	22	1215	1464
3	1365	1556	23	1275	1488
4	1275	1488	24	1465	1632
5	1425	1612	25	1080	1356
6	1310	1516	26	975	1196
7	1365	1556	27	1040	1256

S. No.	Electricity Consumption (in kWh)	Size of House (in sq ft)	S. No.	Electricity Consumption (in kWh)	Size of House (in sq ft)
8	1075	1352	28	1340	1540
9	925	1168	29	865	1144
10	1340	1540	30	1175	1440
11	1425	1612	31	1080	1356
12	1150	1420	32	1500	1652
13	1060	1316	33	1175	1440
14	1545	1680	34	1050	1296
15	1140	1388	35	1365	1580
16	1075	1352	36	1465	1632
17	1620	1736	37	1215	1464
18	1050	1296	38	1365	1580
19	1310	1516	39	1140	1388
20	1645	1760	40	1005	1224

For this data,

- Plot a scatter diagram to get a rough idea about the relationship. If it reveals linear relationship, use this data to develop a simple linear regression model.
- Test the significance at 1% level of significance and find the 99% confidence interval of the regression parameters.
- Also check the linearity and normality assumptions for the regression analysis.

Home Work: Do It Yourself

1) Follow the steps explained in Secs. 11.4 to 11.8 to comprehend the regression analysis for the data of Table 1. Use a different format for the scatter, residual and normal probability plots. Take their screenshots and keep them in your record book.
2) Develop the spreadsheets for the exercise "Continuous Assessment 11" 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.

Multiple Linear Regression

SESSION 12

Structure

12.1 Introduction
Objectives
12.2 Problem Description
12.3 Procedure for Multiple Regression Analysis
12.4 Matrix Plot in Excel 2007
12.5 Fitting and Analysis of Multiple Regression Model in Excel 2007
12.6 Interpretation of the Results of Multiple Regression Analysis
Interpretation of the Regression Statistics Table
Interpretation of the ANOVA Table
Interpretation of the Regression Coefficients Table
Interpretation of the Residuals Output
12.7 Residual Plot in Excel 2007
12.8 Normal Probability Plot in Excel 2007
12.9 Matrix Approach in Excel 2007

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.

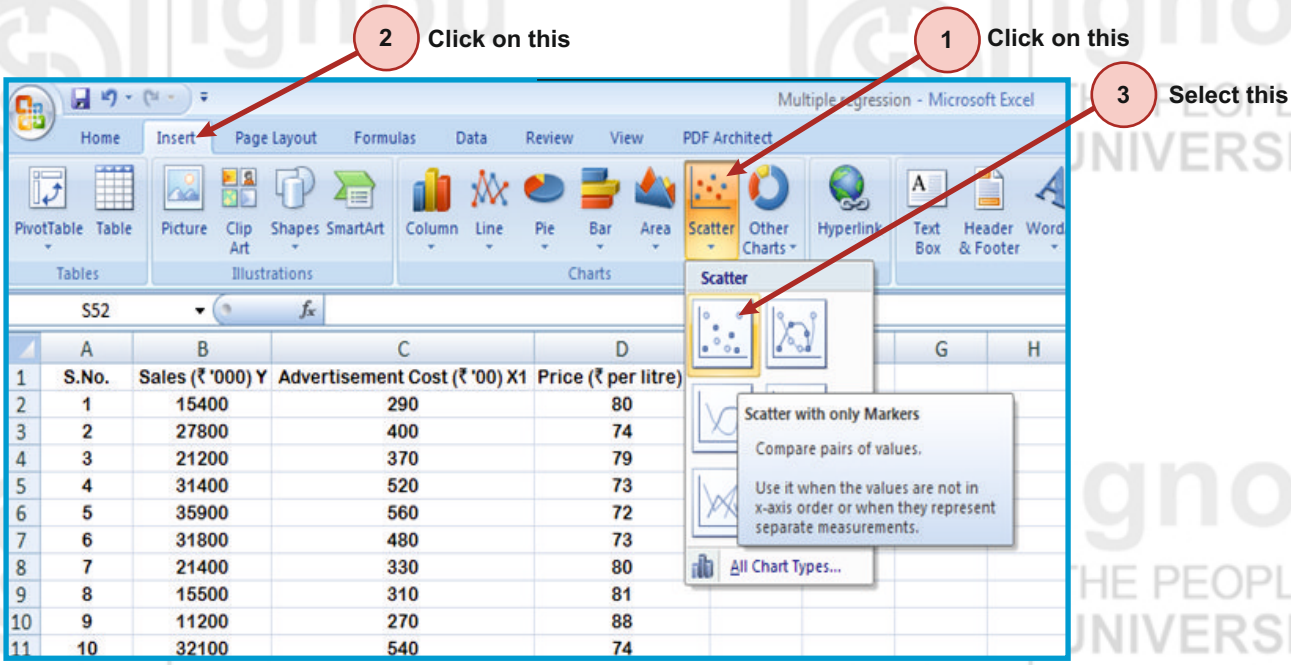


Fig. 12.2

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.

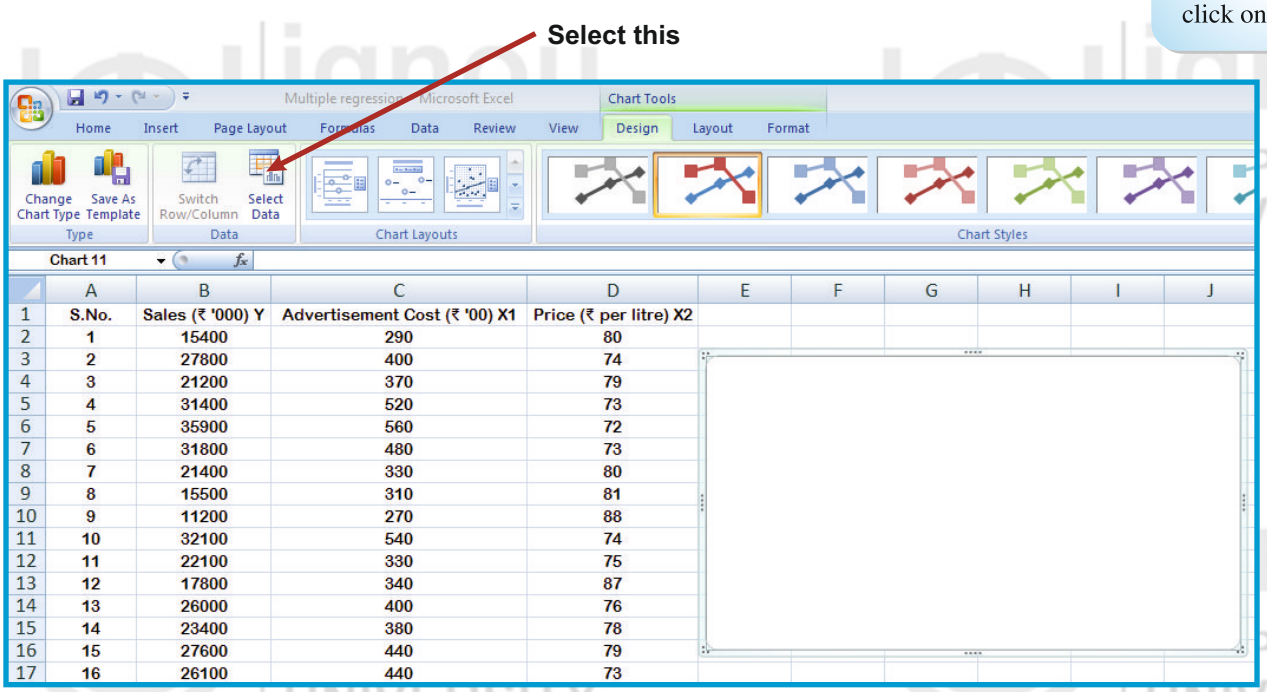


Fig. 12.3

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.

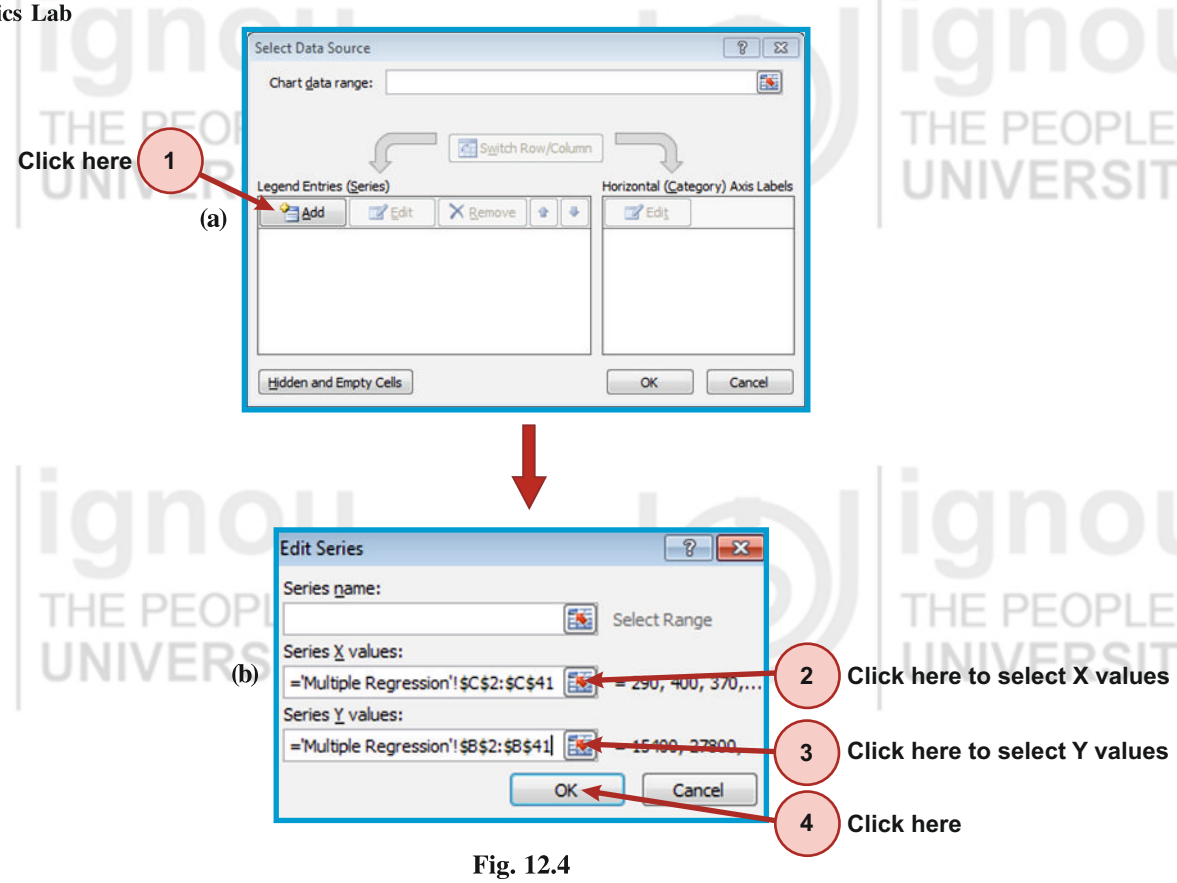


Fig. 12.4

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.

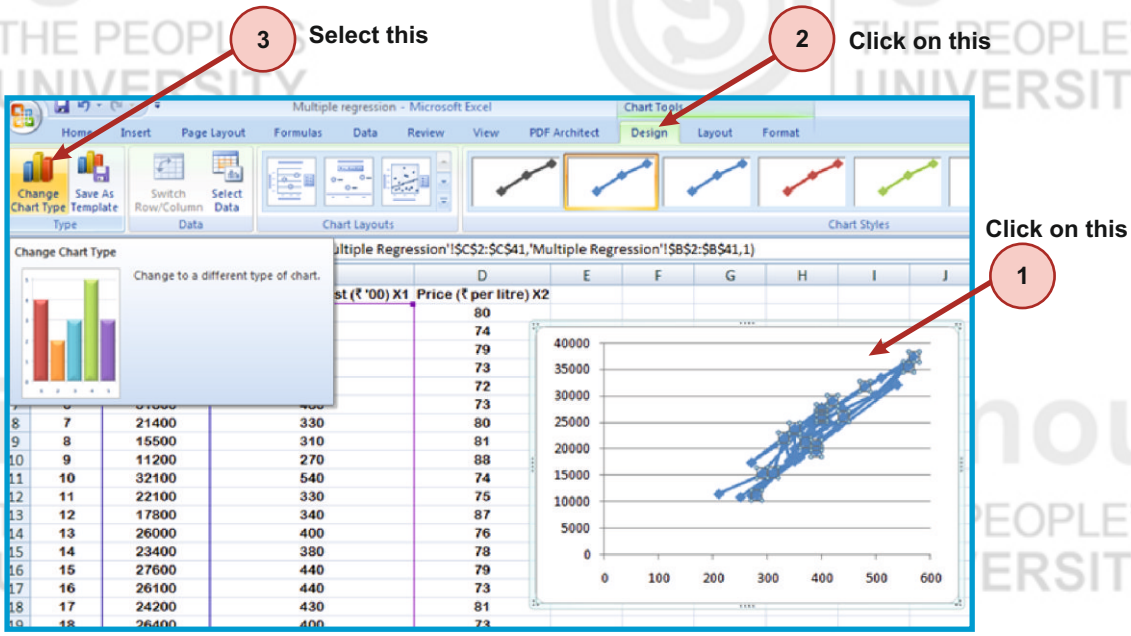


Fig. 12.5

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.

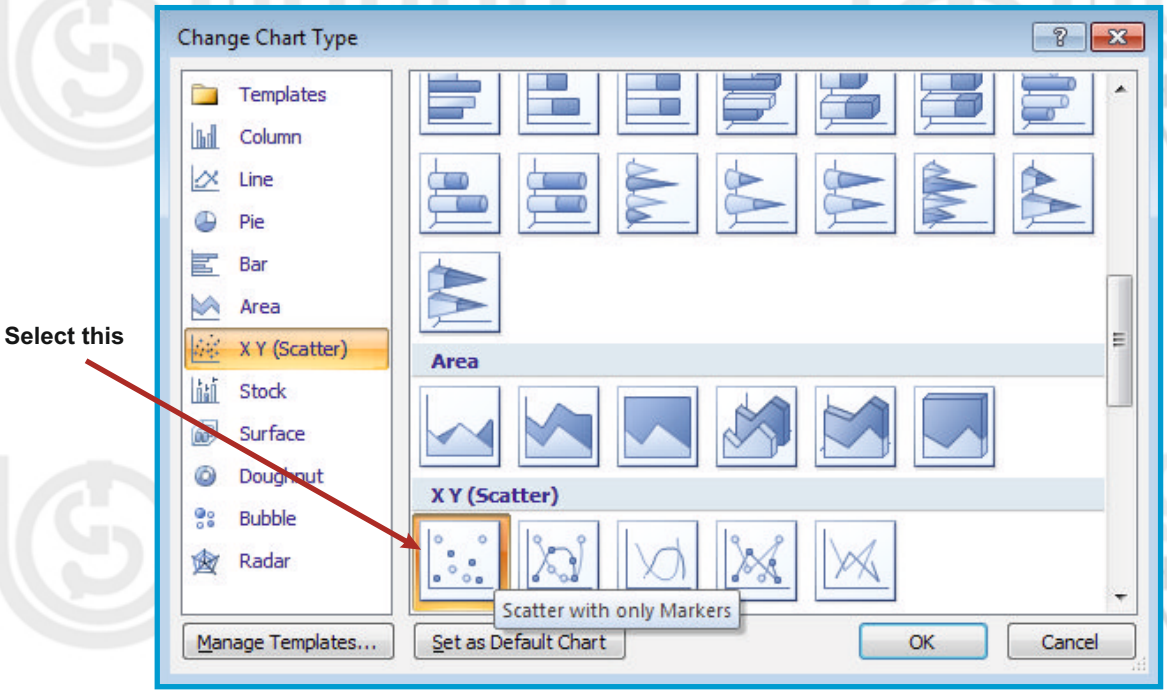


Fig. 12.6

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.

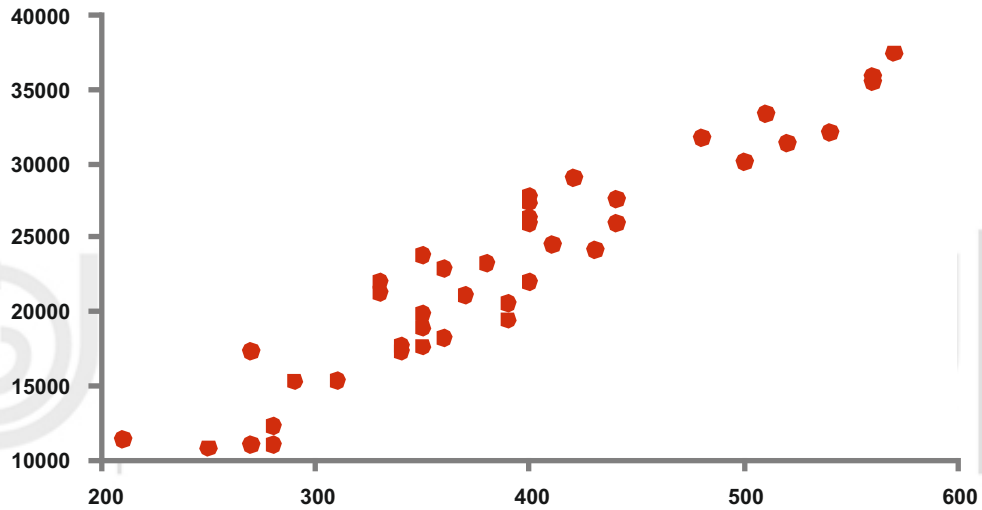


Fig. 12.7

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

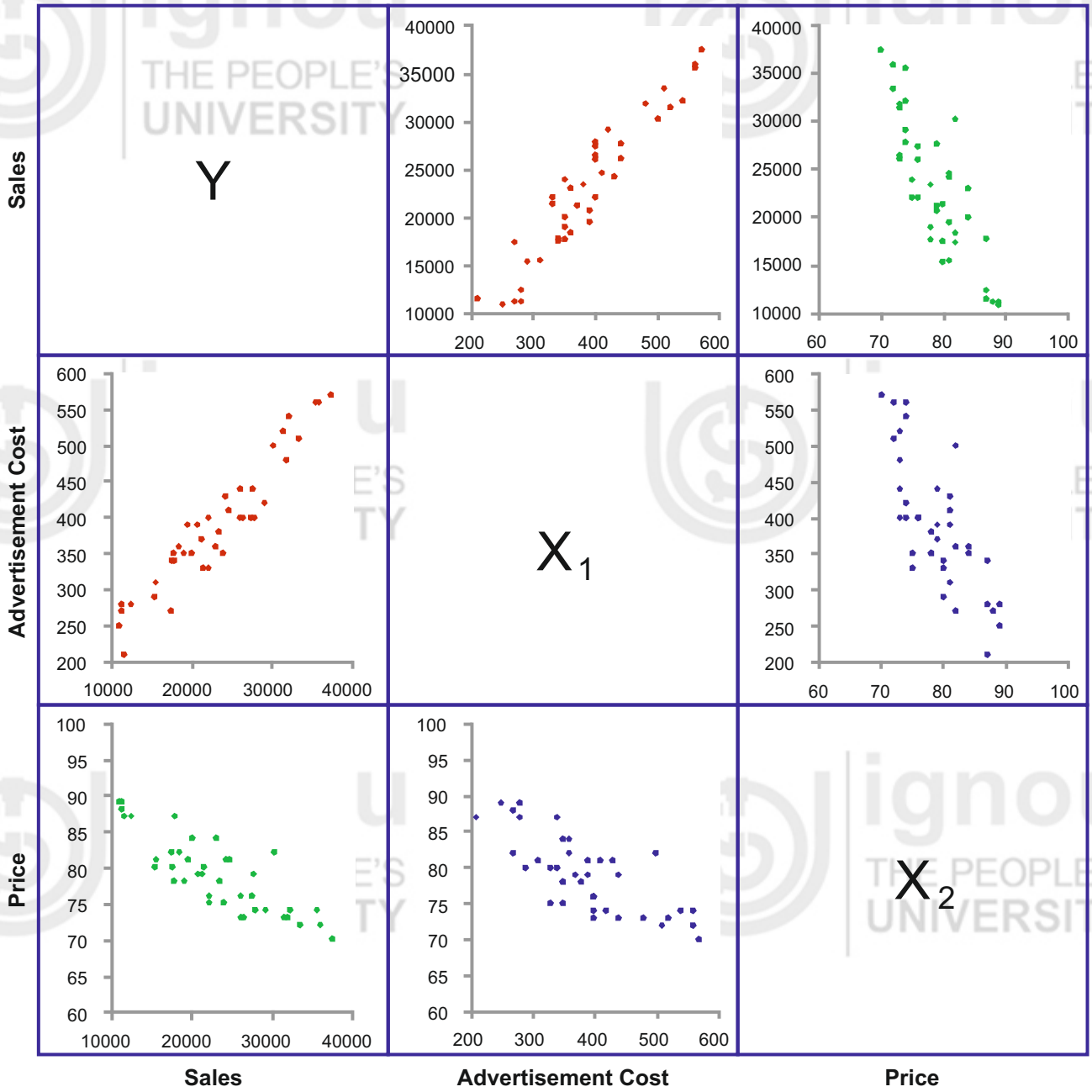


Fig. 12.8: Matrix of scatter plots for the monthly sales data.

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***.

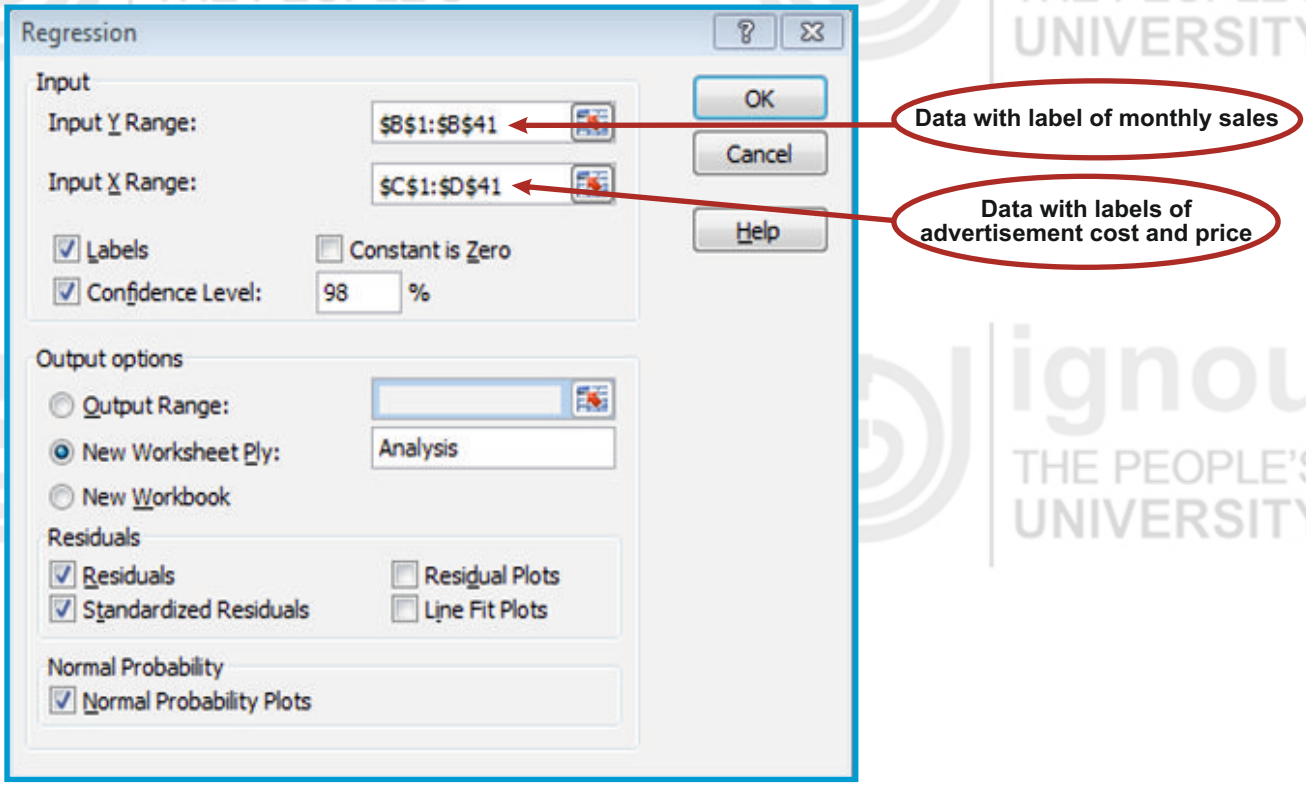


Fig. 12.9

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

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

  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

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

- **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 $(1-\alpha)100\%$ 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.

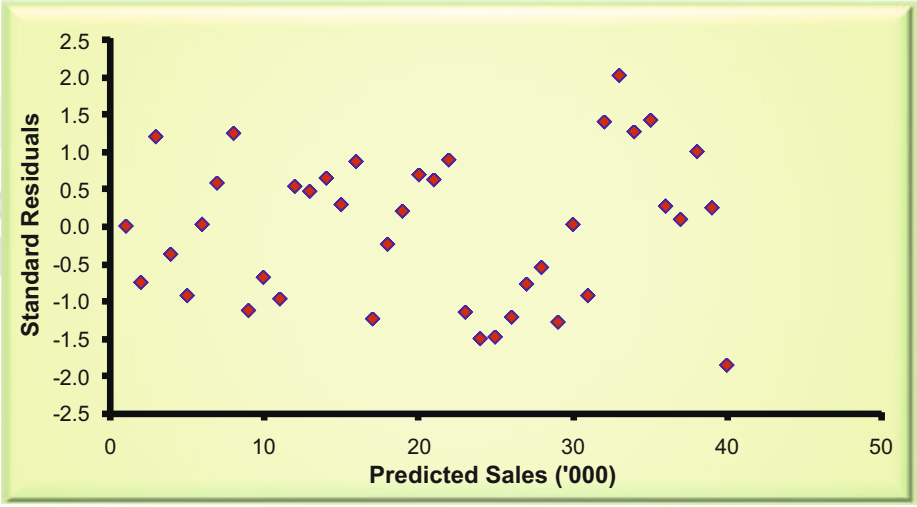


Fig. 12.15: Residual plot.

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.

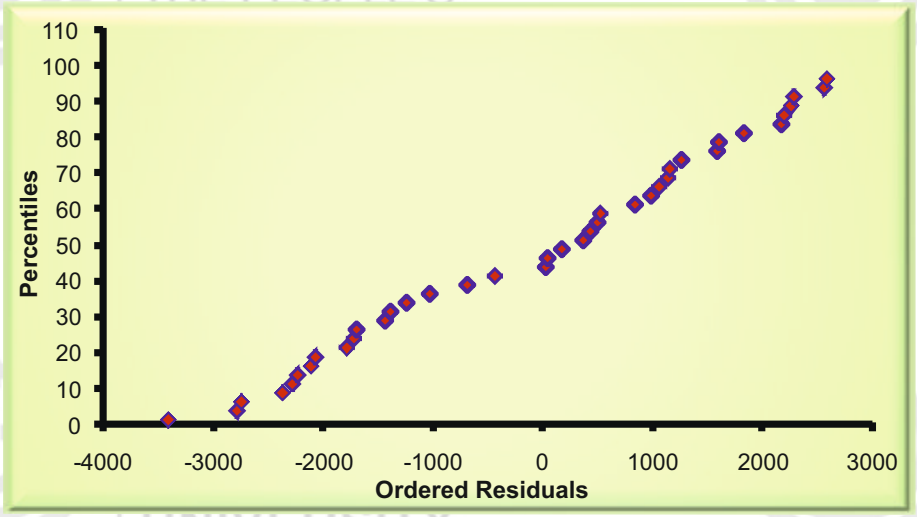


Fig. 12.17: Normal probability plot.

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:

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

where

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

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

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

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

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 **B**, denoted by $\hat{\mathbf{B}}$, can be written as

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

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

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

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. $= n-1$)

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

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

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

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

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

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

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

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

Step 9: F-statistic is given by

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

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

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

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

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

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

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

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

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

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.

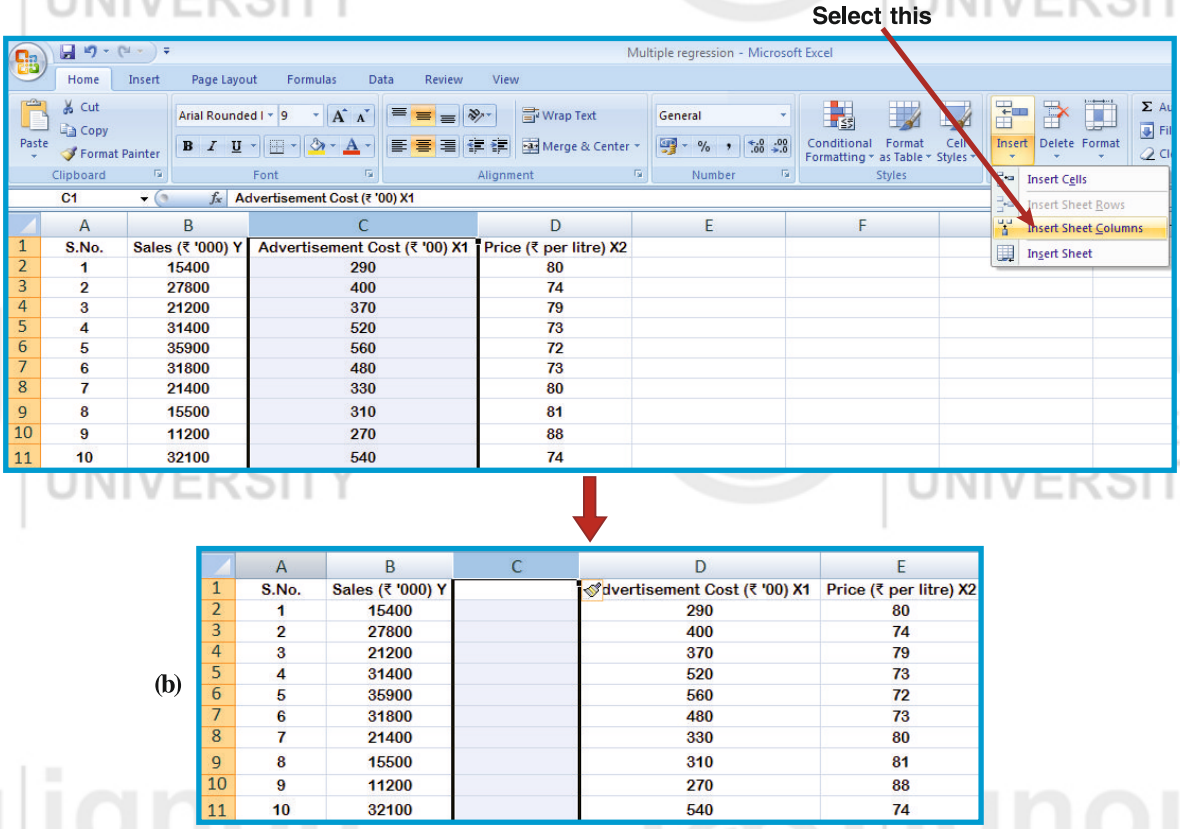
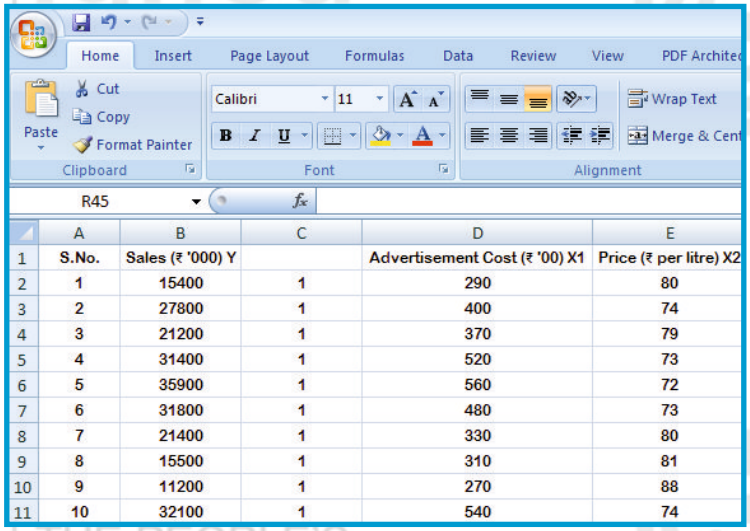


Fig. 12.19

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



	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

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

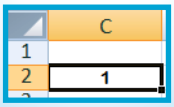


Fig. 12.20

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

Analysis | Multiple Regression | Matrix Approach

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.

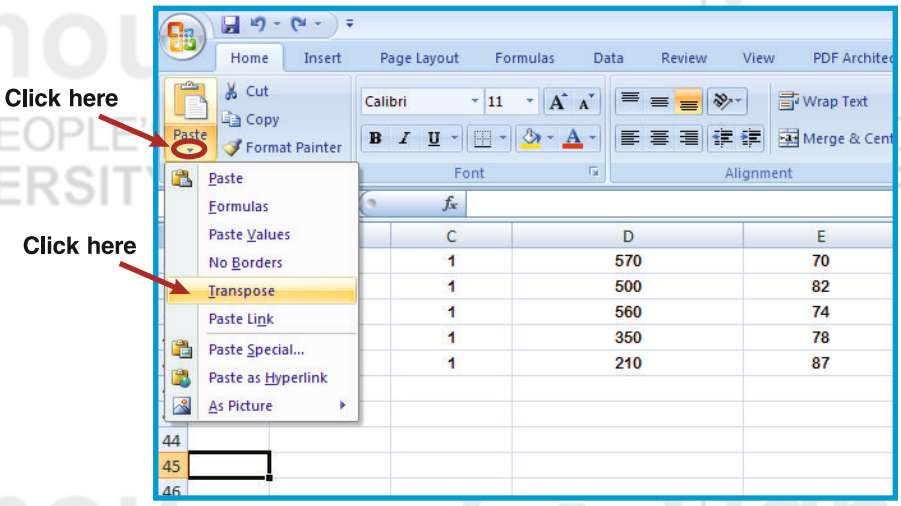


Fig. 12.22

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
44				
45	S.No.	1	2	3
46	Sales (₹ '000) Y	15400	27800	21200
47		1	1	1
48	Advertisement Cost (₹ '00) X1	290	400	370
49	Price (₹ per litre) X2	80	74	79
50				

	AM	AN	AO
44			
45	38	39	40
46	35500	17700	11500
47	1	1	1
48	560	350	210
49	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.

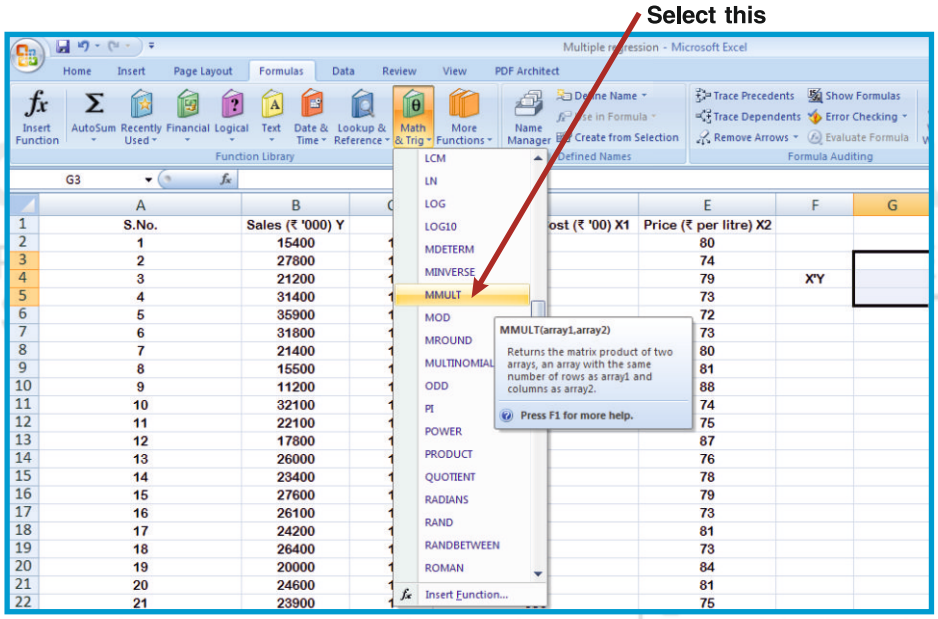


Fig. 12.24

Step 8: We select the range of **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.

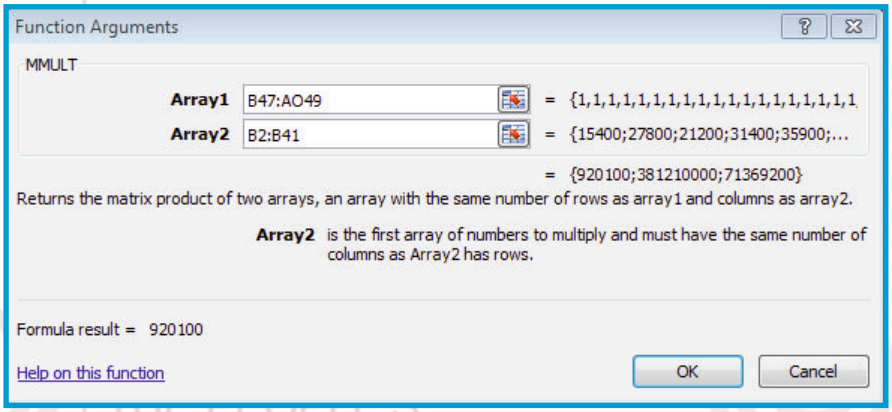


Fig. 12.25

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

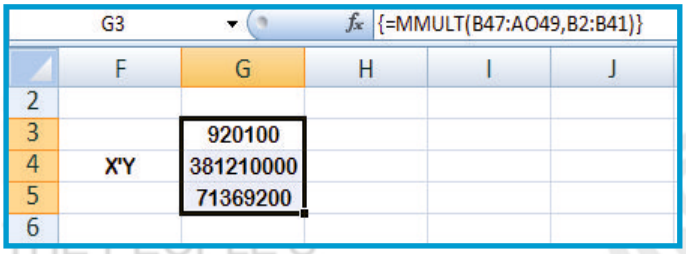


Fig. 12.26

Step 10: **X′** and **X** are matrices of orders (3×40) and (40×3), respectively. So **X′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 **X′X** matrix.

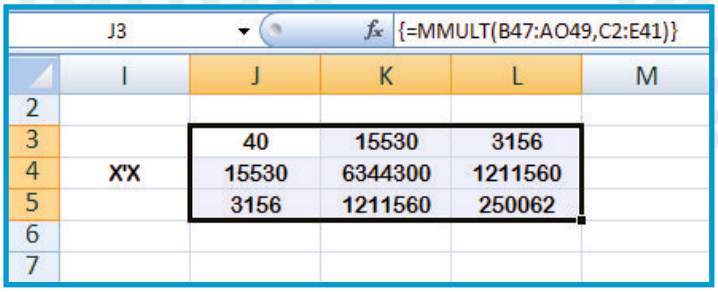


Fig. 12.27

Step 11: For finding the inverse of **X′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).

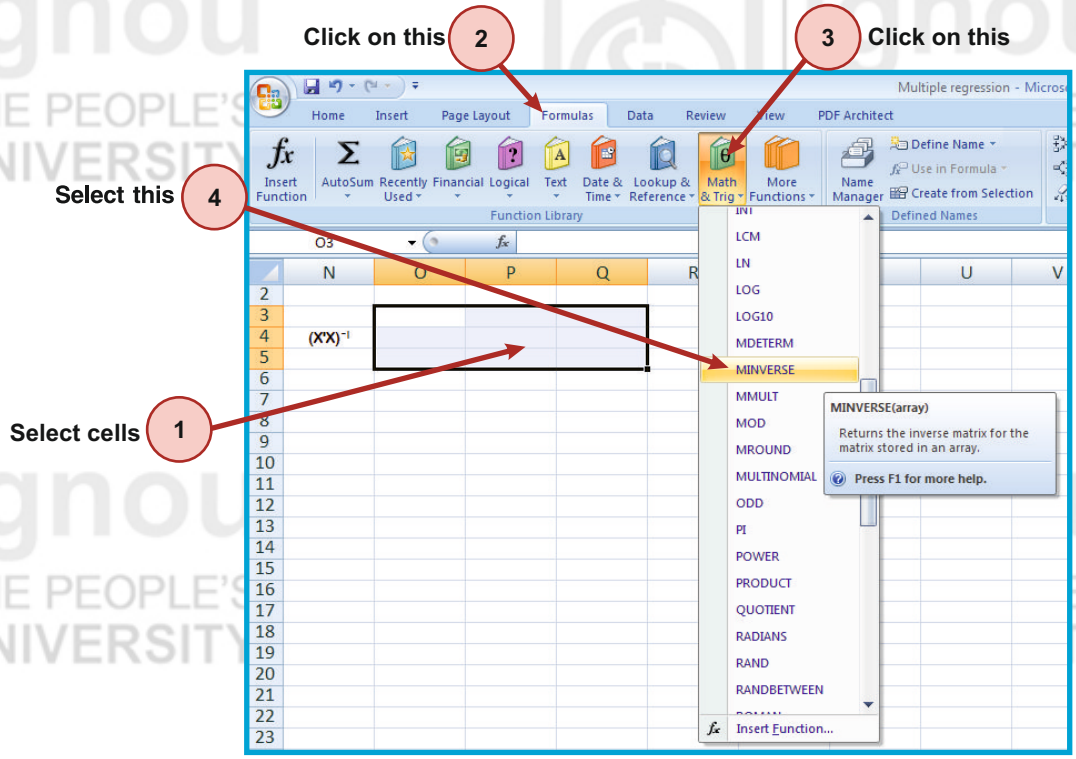
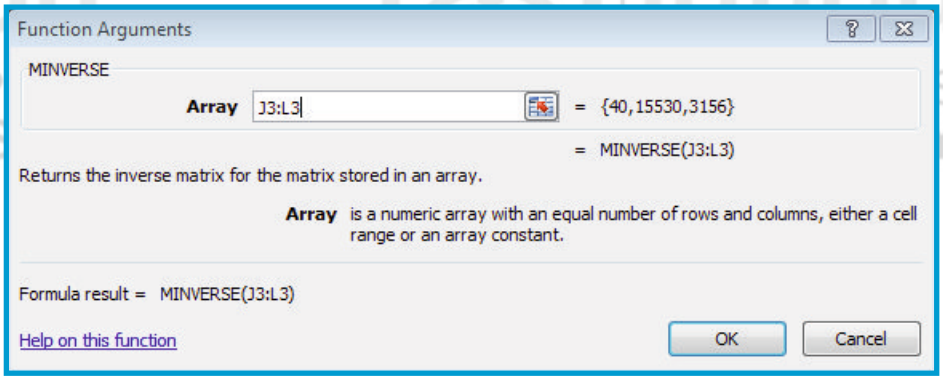


Fig. 12.28

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.

(a)



(b)

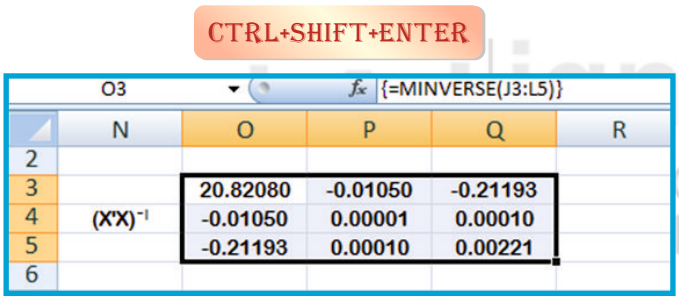


Fig. 12.29

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.

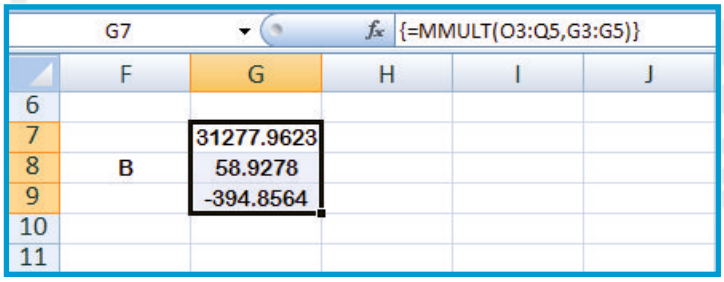


Fig. 12.30

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.

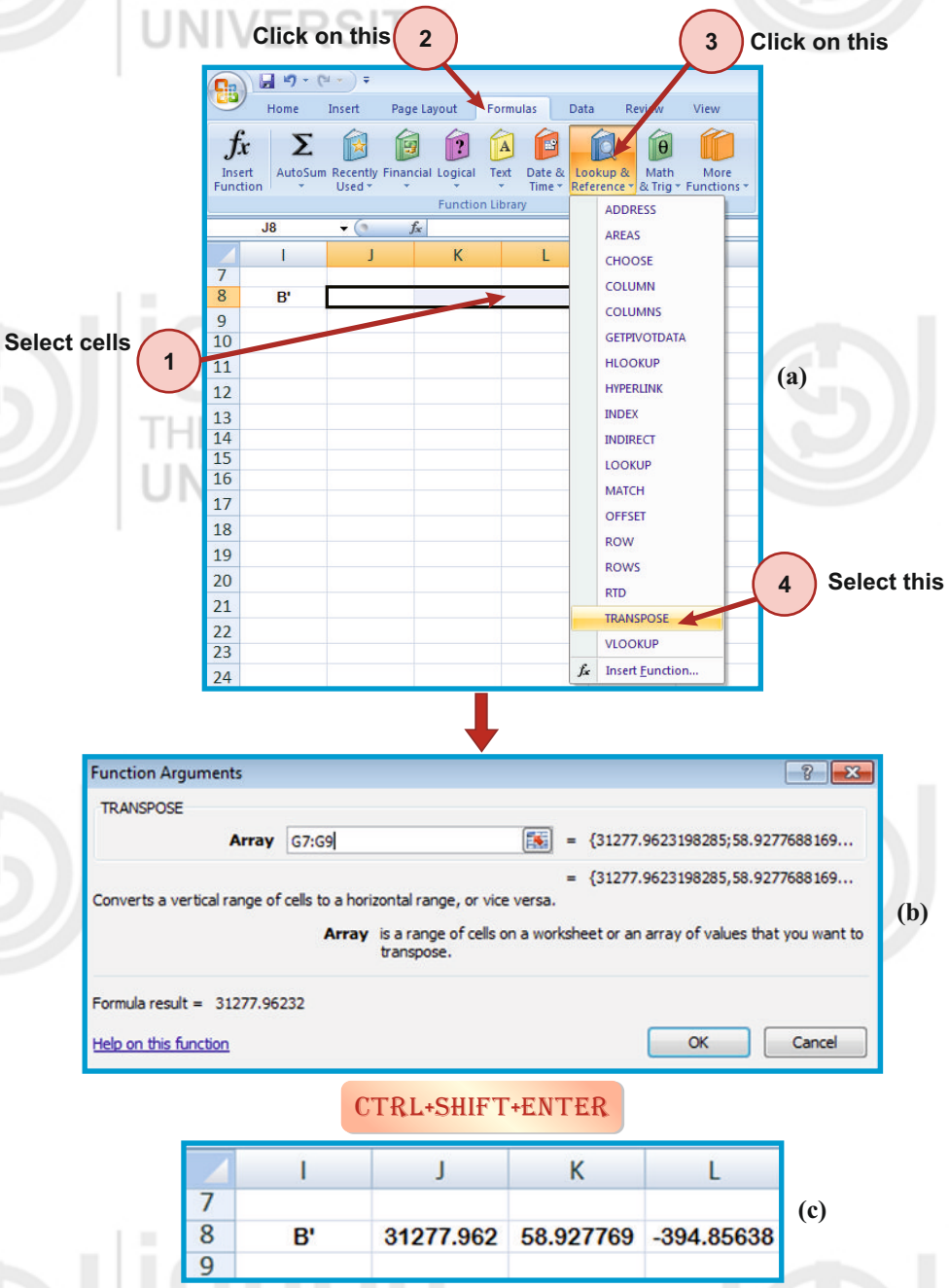


Fig. 12.31

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	$SE(\hat{B}_0) = \sqrt{\sigma_{00}}$	8572.2838	→ = Sqrt(O7)
12	$SE(\hat{B}_1) = \sqrt{\sigma_{11}}$	5.1102	→ = Sqrt(P8)
13	$SE(\hat{B}_2) = \sqrt{\sigma_{22}}$	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.

Regression with Dummy Variable

SESSION 13

Structure

13.1 Introduction
Objectives
13.2 Problem Description
13.3 Procedure for Multiple Regression Analysis with Dummy Variable
13.4 Scatter Plot in Excel 2007
13.5 Fitting and Analysis of Regression with Dummy Variable in Excel 2007
13.6 Fitted Regression Lines

13.1 INTRODUCTION

Prerequisite

- Lab Sessions 11 and 12 of MSTL-002 (Industrial Statistics Lab).
- Unit 11 of MSTE-002 (Industrial Statistics-II).

In Lab Sessions 11 and 12, you have learnt how to solve problems on simple and multiple regression analysis, respectively, where both the independent and dependent variables are quantitative. But in real life, we may come across many situations, in which one or more independent variables are qualitative such as sex (male/female), regions (the East/West/North/South), promotional activities (yes/no), etc. These types of variables are usually known as **indicators** or **dummy variables**.

In this lab session, we use Excel 2007 for multiple regression analysis when at least one of the independent variables is qualitative. We also draw a scatter plot to get an idea about the variation in the relationship between quantitative dependent and independent variables at different levels of the dummy variable.

Objectives

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

- prepare the spreadsheet for regression analysis with dummy variable in MS Excel 2007;
- fit the linear regression lines at different levels of the dummy variable;
- test the significance of the regression parameters;
- determine the confidence intervals of the regression parameters;
- construct the residual plot and normal probability plot; and
- interpret the results of regression analysis with dummy variable.

13.2 PROBLEM DESCRIPTION

For this lab session we consider the monthly sales data where promotional activity is considered as a dummy variable. We use 0 for no promotional activity and 1 for promotional activity for the particular month. In this exercise, you will use Excel 2007 to see how the monthly sales vary with the advertisement cost. We may also check whether the relationship between sales and advertisement cost is the same for both levels of promotional activity or not.

Table 1: Data of a juice manufacturing company

S. No.	Sales (₹'000) Y	Advertisement Cost (₹'00) X_1	Promotion X_2	S. No.	Sales (₹'000) Y	Advertisement Cost (₹'00) X_1	Promotion X_2
1	15400	290	1	21	23900	350	1
2	27800	400	1	22	19000	350	0
3	21200	370	1	23	19500	390	0
4	31400	520	0	24	22100	400	0
5	35900	560	1	25	17500	340	0
6	31800	480	1	26	11200	280	0
7	21400	330	0	27	12400	280	0
8	15500	310	1	28	20700	390	0
9	11200	270	0	29	10900	250	0
10	32100	540	0	30	18400	360	0
11	22100	330	1	31	27400	400	1
12	17800	340	0	32	23000	360	1
13	26000	400	1	33	29100	420	1
14	23400	380	1	34	17400	270	1
15	27600	440	0	35	33400	510	1
16	26100	440	0	36	37400	570	1
17	24200	430	0	37	30200	500	0
18	26400	400	1	38	35500	560	1
19	20000	350	1	39	17700	350	0
20	24600	410	0	40	11500	210	0

In this exercise, we shall

- prepare a scatter plot to get an idea about the relationship among the variables;
- use this data to develop a linear regression model and its related analysis at 1% level of significance;
- check the linearity and normality assumptions for the regression analysis; and
- draw both fitted regression lines on the scatter plot.

13.3 PROCEDURE FOR MULTIPLE REGRESSION ANALYSIS WITH DUMMY VARIABLE

You have learnt about regression analysis with dummy variable in Unit 11 of MSTE-002. So we list the formulae and briefly explain the procedure for regression analysis with dummy variable as follows:

Step 1: Let Y and X_1 be the quantitative dependent and independent variables, respectively. If X_2 is a qualitative independent variable with two levels 0 and 1, the regression model defined in equation (1) of Lab Session 12 represents the two different equations of the regression line as follows:

$$Y = B_0 + B_1X_1 + e, \qquad \text{for } X_2 = 0 \qquad ...(1)$$

and $$Y = (B_0 + B_2) + B_1X_1 + e, \quad \text{for } X_2 = 1 \qquad ...(2)$$

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: Fitting of regression model with dummy variable and other related analysis can be done in the same way as in Lab Session 12.

First of all, we explain how to draw a scatter plot in the case of dummy variable.

13.4 SCATTER PLOT IN EXCEL 2007

In Lab Sessions 11 and 12, we have plotted the scatter diagram to get an idea of the relationship between sales and advertisement cost. The method of plotting the scatter diagram is the same as described in Lab Sessions 11 and 12. The difference is that we show both levels of the promotional activity differently in this lab session so that we can distinguish between the data points corresponding to both levels. For the data given in Table 1, the promotional activity is the dummy variable.

Let $X_2 = \begin{cases} 0, & \text{No} \\ 1, & \text{Yes} \end{cases}$

The main steps for drawing a scatter plot in the case of dummy variable with the help of MS Excel 2007 are given below:

Step 1: We enter the given data in Excel spreadsheet and name it "**Regression_Dummy**" as shown in Fig. 13.1.

	A	B	C	D
1	S.No.	Sales (₹ '000) Y	Advertisement Cost (₹ '00) X1	Promotion X2
2	1	15400	290	1
3	2	27800	400	1
4	3	21200	370	1
5	4	31400	520	0
6	5	35900	560	1
7	6	31800	480	1
8	7	21400	330	0
9	8	15500	310	1
10	9	11200	270	0
11	10	32100	540	0
12	11	22100	330	1
13	12	17800	340	0
14	13	26000	400	1
15	14	23400	380	1
16	15	27600	440	0
17	16	26100	440	0
18	17	24200	430	0
19	18	26400	400	1
20	19	20000	350	1
21	20	24600	410	0
22	21	23900	350	1
23	22	19000	350	0
24	23	19500	390	0
25	24	22100	400	0
26	25	17500	340	0

Regression_Dummy

Fig. 13.1

Step 2: Before choosing the scatter plot, we need to sort the given data on the basis of the promotional activities as shown in Fig. 13.2. For this purpose, we

1. select the entire data of monthly sales, advertisement cost and promotion given in Cells B1:D41,
2. select the ***Sort & Filter*** option from ***Home*** tab as shown in Fig. 13.2a,
3. select ***Custom Sort***,
4. choose promotion X_2 in ***Column Sort by***, and
5. click on ***OK*** as shown in Fig. 13.2b.

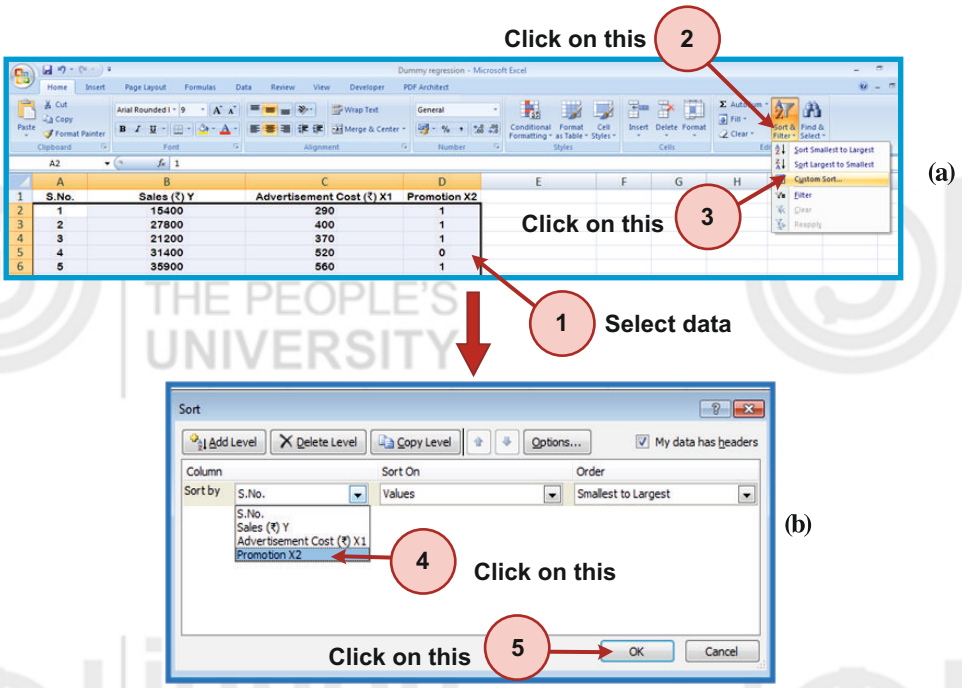


Fig. 13.2

Step 3: After clicking ***OK***, we obtain the sorted data on the basis of promotional activities as shown in Fig. 13.3.

	A	B	C	D
1	S.No.	Sales (₹) Y	Advertisement Cost (₹) X1	Promotion X2
2	4	31400	520	0
3	7	21400	330	0
4	9	11200	270	0
5	10	32100	540	0
6	12	17800	340	0
7	15	27600	440	0
8	16	26100	440	0
9	17	24200	430	0
10	20	24600	410	0
11	22	19000	350	0
12	23	19500	390	0
13	24	22100	400	0
14	25	17500	340	0
15	26	11200	280	0
16	27	12400	280	0
17	28	20700	390	0
18	29	10900	250	0
19	30	18400	360	0
20	37	30200	500	0
21	39	17700	350	0
22	40	11500	210	0
23	1	15400	290	1
24	2	27800	400	1
25	3	21200	370	1
26	5	35900	560	1

Regression_Dummy

Fig. 13.3

Step 4: For plotting the scatter diagram, we select ***Scatter with only Markers*** from ***Insert*** tab without selecting any data point. We get a blank chart as shown in Fig. 13.4. We now click on the chart and select the ***Select Data*** option under ***Design*** tab of the ***Chart Tools*** (Fig. 13.4).

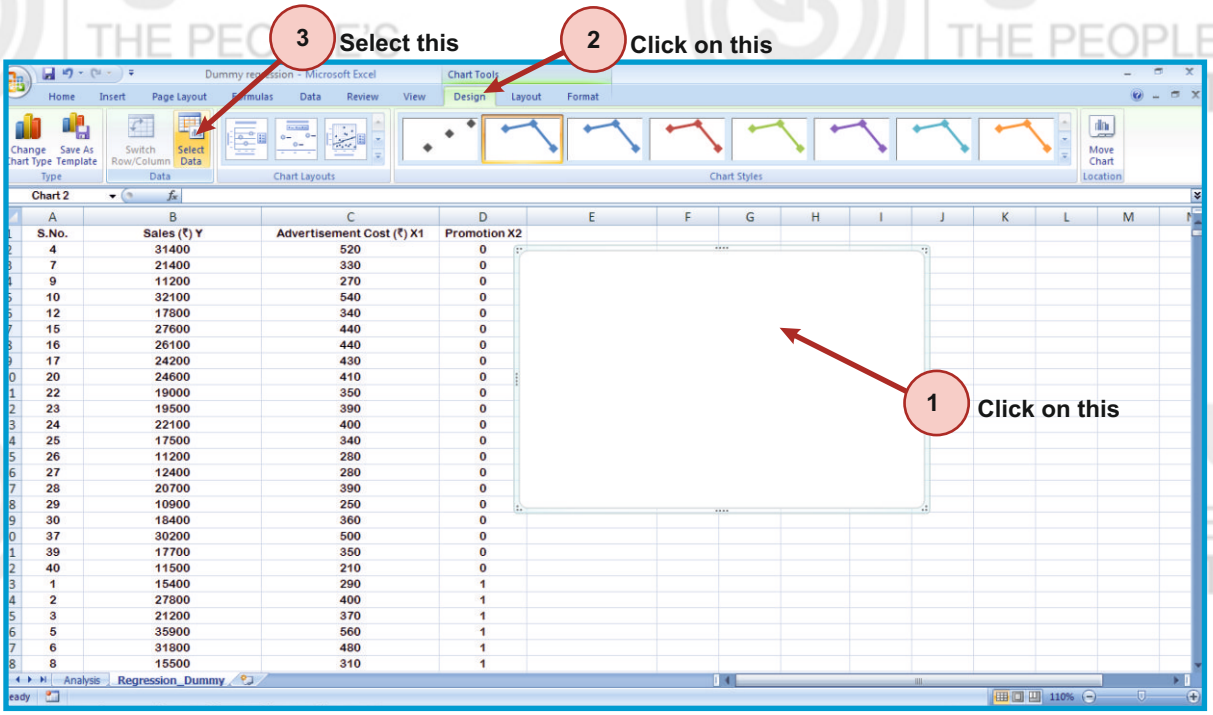


Fig. 13.4

Step 5: We plot the scatter diagram in two parts. We select the data points of the advertisement cost and sales for (i) no promotional activities and (ii) promotional activities, respectively, as separate data series. To add the data series, we

1. click on ***Add*** button shown in Fig. 13.5a and get a new dialog box opened as shown in Fig. 13.5b,
2. select the values of advertisement cost in ***Series X value***, i.e., Cells C2:C22 and sales in ***Series Y value***, i.e., Cells B2:B22 corresponding to the value "0" of the promotional activity. We name the series as "**No Promotional Activity**" as shown in Fig. 13.5b,
3. click on ***OK*** as shown in Fig. 13.5b,
4. select again the ***Add*** button shown in Fig. 13.5c,
5. select the values of the advertisement cost in ***Series X value***, i.e., Cells C23:C41 and sales in ***Series Y value***, i.e., Cells B23:B41 corresponding to the value "1" of the promotional activity. We name the series "**Promotional Activity**" as shown in Fig. 13.5d,
6. click on ***OK*** as shown in Fig. 13.5d, and
7. click on ***OK*** as shown in Fig. 13.5e.

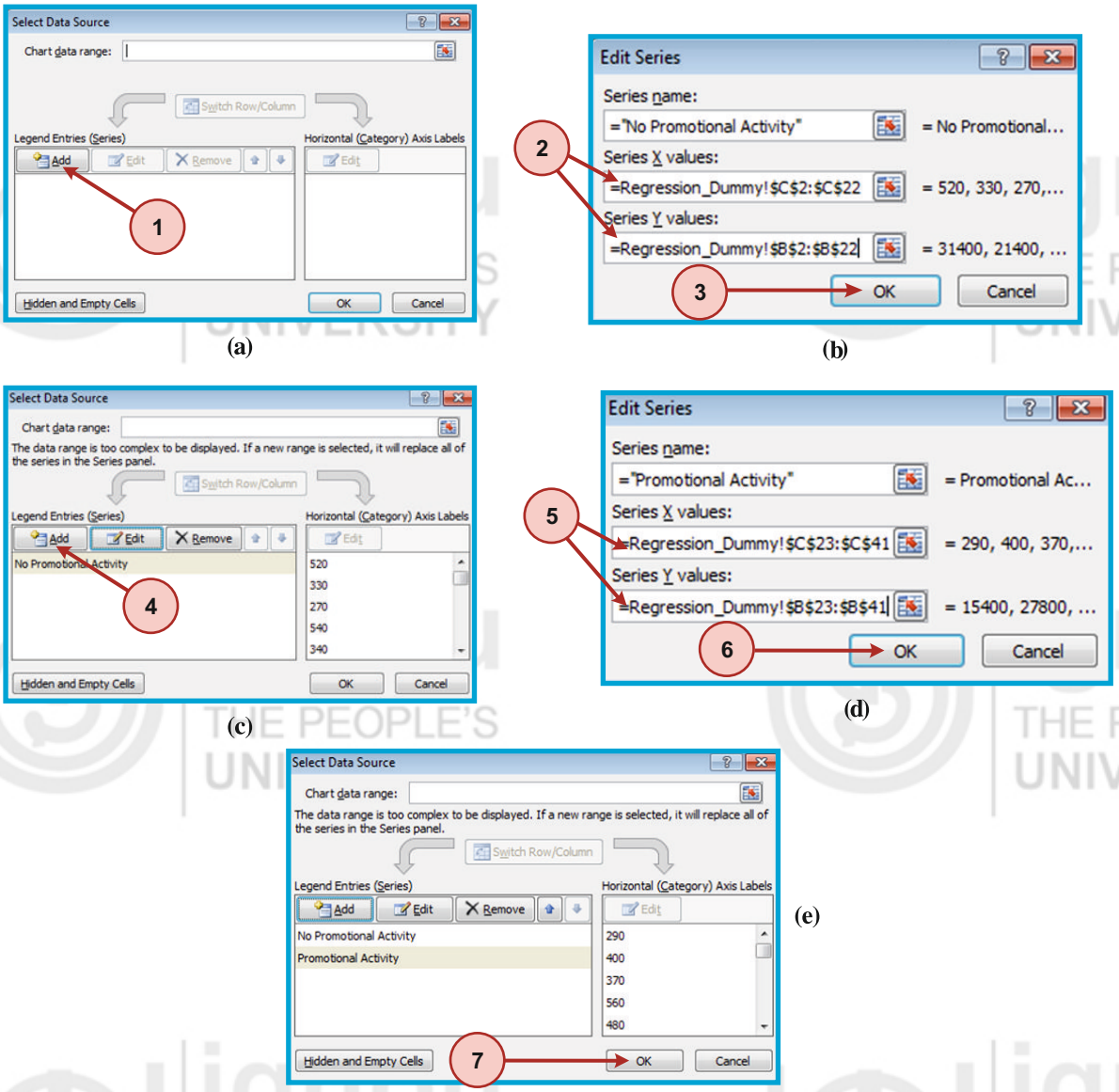


Fig. 13.5

The resulting chart is shown in Fig. 13.6.

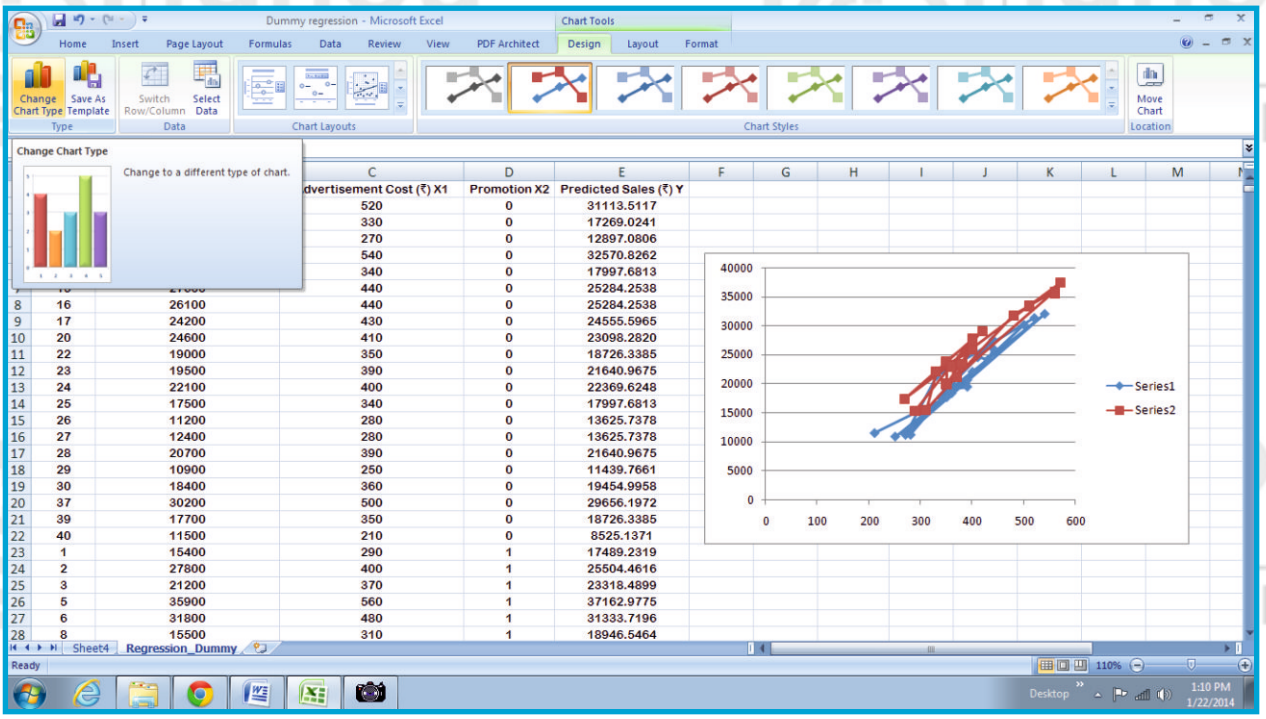


Fig. 13.6

Step 6: We now repeat Steps 5 and 6 of Sec. 12.4 of Lab Session 12 to obtain the required scatter plot. We format the chart by considering green colour markers for the data with promotional activities and red colour markers for no promotional activities as discussed in Step 7 of Sec. 12.4 of Lab Session 12. The resulting scatter plot is shown in Fig. 13.7.

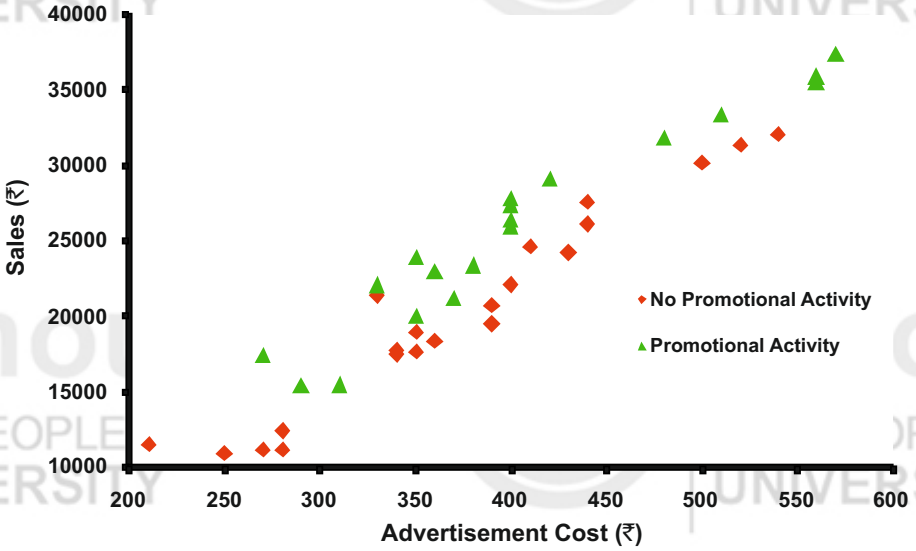


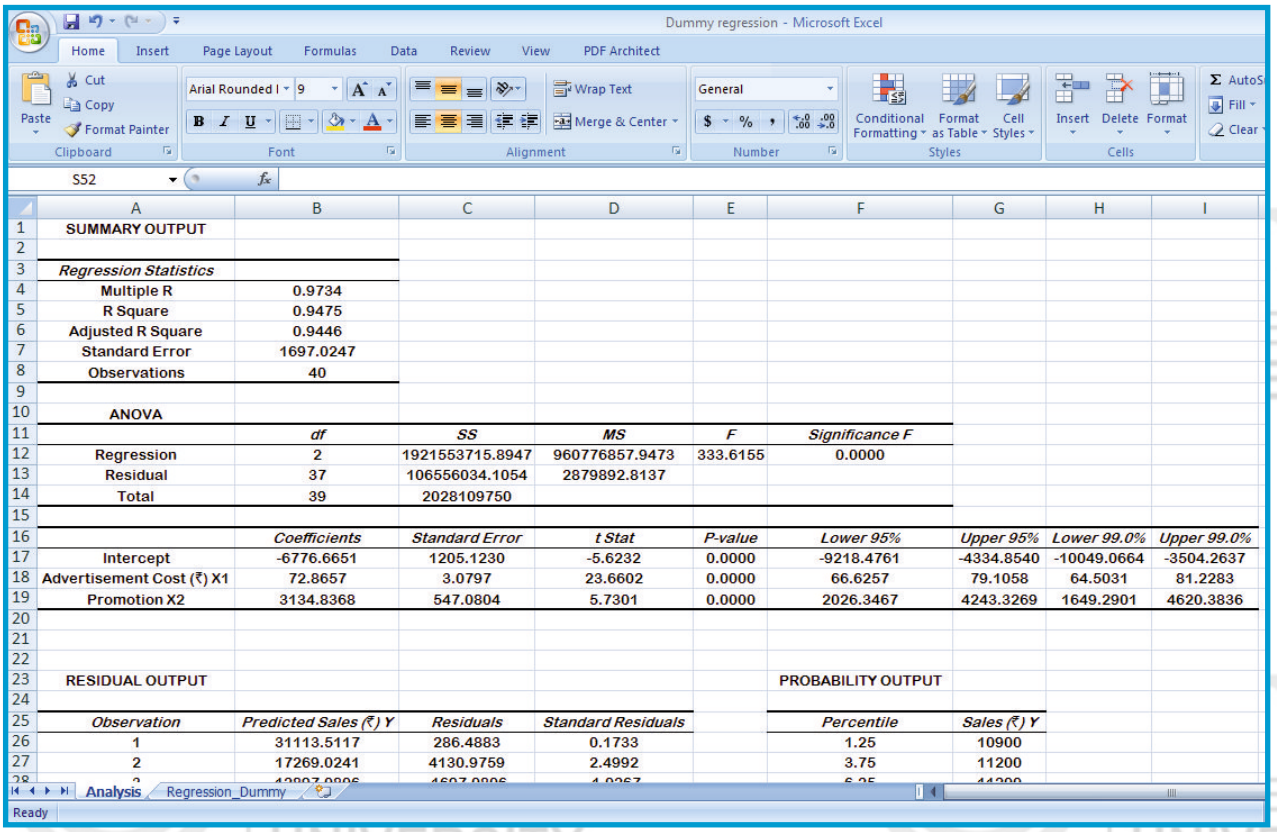
Fig. 13.7

Interpretation

Fig. 13.7 reveals the relationship between monthly sales and advertisement cost for both levels of promotional activity. It shows approximately linear relationship between monthly sales and advertisement cost for both levels of promotional activity. Having obtained an idea of the relationship between the variables, we now carry out the regression analysis for the given data.

13.5 FITTING AND ANALYSIS OF REGRESSION WITH DUMMY VARIABLE IN EXCEL 2007

We fit a multiple linear regression on the given data by choosing ***Data Analysis*** under the ***Data*** tab and subsequently selecting ***Regression*** in the same manner as discussed in Sec. 12.5 of Lab Session 12. The partial output is given in Fig. 13.8.



	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.9734							
5	R Square	0.9475							
6	Adjusted R Square	0.9446							
7	Standard Error	1697.0247							
8	Observations	40							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	2	1921553715.8947	960776857.9473	333.6155	0.0000			
13	Residual	37	106556034.1054	2879892.8137					
14	Total	39	2028109750						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
17	Intercept	-6776.6651	1205.1230	-5.6232	0.0000	-9218.4761	-4334.8540	-10049.0664	-3504.2637
18	Advertisement Cost (₹) X1	72.8657	3.0797	23.6602	0.0000	66.6257	79.1058	64.5031	81.2283
19	Promotion X2	3134.8368	547.0804	5.7301	0.0000	2026.3467	4243.3269	1649.2901	4620.3836
20									
21									
22									
23	RESIDUAL OUTPUT					PROBABILITY OUTPUT			
24									
25	*Observation*	*Predicted Sales (₹) Y*	*Residuals*	*Standard Residuals*		*Percentile*	*Sales (₹) Y*		
26	1	31113.5117	286.4883	0.1733		1.25	10900		
27	2	17269.0241	4130.9759	2.4992		3.75	11200		

Fig. 13.8

Interpretation of the Results of Regression Analysis

The interpretation of the results given in Fig. 13.8 is the same as explained in Lab Sessions 11 and 12. Let us briefly interpret the results:

- The regression coefficients of the fitted model are $\hat{B}_0 = -6776.6651$, $\hat{B}_1 = 72.8657$ and $\hat{B}_2 = 3134.8368$ in Cells B17, B18 and B19, respectively.
- The standard error of e is $\hat{\sigma} = \sqrt{MSS_{Res}} = 1697.0247$ (Cell B7).
- $R^2 = 0.9475$ means that 94.75% of the variation in Y is explained by the regressors, i.e., advertisement cost (X_1) and promotional activity (X_2). We can say that 94.75% of the variability in the monthly sales is accounted for by the regression model.
- From the ANOVA table, the F-statistic is 333.6155 given in Cell E12. The associated p-value is given in Cell F12, i.e., 0.0000. 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 promotional activity (regressors) are statistically significant in predicting the sales at 5% significance level. This implies that there is a linear relationship between sales and the combination of the advertisement cost and promotional activity.

- Standard errors of $\hat{B}_0$, $\hat{B}_1$ and $\hat{B}_2$ given in Cells C17, C18 and C19, respectively, are $SE(\hat{B}_0) = 1205.1230$, $SE(\hat{B}_1) = 3.0797$ and $SE(\hat{B}_2) = 547.0804$.
- The t-statistic for the intercept ($\hat{B}_0$) is -5.6232 (Cell D17) and its p-value is 0.0000 (Cell E17). Since 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.
- The t-statistic for $\hat{B}_1$ is 23.6602 (Cell D18) and its p-value is 0.0000 (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., advertisement cost affects the monthly sales of the juice on the basis of given data.
- The t-statistic for $\hat{B}_2$ is 5.7301 (Cell D19) and its p-value is 0.0000 (Cell E19), which is less than 0.05. 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., promotional activities also affect the monthly sales of the juice for the given data.
- 95% confidence intervals for the regression coefficients $\hat{B}_0$, $\hat{B}_1$ and $\hat{B}_2$ are $(-9218.4761, -4334.8540)$, $(66.6257, 79.1058)$ and $(2026.3467, 4243.3269)$, respectively.
- 99% confidence intervals for the regression coefficients $\hat{B}_0$, $\hat{B}_1$ and $\hat{B}_2$ are $(-10049.0664, -3504.2637)$, $(64.5031, 81.2283)$ and $(1649.2901, 4620.3836)$, respectively.
- We can also draw the residual and normal probability plots in the same way as explained in Lab Sessions 11 and 12.

The resulting plots are shown in Figs. 13.9 and 13.10, respectively.

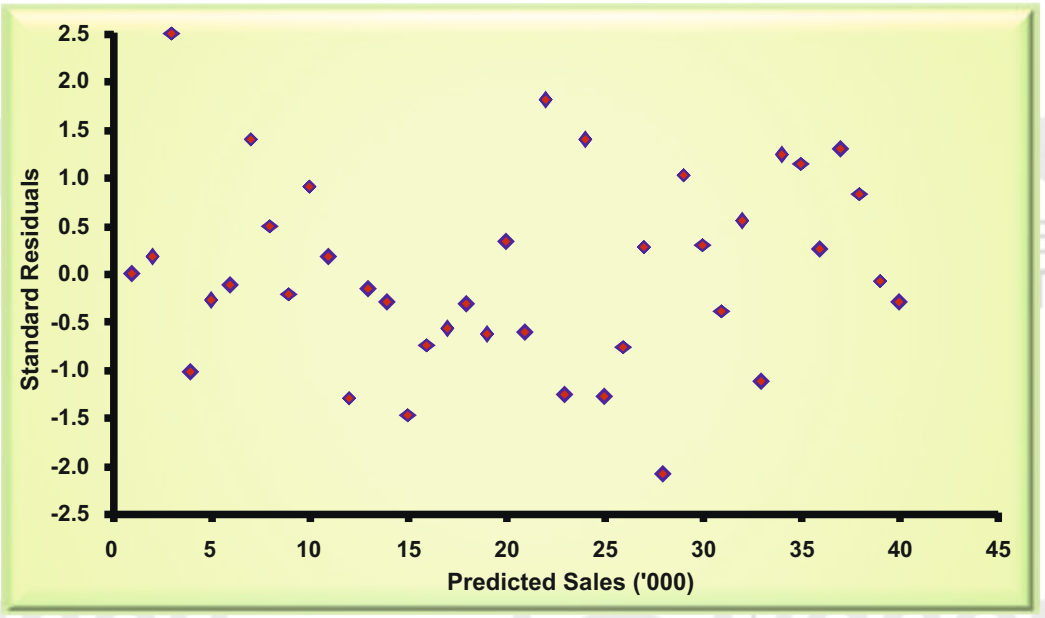


Fig. 13.9: Residual Plot.

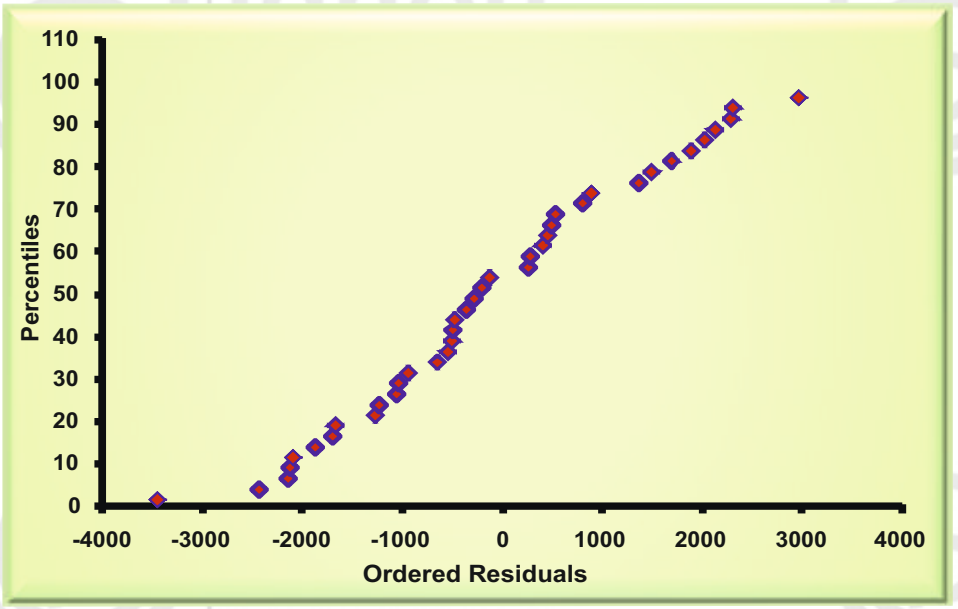


Fig. 13.10: Normal Probability Plot.

Note that Figs. 13.9 and 13.10 confirm the assumptions of linearity and normality, respectively. In the next section, we obtain fitted regression lines.

13.6 FITTED REGRESSION LINES

In Unit 11 in MSTE-002, you have learnt that we use one dummy variable for two categories, two dummy variables for three categories of the qualitative variables, and so on. In this lab session, since the qualitative variable, i.e., the promotional activity has two categories (0 and 1), we use one dummy variable X_2. The fitted regression model is, therefore,

$$\hat{Y} = \hat{B}_0 + \hat{B}_1 X_1 + \hat{B}_2 X_2$$

$$\Rightarrow \quad \hat{Y} = -6776.6651 + 72.8657 X_1 + 3134.8368 X_2$$

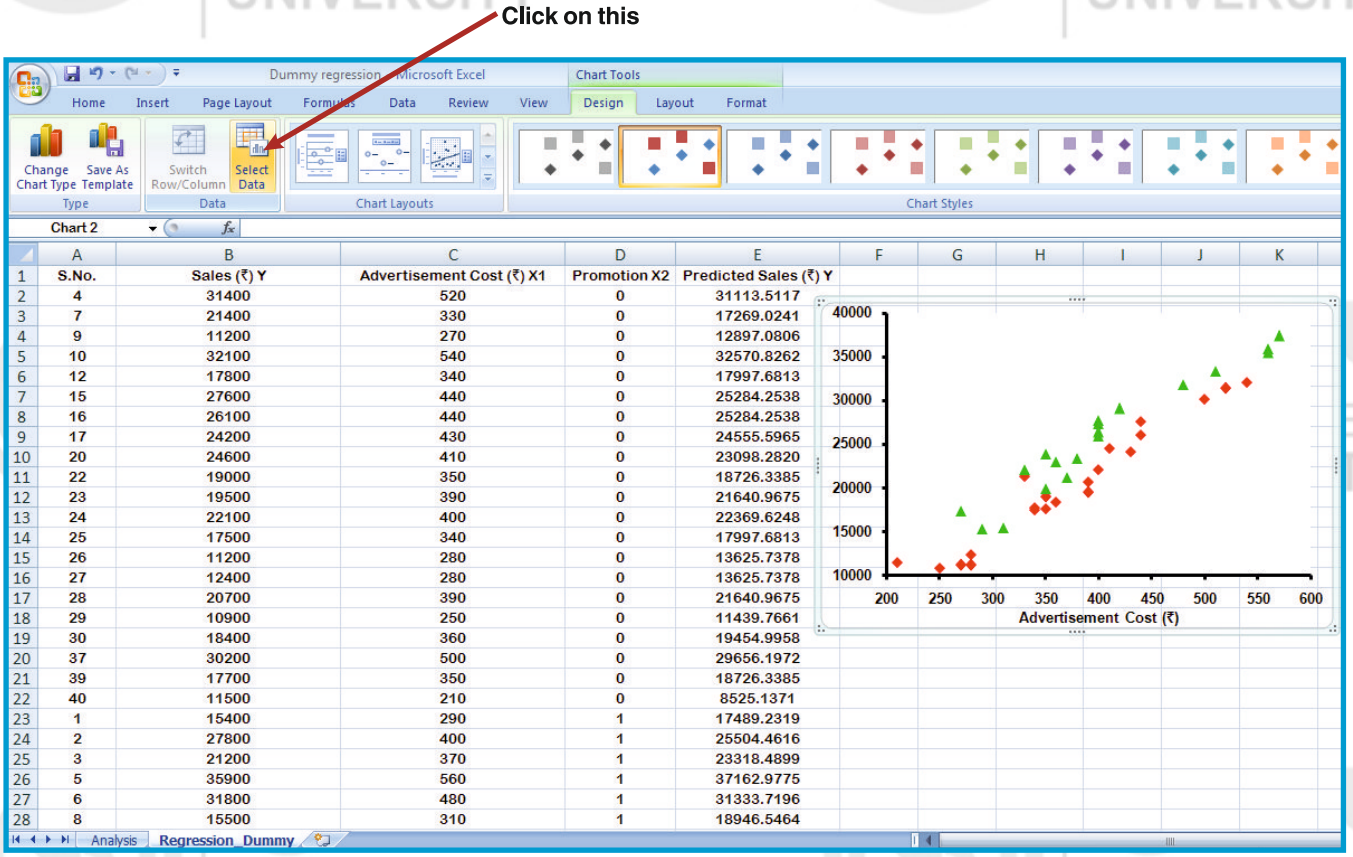
This model represents two different regression equations for different categories of promotional activities as given below:

Promotional Activity	Monthly Sales (₹)
0	$\hat{Y} = \hat{B}_0 + \hat{B}_1 X_1 = -6776.6651 + 72.8657\ X_1$
1	$\hat{Y} = (\hat{B}_0 + \hat{B}_2) + \hat{B}_1 X_1 = -3641.8283 + 72.8657\ X_1$

The steps involved in plotting fitted regression lines are given below:

Step 1: To plot the fitted regression lines on the scatter plot given in Fig. 13.7, we copy the values of the **predicted sales** given in Cells B25:B64 of the sheet named "**Analysis**" (Fig. 13.8) and paste in Cells E1:E41 in the sheet named "**Regression_Dummy**" (Fig. 13.11).

We now select the chart plotted on Excel sheet and click on ***Select Data*** option as shown in Fig. 13.11. It opens a new dialog box as shown in Fig. 13.12a.



	A	B	C	D	E
1	S.No.	Sales (₹) Y	Advertisement Cost (₹) X1	Promotion X2	Predicted Sales (₹) Y
2	4	31400	520	0	31113.5117
3	7	21400	330	0	17269.0241
4	9	11200	270	0	12897.0806
5	10	32100	540	0	32570.8262
6	12	17800	340	0	17997.6813
7	15	27600	440	0	25284.2538
8	16	26100	440	0	25284.2538
9	17	24200	430	0	24555.5965
10	20	24600	410	0	23098.2820
11	22	19000	350	0	18726.3385
12	23	19500	390	0	21640.9675
13	24	22100	400	0	22369.6248
14	25	17500	340	0	17997.6813
15	26	11200	280	0	13625.7378
16	27	12400	280	0	13625.7378
17	28	20700	390	0	21640.9675
18	29	10900	250	0	11439.7661
19	30	18400	360	0	19454.9958
20	37	30200	500	0	29656.1972
21	39	17700	350	0	18726.3385
22	40	11500	210	0	8525.1371
23	1	15400	290	1	17489.2319
24	2	27800	400	1	25504.4616
25	3	21200	370	1	23318.4899
26	5	35900	560	1	37162.9775
27	6	31800	480	1	31333.7196
28	8	15500	310	1	18946.5464

Fig. 13.11

Step 2: To draw the regression lines on the scatter plot, we

1. click on ***Add*** button shown in Fig. 13.12a so that a new dialog box opens as shown in Fig. 13.12b,
2. add values of the advertisement cost in ***Series X value***, i.e., Cells C2:C22 and the predicted sales in ***Series Y value***, i.e., Cell E2:E22 corresponding to "0" value of promotional activity. We name the series as "**Fitted Line for No Promotional Activity**" as shown in Fig. 13.12b,
3. click on ***OK*** as shown in Fig. 13.12b,
4. select again the ***Add*** button shown in Fig. 13.12c,
5. select the values of the advertisement cost in ***Series X value***, i.e., Cells C23:C41 and the predicted sales in ***Series Y value***, i.e., Cells E23:E41 corresponding to "1" value of the promotional activity. We name the series as "**Fitted Line for Promotional Activity**" as shown in Fig. 13.12d,
6. click on ***OK*** as shown in Fig. 13.12d, and
7. click on ***OK*** as shown in Fig. 13.12e.

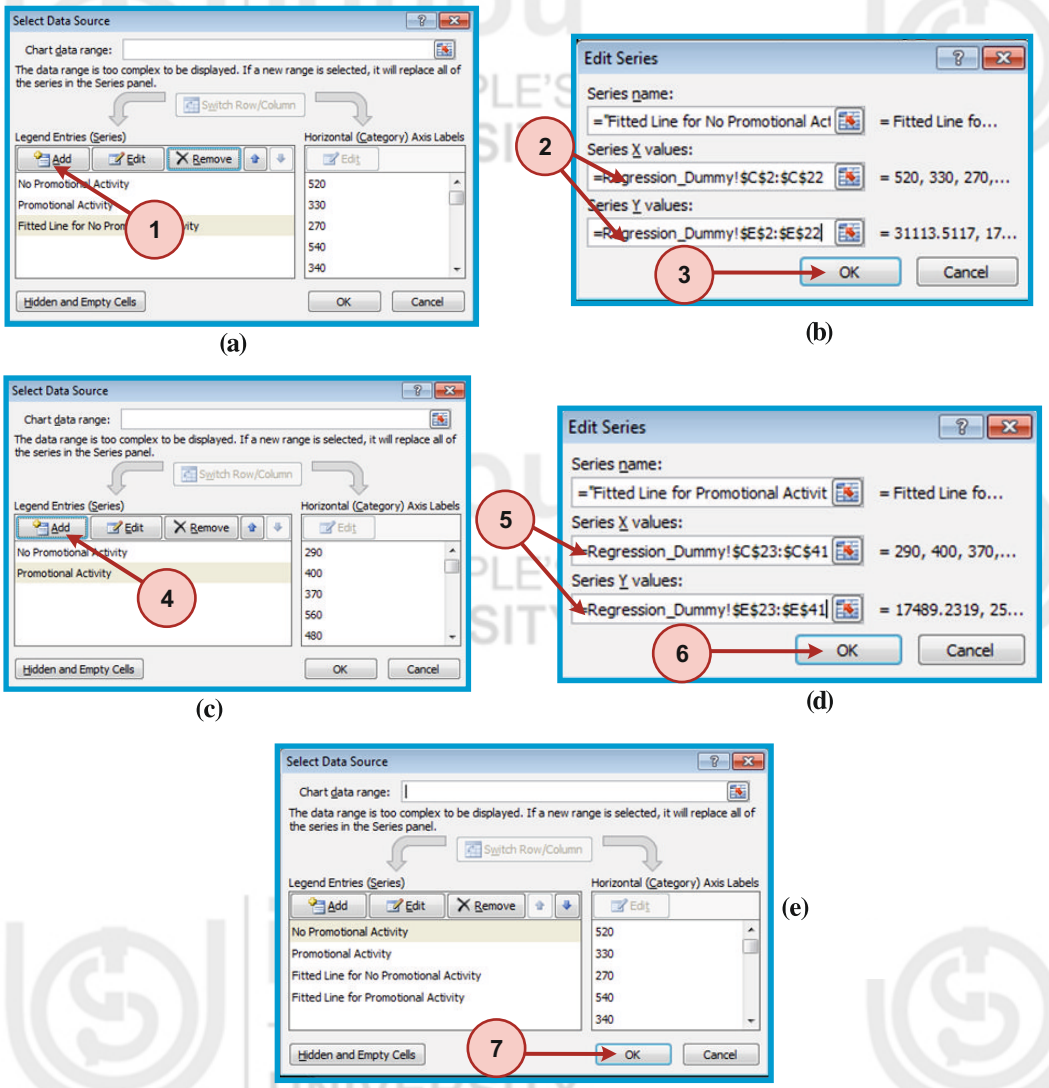


Fig. 13.12

The resulting chart is shown in Fig. 13.13.

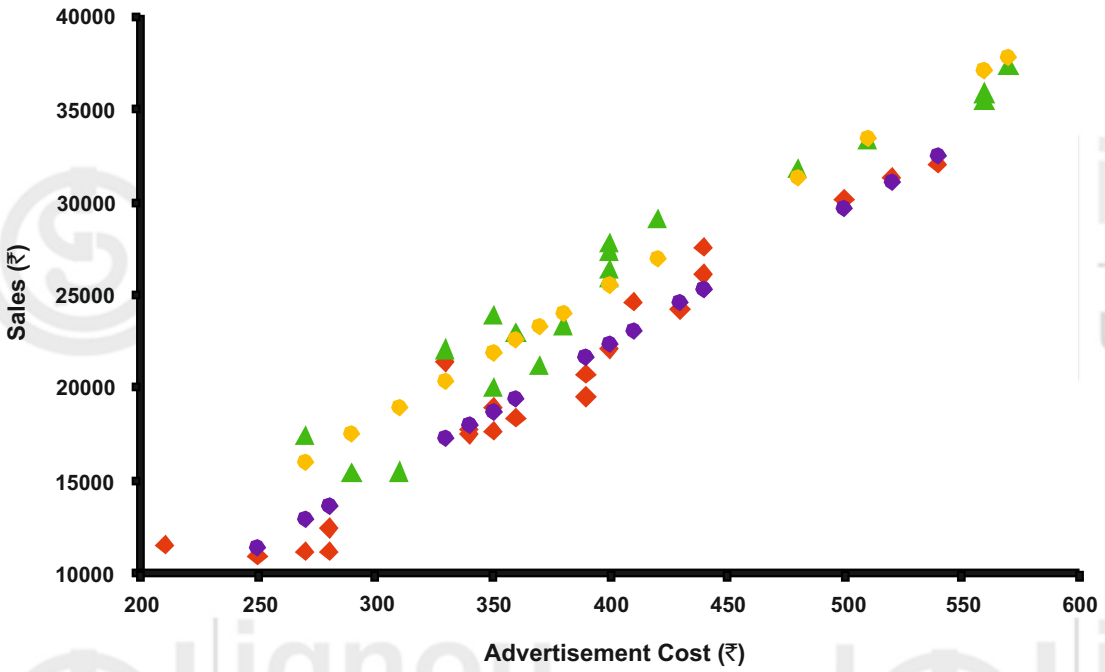
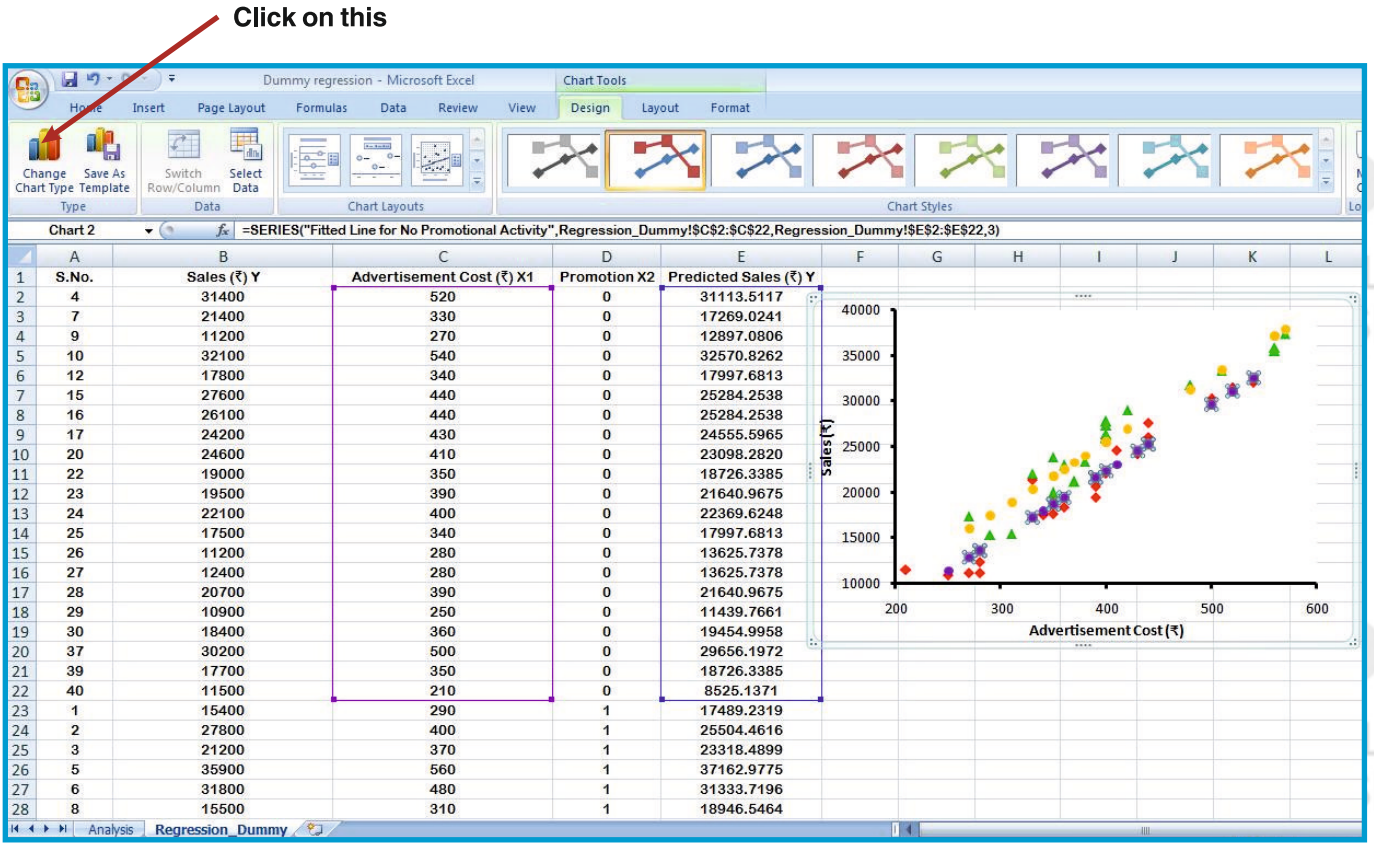


Fig. 13.13

Step 3: By default, the scatter chart shown in Fig. 13.13 shows all series in a **scattered** form **with only markers**. The data points for fitted regression lines are shown with purple and yellow coloured markers for no promotional activities (0) and promotional activity (1), respectively, as shown in Fig. 13.13. For changing the purple coloured data points in the form of the regression line, we select the purple coloured points and select ***Change Chart Type*** as shown in Fig. 13.14.



S.No.	Sales (₹) Y	Advertisement Cost (₹) X1	Promotion X2	Predicted Sales (₹) Y
4	31400	520	0	31113.5117
7	21400	330	0	17269.0241
9	11200	270	0	12897.0806
10	32100	540	0	32570.8262
12	17800	340	0	17997.6813
15	27600	440	0	25284.2538
16	26100	440	0	25284.2538
17	24200	430	0	24555.5965
20	24600	410	0	23098.2820
22	19000	350	0	18726.3385
23	19500	390	0	21640.9675
24	22100	400	0	22369.6248
25	17500	340	0	17997.6813
26	11200	280	0	13625.7378
27	12400	280	0	13625.7378
28	20700	390	0	21640.9675
29	10900	250	0	11439.7661
30	18400	360	0	19454.9958
37	30200	500	0	29656.1972
39	17700	350	0	18726.3385
40	11500	210	0	8525.1371
1	15400	290	1	17489.2319
2	27800	400	1	25504.4616
3	21200	370	1	23318.4899
5	35900	560	1	37162.9775
6	31800	480	1	31333.7196
8	15500	310	1	18946.5464

Fig. 13.14

Step 4: We select ***Scatter with Straight Lines*** and click on ***OK*** and repeat the same procedure for the yellow coloured points.

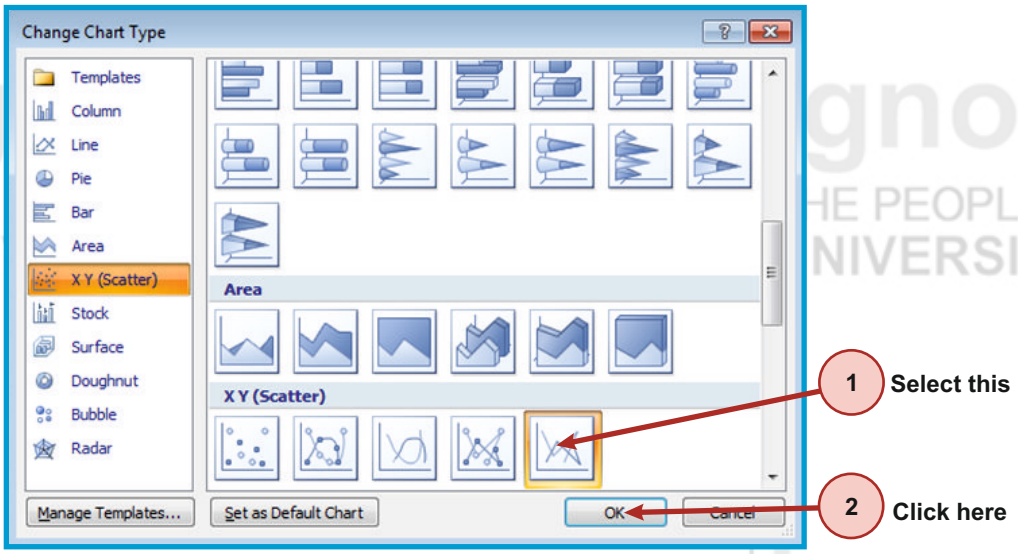


Fig. 13.15

Step 5: We can also format the fitted lines in any way. Here, we have used **dashed red** and **green** coloured lines, respectively, for the "0" and "1" values of the promotional activity with weight **1.5 pt**.

The scatter plots in the case of two different values of a dummy variable with both fitted regression lines are shown in Fig. 13.16.

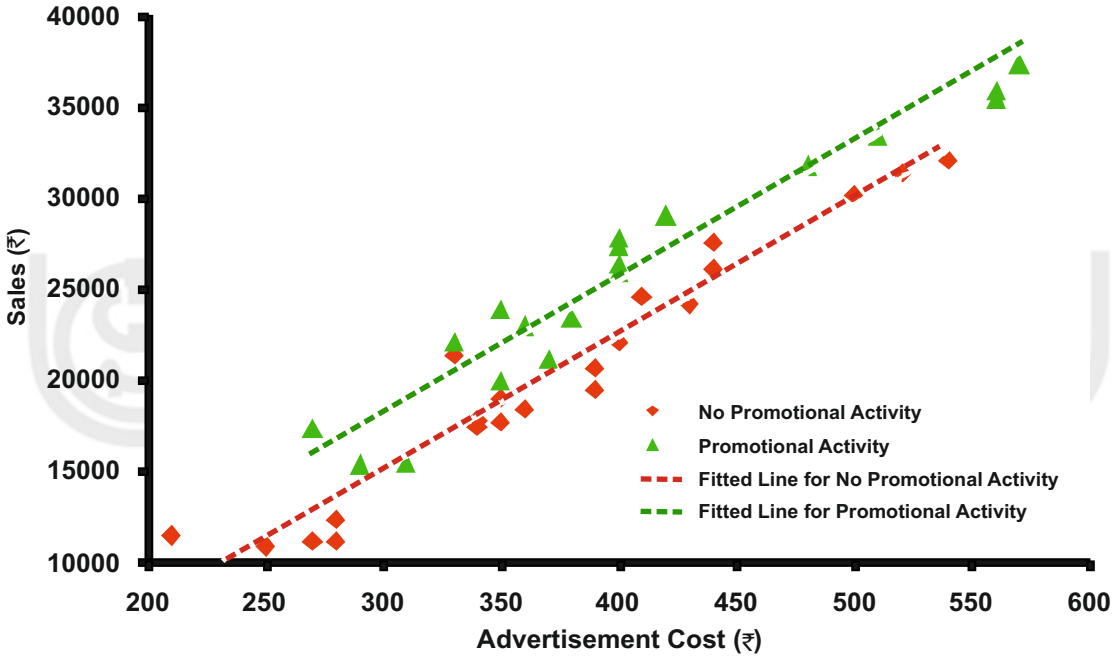


Fig. 13.16

You should now do the following activities for practice.

Activity

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

A1) Example 5 given in Unit 11 of MSTE-002.

A2) Exercise E3 given in Unit 11 of MSTE-002.

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

Continuous Assessment 13

Suppose we are interested in developing a linear model for the electricity consumption of a household 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 (0 for no AC and 1 for having AC) in Table 2.

Table 2: Electricity consumption data

S.No.	Unit (in kWh)	Area (in sq ft)	AC (in hours)	S.No.	Unit (in kWh)	Area (in sq ft)	AC (in hours)
1	513	725	1	21	796	875	1
2	926	1000	1	22	633	875	0
3	706	925	1	23	650	975	0
4	1046	1300	0	24	736	1000	0
5	1196	1400	1	25	583	850	0
6	1060	1200	1	26	373	700	0
7	713	825	0	27	413	700	0
8	516	775	1	28	690	975	0
9	373	675	0	29	363	625	0
10	1070	1350	0	30	613	900	0
11	736	825	1	31	913	1000	1
12	593	850	0	32	766	900	1
13	866	1000	1	33	970	1050	1
14	780	950	1	34	580	675	1
15	920	1100	0	35	1113	1275	1
16	870	1100	0	36	1246	1425	1
17	806	1075	0	37	1006	1250	0
18	880	1000	1	38	1183	1400	1
19	666	875	1	39	590	875	0
20	820	1025	0	40	383	525	0

For this data,

- Prepare a scatter plot to get an idea about the relationship among the variables and develop a linear regression model and its related analysis at 2% level of significance.
- Check the linearity and normality assumptions for the regression analysis.
- Draw both fitted regression lines on the scatter plot.

Home Work: Do It Yourself

1) Follow the steps explained in Secs. 13.4, 13.5 and 13.6 to comprehend the regression analysis with dummy variable for the data of Table 1. Use a different format for the scatter, residual and normal probability plots. Take their screenshots and keep them in your record book.

2) Develop the spreadsheets for the exercise "Continuous Assessment 13" 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.

Variable Selection Methods

SESSION 14

Structure

14.1 Introduction
Objectives
14.2 Problem Description
14.3 Forward Selection Method
14.4 Backward Elimination Method
14.5 Stepwise Regression Method

14.1 INTRODUCTION

In Lab Sessions 11, 12 and 13, you have learnt how to fit linear regression models for one and more than one predictors and the related inferential analysis. In real life, we may come across a long list of regressors available in the data which may be related to the response variable. In this situation, it is difficult to choose some important regressors, which can really contribute to the model. The basic screening may be done on the basis of prior experience, but some standard methods known as **variable selection methods** are also used to decide the appropriate regressors to be entered in the model.

Prerequisite

- Lab Session 12 of MSTL-002 (Industrial Statistics Lab).
- Unit 12 of MSTE-002 (Industrial Statistics-II).

You have learnt in Unit 12 of MSTE-002 (Industrial Statistics-II) that three methods are commonly used for variable selection: (i) forward selection, (ii) backward elimination and (iii) stepwise regression. These methods of variable selection are available in most statistical software packages. In this lab session, our aim is to train you in obtaining the best regression model using these three methods of variable selection in Excel 2007.

Objectives

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

- prepare the spreadsheet for the variable selection methods with MS Excel 2007;
- test the significance of the regression parameters;
- choose the appropriate best fitted regressors for the regression model; and
- interpret the results of the forward selection, backward elimination and stepwise methods.

14.2 PROBLEM DESCRIPTION

For applying the variable selection methods, we consider the data of a juice manufacturing company. In Lab Session 12, we fitted a multiple regression model for the monthly sales as a linear function of advertisement cost and price. Here

we have also considered two other variables: temperature and number of stores. The data are recorded in Table 1.

Table 1: Data of a juice manufacturing company

S. No.	Monthly Sales	Advertisement Cost	Price	Temperature	Number of Stores
1	212000	3240	85	24	14
2	230000	3550	94	28	15
3	273000	3890	77	33	16
4	255000	3720	84	31	16
5	285000	4030	75	35	17
6	262000	3790	81	32	16
7	273000	3950	76	34	16
8	215000	3290	89	24	15
9	185000	2920	79	21	13
10	268000	3850	78	33	16
11	285000	4030	75	35	17
12	230000	3550	94	28	15
13	212000	3240	85	24	14
14	309000	4200	68	39	17
15	228000	3470	94	27	15
16	215000	3290	89	24	15
17	324000	4340	64	42	18
18	210000	3330	94	24	14
19	262000	3790	81	32	16
20	329000	4400	62	44	18
21	313000	4240	67	40	17
22	243000	3660	88	30	16
23	255000	3720	84	31	16
24	293000	4080	73	36	17
25	216000	3380	94	25	15
26	195000	2990	83	22	14
27	208000	3140	82	24	14
28	268000	3850	78	33	16
29	173000	2860	78	20	13
30	235000	3600	92	29	15
31	216000	3380	94	25	15
32	300000	4130	71	37	17
33	235000	3600	92	29	15
34	210000	3330	94	24	14
35	273000	3890	77	33	16
36	293000	4080	73	36	17
37	243000	3660	88	30	16
38	273000	3950	76	34	16
39	228000	3470	94	27	15
40	201000	3060	83	23	14

We can also decide on different values for α_{IN} and α_{OUT} as these show the level of significance for entering and eliminating a regressor, respectively. For example, we can use $\alpha_{IN} = 0.05$ and $\alpha_{OUT} = 0.03$.

We now build appropriate linear regression models for the monthly sales using (i) forward selection, (ii) backward elimination and (iii) stepwise approaches to determine the most appropriate regressors at 5% level of significance, i.e., $\alpha_{IN} = \alpha_{OUT} = 0.05$.

14.3 FORWARD SELECTION METHOD

You have already studied the forward selection method in Unit 12 of MSTE-002. Therefore, before describing how to apply the forward selection procedure in Excel, we provide a brief overview of the steps involved in this method. First, we assume a model without any regressor. Then we enter one regressor at each step on the basis of partial F-test into the model. We stop when no more regressors can be significantly entered in the model. In this way, we get a **final model**. This process consists of the following steps:

If we add an important regressor (X) to the model, it increases the regression sum of square and decreases the residual sum of squares, which result in increasing the F-value. On the other hand, adding an unimportant (unnecessary) regressor may decrease the regression sum of squares and increase the residual sum of squares, which results in decreasing the F value.

Step 1: Before starting the procedure for forward selection, we decide the level of significance to enter the regressor in the model, i.e., **Alfa-to-In** (α_{IN}) to compare the p-values calculated at each stage of selection.

Step 2: If Y is a dependent variable, we start by assuming a model without any regressor, i.e.,

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

where B_0 is the intercept and e is a normally distributed random error component with mean zero and variance σ^2.

Step 3: We perform the regression analysis for all possible one-regressor models and decide the entering variable in the model. Then we identify the regressor having the largest absolute t value of the coefficient (or partial F value) and also calculate its p-value. If its p-value is significant, the corresponding regressor will enter in the model. We can also find the first variable to be entered in the model by calculating the correlation coefficient (r) for each pair of the dependent and independent variables (X, Y). We check the significance of the model for the pair having the highest correlation coefficient.

Step 4: We now perform the regression analysis for all possible two-regressor models containing one regressor determined from Step 3 and the second from the remaining regressors. We compute the p-value for the coefficients of the remaining regressors and include the most significant regressor in the model. In this way, we get 2 regressors in the model.

For example, if we add the regressor x_q in the model, which already contains the regressor x_p, the extra sum of squares measures the increase in the regression sum of squares due to the addition of the one more regressor. For example, the extra sum of squares due to x_q is

$$SS_{Reg}\left(b_q / b_p\right) = SS_{Reg}(b_p, b_q) - SS_{Reg}\left(b_p\right)$$

We can also define partial F-statistic to check the effect of x_q after eliminating the effect of x_p as follows:

$$F\left(b_q / b_p\right) = \frac{SS_{Reg}\left(b_q / b_p\right)}{SS_{Res}} \sim F_{(\alpha),\,[1,(n-p)]}$$

Step 5: Similarly, we carry out the regression analysis for all possible three-regressor models containing two regressors determined from Steps 3 and 4 and one from the remaining regressors. We select the most significant regressor from the remaining regressors.

For example, if we add one more regressor x_r in the model which already contains the regressors x_p and x_q then the extra sum of squares due to x_r is

$$SS_{Reg}\left(b_r / b_p, b_q\right) = SS_{Reg}\left(b_p, b_q, b_r\right) - SS_{Reg}\left(b_p, b_q\right)$$

and $$F\left(b_r / b_p, b_q\right) = \frac{SS_{Reg}\left(b_r / b_p, b_q\right)/2}{SS_{Res}} \sim F_{(\alpha),[2,(n-p)]}$$

Step 6: We repeat the procedure till all regressors are included in the model or no additional regressor gives a significant p-value. In the same manner, we can compute the extra sum of squares for every addition of a regressor.

Steps in Excel

Before starting the procedure, the first thing we need to do is to set a significance level for deciding a regressor to be entered into the model. We call this significance level **Alpha-to-In** and denote it as α_{IN}. In this session, we are considering $\alpha_{IN} = 0.05$. If the p-value is less than or equal to α_{IN}, we consider the corresponding regressor as significant, otherwise it is insignificant.

Note that the p-value associated with partial F-statistic is the same as the p-value associated with the t-statistic of the corresponding regressor. For the data given in Sec. 14.2, we denote the monthly sales, advertisement cost, price, temperature and number of stores by Y, X_1, X_2, X_3 and X_4, respectively.

The steps involved in the forward selection method to select the appropriate regressors in the model using Excel 2007 are given below:

This partial F test is equivalent to the t-test because it involves the testing of only the single regressor. We know that for single variable

$$F_{cal} = t^2_{cal}$$

Step 1: We enter the given data in Excel 2007 spreadsheet as shown in Fig. 14.1.

	A	B	C	D	E	F
1	S.No.	Y	X1	X2	X3	X4
2	1	212000	3240	85	24	14
3	2	230000	3550	94	28	15
4	3	273000	3890	77	33	16
5	4	255000	3720	84	31	16
6	5	285000	4030	75	35	17
7	6	262000	3790	81	32	16
8	7	273000	3950	76	34	16
9	8	215000	3290	89	24	15
10	9	185000	2920	79	21	13
11	10	268000	3850	78	33	16
12	11	285000	4030	75	35	17
13	12	230000	3550	94	28	15
14	13	212000	3240	85	24	14
15	14	309000	4200	68	39	17
16	15	228000	3470	94	27	15
17	16	215000	3290	89	24	15
18	17	324000	4340	64	42	18
19	18	210000	3330	94	24	14
20	19	262000	3790	81	32	16
21	20	329000	4400	62	44	18

Forward Selection

Fig. 14.1: Partial screenshot of the spreadsheet for the given data.

Step 2: To select the first regressor in the model

We perform the regression analysis for each regressor by considering only one regressor at a time as discussed in Step 4 of Sec. 11.4 in Lab Session 11. In variable selection methods, we only need the p-value related to the regression coefficients. So we specify only the following options as shown in Fig. 14.2:

- ✓ ***Input Y Range***.
- ✓ ***Input X Range***.
- ✓ Tick the ***Labels***.
- ✓ Specify the ***Output Range***.

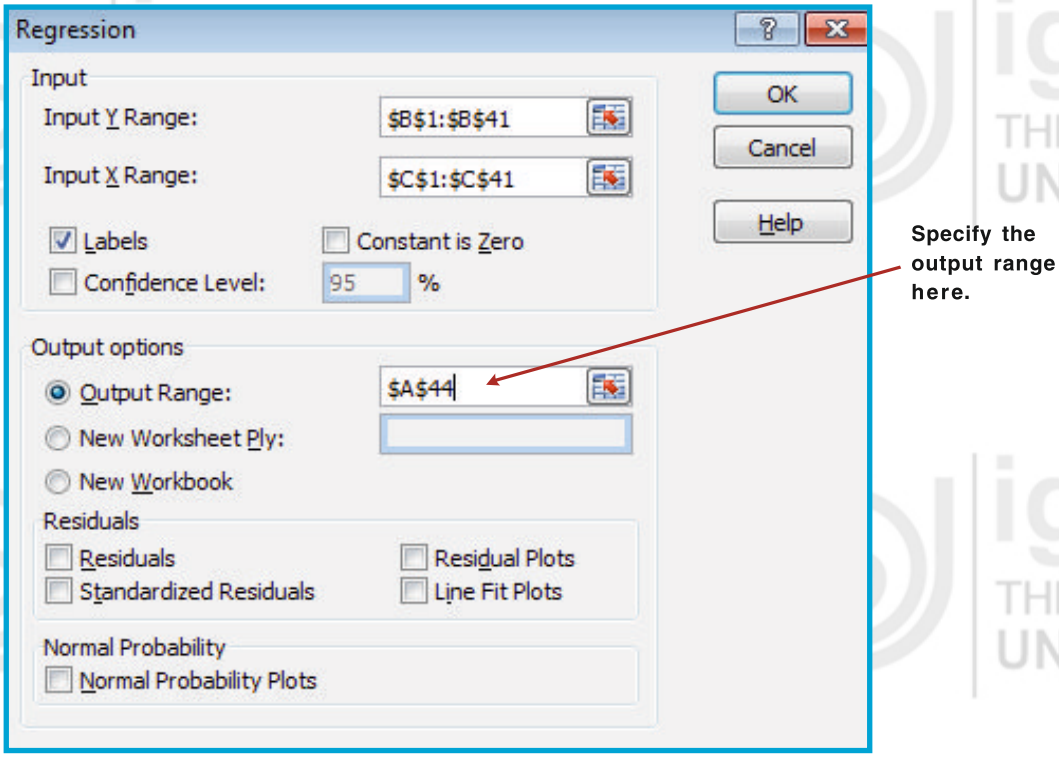


Fig. 14.2

We obtain the results after clicking ***OK***. We perform the regression analysis for X_1, X_2, X_3 and X_4 one at a time as given below:

For X_1: We consider **Y range** as Cells B1:B41, **X range** as Cells C1:C41 and **output range** as Cell A44 as shown in Fig. 14.2. The result is shown in Fig. 14.3.

	A	B	C	D	E	F	G
53	ANOVA						
54		*df*	*SS*	*MS*	*F*	*Significance F*	
55	Regression	1.0000	60035464734.5936	60035464734.5936	1783.2638	0.0000	
56	Residual	38.0000	1279310265.4064	33666059.6160			
57	Total	39.0000	61314775000.0000				
58							
59		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
60	Intercept	-109571.1015	8524.6940	-12.8534	0.0000	-126828.4421	-92313.7609
61	X1	98.0940	2.3229	42.2287	0.0000	93.3915	102.7966
62							

Fig. 14.3

For X_2: We consider **Y range** as Cells B1:B41, **X range** as Cells D1:D41 and **output range** as Cell K44. The result is shown in Fig. 14.4.

	K	L	M	N	O	P	Q
53	ANOVA						
54		*df*	*SS*	*MS*	*F*	*Significance F*	
55	Regression	1.0000	34697762892.9791	34697762892.9791	49.5366	0.0000	
56	Residual	38.0000	26617012107.0209	700447687.0269			
57	Total	39.0000	61314775000.0000				
58							
59		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
60	Intercept	516401.6045	38317.8738	13.4768	0.0000	438831.1251	593972.0840
61	X2	-3264.2509	463.7892	-7.0382	0.0000	-4203.1431	-2325.3587
62							

Fig. 14.4

For X_3: We consider **Y range** as Cells B1:B41, **X range** as Cells E1:E41 and **output range** as Cell U44. The result is shown in Fig. 14.5.

	U	V	W	X	Y	Z	AA
53	ANOVA						
54		*df*	*SS*	*MS*	*F*	*Significance F*	
55	Regression	1.0000	60594399255.6855	60594399255.6855	3196.3697	0.0000	
56	Residual	38.0000	720375744.3145	18957256.4293			
57	Total	39.0000	61314775000.0000				
58							
59		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
60	Intercept	50197.4317	3571.4011	14.0554	0.0000	42967.5082	57427.3551
61	X3	6593.2635	116.6197	56.5364	0.0000	6357.1793	6829.3478
62							

Fig. 14.5

For X_4: We consider **Y range** as Cells B1:B41, **X range** as Cells F1:F41 and **output range** as Cell AE44. The result is given in Fig. 14.6.

	AE	AF	AG	AH	AI	AJ	AK
53	ANOVA						
54		*df*	*SS*	*MS*	*F*	*Significance F*	
55	Regression	1.0000	57283024294.0702	57283024294.0702	539.9032	0.0000	
56	Residual	38.0000	4031750705.9298	106098702.7876			
57	Total	39.0000	61314775000.0000				
58							
59		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
60	Intercept	-223668.8181	20378.3862	-10.9758	0.0000	-264922.7039	-182414.9322
61	X4	30402.1783	1308.4188	23.2358	0.0000	27753.4230	33050.9336
62							

Fig. 14.6

- The first regressor to be entered in the model is temperature (X_3).
- The lowest p-value of t test = 0.0000.
- From Fig. 14.5, the model is:

  $\hat{Y} = 50197.432 + 6593.264X_3$
- F test p-value = 0.0000.

Interpretation

We now compare Figs. 14.3 to 14.6 and note from Fig. 14.5 that the highest absolute value of t-statistic is 56.5364 given in Cell X61 (Fig. 14.5), which is associated with regressor X_3. The p-value of t-statistic associated with the X_3 is 0.0000 given in Cell Y61, which is less than $\alpha_{IN} = 0.05$. We can conclude that the regressor X_3 is contributing significantly to the model. We can include the regressor X_3 (temperature) into the model.

Step 3: To select the second regressor in the model

The variable X_3 is already entered in the model as shown in Step 2. We now repeat Step 2 for all possible 2 regressors in the models that include X_3 as one of the regressors, i.e., (X_3, X_1),

(X_3, X_2) and (X_3, X_4) to find the second regressor to be entered in the model. As we have already explained in Lab Session 12, we need to put independent variables in the adjacent columns in Excel 2007 to run regression analysis for more than one independent variable. So we copy and paste the values of (X_3, X_1), (X_3, X_2) and (X_3, X_4) in Cells H1:I41, K1:L41, and N1:O41, respectively, as shown in Fig. 14.7.

	H	I	J	K	L	M	N	O
1	X3	X1		X3	X2		X3	X4
2	24	3240		24	85		24	14
3	28	3550		28	94		28	15
4	33	3890		33	77		33	16
5	31	3720		31	84		31	16
6	35	4030		35	75		35	17
7	32	3790		32	81		32	16
8	34	3950		34	76		34	16
9	24	3290		24	89		24	15
10	21	2920		21	79		21	13
11	33	3850		33	78		33	16
12	35	4030		35	75		35	17

Forward Selection

Fig. 14.7

To find the second variable to be entered in the model, we perform the regression analysis for these combinations of the two regressors in the model one at a time as discussed in Step 2.

For X_3 and X_1: We consider **Y range** as Cells B1:B41, **X range** as Cells H1:I41 and **output range** as Cell A64. The result is shown in Fig. 14.8.

	A	B	C	D	E	F	G
73	ANOVA						
74		*df*	*SS*	*MS*	*F*	*Significance F*	
75	Regression	2.0000	60773651603.4472	30386825801.7236	2077.7378	0.0000	
76	Residual	37.0000	541123396.5528	14624956.6636			
77	Total	39.0000	61314775000.0000				
78							
79		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
80	Intercept	-4221.1064	15857.3263	-0.2662	0.7916	-36351.1012	27908.8883
81	X3	4436.6174	624.4765	7.1045	0.0000	3171.3078	5701.9270
82	X1	32.6780	9.3341	3.5009	0.0012	13.7654	51.5906
83							

Fig. 14.8

For X_3 and X_2: We consider **Y range** as Cells B1:B41, **X range** as Cells K1:L41 and **output range** as Cell K64. The result is shown in Fig. 14.9.

	K	L	M	N	O	P	Q
73	ANOVA						
74		*df*	*SS*	*MS*	*F*	*Significance F*	
75	Regression	2.0000	60598074331.3543	30299037165.6772	1564.2017	0.0000	
76	Residual	37.0000	716700668.6457	19370288.3418			
77	Total	39.0000	61314775000.0000				
78							
79		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
80	Intercept	56137.9637	14108.0130	3.9792	0.0003	27552.4142	84723.5132
81	X3	6534.7594	178.7085	36.5666	0.0000	6172.6616	6896.8572
82	X2	-50.9283	116.9213	-0.4356	0.6657	-287.8334	185.9769
83							

Fig. 14.9

For X_3 and X_4: We consider **Y range** as Cells B1:B41 **X range** as Cells N1:O41 and **output range** as Cell U64. The result is shown in Fig. 14.10.

	U	V	W	X	Y	Z	AA
73	ANOVA						
74		*df*	*SS*	*MS*	*F*	*Significance F*	
75	Regression	2.0000	60722895691.0313	30361447845.5156	1897.9774	0.0000	
76	Residual	37.0000	591879308.9688	15996738.0802			
77	Total	39.0000	61314775000.0000				
78							
79		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
80	Intercept	2036.6194	17306.5584	0.1177	0.9070	-33029.7985	37103.0373
81	X3	5562.0325	379.2959	14.6641	0.0000	4793.5060	6330.5589
82	X4	5098.1839	1798.8116	2.8342	0.0074	1453.4455	8742.9223
83							

Fig. 14.10

Interpretation

We now study Figs. 14.8, 14.9 and 14.10 and notice that the highest absolute value of t-statistic is 3.5009 given in Cell D82 (Fig. 14.8), which is associated with regressor X_1. The p-value of t-statistic associated with the regressor X_1 is 0.0012 given in Cell E82, which is less than $\alpha_{IN} = 0.05$. Thus, the regressor X_1 is contributing significantly to the model. We now conclude that the regressor X_1 (advertising cost) enters in the model as a second regressor.

- The second regressor to be entered in the model is the advertising cost (X_1).
- The lowest p-value of t test = 0.0012.
- From Fig. 14.8, the model is:

  $\hat{y} = -4221.106 + 4436.617X_3 + 32.678X_1$
- F test p-value = 0.0000.

Step 4: To select the third regressor in the model

After we select two regressors in the model, we consider all possible models with 3 regressors, keeping in mind that X_3 and X_1 are already entered in the model. The possible combinations of the regressors will be (X_3, X_1, X_2) and (X_3, X_1, X_4). For performing the regression analysis in Excel, we keep the values of (X_3, X_1, X_2) and (X_3, X_1, X_4) in adjacent columns, i.e., Cells Q1:S41 and U1:W41, respectively, as shown in Fig. 14.11.

	Q	R	S	T	U	V	W
1	**X3**	**X1**	**X2**		**X3**	**X1**	**X4**
2	24	3240	85		24	3240	14
3	28	3550	94		28	3550	15
4	33	3890	77		33	3890	16
5	31	3720	84		31	3720	16
6	35	4030	75		35	4030	17
7	32	3790	81		32	3790	16
8	34	3950	76		34	3950	16
9	24	3290	89		24	3290	15
10	21	2920	79		21	2920	13
11	33	3850	78		33	3850	16
12	35	4030	75		35	4030	17
13	28	3550	94		28	3550	15
14	24	3240	85		24	3240	14
15	39	4200	68		39	4200	17
16	27	3470	94		27	3470	15
17	24	3290	89		24	3290	15
18	42	4340	64		42	4340	18
19	24	3330	94		24	3330	14
20	32	3790	81		32	3790	16
21	44	4400	62		44	4400	18
22	40	4240	67		40	4240	17
23	30	3660	88		30	3660	16
24	31	3720	84		31	3720	16
25	36	4080	73		36	4080	17
26	25	3380	94		25	3380	15

Forward Selection

Fig. 14.11

To discover the third regressor to be entered in the model, we execute the regression analysis for combinations of the three regressors in the model one at a time as discussed in Steps 2 and 3.

For X_3, X_1 and X_2: We consider **Y range** as Cells B1:B41, **X range** as Cells Q1:S41 and **output range** as Cell A85. The result is shown in Fig. 14.12.

	A	B	C	D	E	F	G
94	ANOVA						
95		*df*	*SS*	*MS*	*F*	*Significance F*	
96	Regression	3.0000	61009129636.2336	20336376545.4112	2395.2909	0.0000	
97	Residual	36.0000	305645363.7664	8490148.9935			
98	Total	39.0000	61314775000.0000				
99							
100		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
101	Intercept	2150.5738	12142.4667	0.1771	0.8604	-22475.4899	26776.6376
102	X3	1445.8446	740.8706	1.9515	0.0588	-56.7106	2948.3998
103	X1	68.2135	9.8034	6.9581	0.0000	48.3312	88.0958
104	X2	-561.9502	106.7039	-5.2664	0.0000	-778.3557	-345.5446
105							

Fig. 14.12

For X_3, X_1 and X_4: We consider **Y range** as Cells B1:B41, **X range** as Cells U1:W41 and **output range** as Cell K85. The result is shown in Fig. 14.13.

	K	L	M	N	O	P	Q
94	ANOVA						
95		*df*	*SS*	*MS*	*F*	*Significance F*	
96	Regression	3.0000	60809972541.5603	20269990847.1868	1445.5549	0.0000	
97	Residual	36.0000	504802458.4397	14022290.5122			
98	Total	39.0000	61314775000.0000				
99							
100		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
101	Intercept	-20674.8643	18590.6180	-1.1121	0.2735	-58378.3847	17028.6562
102	X3	4304.9125	616.9260	6.9780	0.0000	3053.7285	5556.0965
103	X1	25.4067	10.1954	2.4920	0.0174	4.7294	46.0839
104	X4	3023.5790	1878.6777	1.6094	0.1163	-786.5559	6833.7138
105							

Fig. 14.13

Interpretation

It is the same as for Step 3. We find the regressor having the highest absolute value of t-statistic. It is 5.2664 given in Cell D104 corresponding to the regressor X_2 (Fig. 14.12) and its p-value is 0.0000 given in Cell E104, which is less than $\alpha_{IN} = 0.05$. Hence, the regressor X_2 is contributing significantly to the model. We conclude that the regressor X_2 (price) enters in the model as a third regressor.

- The third regressor to be entered in the model is price (X_2).
- The lowest p-value of t test = 0.0000.
- From Fig.14.12, the Model is

  $$\hat{Y} = 2150.574 + 1445.845X_3 + 68.213X_1 - 561.95X_2$$

- F test p-value = 0.0000.

Step 5: To select the fourth regressor in the model

We repeat the previous steps by considering the combination of 4 regressors, i.e., (X_3, X_1, X_2, X_4). Here we have kept these regressors in Cells Y1:AB41 as shown in Fig. 14.14 to keep them in adjacent columns.

	Y	Z	AA	AB
1	X3	X1	X2	X4
2	24	3240	85	14
3	28	3550	94	15
4	33	3890	77	16
5	31	3720	84	16
6	35	4030	75	17
7	32	3790	81	16
8	34	3950	76	16
9	24	3290	89	15
10	21	2920	79	13
11	33	3850	78	16
12	35	4030	75	17
13	28	3550	94	15
14	24	3240	85	14
15	39	4200	68	17
16	27	3470	94	15
17	24	3290	89	15
18	42	4340	64	18
19	24	3330	94	14
20	32	3790	81	16
21	44	4400	62	18
22	40	4240	67	17
23	30	3660	88	16
24	31	3720	84	16
25	36	4080	73	17
26	25	3380	94	15

Forward Selection

Fig. 14.14

For X_3, X_1, X_2 and X_4 : We consider **Y range** as Cells B1:B41, **X range** as Cells Y1:AB41 and **output range** as Cell A107. The result is shown in Fig. 14.15.

	A	B	C	D	E	F	G
116	ANOVA						
117		*df*	*SS*	*MS*	*F*	*Significance F*	
118	Regression	4.0000	61059504597.6716	15264876149.4179	2092.9597	0.0000	
119	Residual	35.0000	255270402.3284	7293440.0665			
120	Total	39.0000	61314775000.0000				
121							
122		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
123	Intercept	-17069.7198	13421.7442	-1.2718	0.2118	-44317.3089	10177.8693
124	X3	1195.1257	693.2697	1.7239	0.0936	-212.2867	2602.5381
125	X1	60.7615	9.5184	6.3836	0.0000	41.4381	80.0850
126	X2	-579.8462	99.1325	-5.8492	0.0000	-781.0958	-378.5966
127	X4	3569.2517	1358.1134	2.6281	0.0127	812.1350	6326.3684
128							

Fig. 14.15

- The fourth regressor to be entered in the model is the number of stores (X_4).
- The lowest p-value of t test = 0.0127.
- From Fig. 14.15, the model is:

  $\hat{y} = -17069.720 + 1195.126X_3 + 60.762X_1 - 579.846X_2 + 3569.252X_4$
- F test p-value = 0.0000.

Interpretation

Fig. 14.15 reveals that the absolute value of t-statistic corresponding to the regressor X_4 is 2.6281 (given in Cell D127) and its p-value is 0.0127 (given in Cell E127), which is less than α_{IN} = 0.05. So we can keep the last regressor X_4 (number of stores) in the model.

In this way, the forward selection method is used to build an appropriate model. The final model is

$$Y = -17069.7198 + 1195.1257X_3 + 60.7615X_1 - 579.8462X_2 + 3569.2517X_4$$

We can also note from Fig. 14.15 that the p-value corresponding to F-statistic is 0.0000 given in Cell F118, which is less than 0.05. We can conclude that the overall model is also significant at 5% level of significance.

14.4 BACKWARD ELIMINATION METHOD

In Sec. 14.3, you have learnt how to choose the important regressors which can really contribute in the model using the forward selection method in Excel 2007. In this section, you will learn the backward elimination method of variable selection. As the name itself suggests, this method goes in the reverse direction of the forward selection method.

As you have already studied the backward elimination method in Unit 12 of MSTE-002, we provide only a brief overview of the steps involved. In this method, we assume a model with all regressors and eliminate (drop) one regressor from the model at each step on the basis of p-values of the t-statistic(s) of the regression coefficients. We stop when no more regressors can be significantly eliminated from the model. In this way, we get a **final model**. This process consists of the following steps:

Step 1: Before starting the procedure for the backward elimination, we decide the level of significance to eliminate the regressor from the model, which is called **Alfa-to-Out** (α_{OUT}) to compare the p-values calculated at each stage of the elimination.

Step 2: Let $X_1, X_2, ..., X_p$ be the p regressors. Here, we start by considering all regressors in the model:

$$\hat{Y} = \hat{B}_0 + \hat{B}_1 X_1 + \hat{B}_2 X_2 + ... + \hat{B}_p X_p \qquad ...(2)$$

Step 3: We perform the regression analysis by considering all regressors in the model. To decide the variable that has to be eliminated from the model, we identify the regressor having the lowest absolute t value of the regression coefficient (or partial F value) and compare its p-value with α_{OUT}. If p-value is insignificant, the corresponding regressor will be eliminated from the model.

Step 4: We now leave the regressor eliminated in Step 3 and carry out the regression analysis for the remaining $(p-1)$ regressors in the model. We eliminate the most insignificant regressor from the model. We now have $(p-2)$ regressors in the model.

Step 5: Similarly, we carry out the regression analysis for the remaining $(p-2)$ regressors in the model excluding the two regressors eliminated in Steps 3 and 4. Then we eliminate the most insignificant regressor from the remaining regressors.

Step 6: We repeat the procedure till no regressor gives insignificant p-value.

Steps in Excel

For eliminating a regressor from the model, the significance level or **Alpha-to-Out** denoted by α_{OUT} is given as 0.05. If the p-value is less than α_{OUT}, we consider the corresponding regressor as significant and retain it. Otherwise, we consider it as insignificant and eliminate that regressor.

Step 1: We enter the given data in a new Excel sheet and name it "**Backward Elimination**" as shown in Fig. 14.16.

	A	B	C	D	E	F
1	**S.No.**	**Y**	**X1**	**X2**	**X3**	**X4**
2	1	212000	3240	85	24	14
3	2	230000	3550	94	28	15
4	3	273000	3890	77	33	16
5	4	255000	3720	84	31	16
6	5	285000	4030	75	35	17
7	6	262000	3790	81	32	16
8	7	273000	3950	76	34	16
9	8	215000	3290	89	24	15
10	9	185000	2920	79	21	13
11	10	268000	3850	78	33	16
12	11	285000	4030	75	35	17
13	12	230000	3550	94	28	15
14	13	212000	3240	85	24	14
15	14	309000	4200	68	39	17
16	15	228000	3470	94	27	15
17	16	215000	3290	89	24	15
18	17	324000	4340	64	42	18
19	18	210000	3330	94	24	14
20	19	262000	3790	81	32	16
21	20	329000	4400	62	44	18

Backward Selection

Fig. 14.16: Partial screenshot of the spreadsheet for the given data.

Step 2: To eliminate the first regressor from the model

We perform the regression analysis by considering all the regressors, i.e., X_1, X_2, X_3 and X_4 in the model as discussed in Sec. 14.3. Here we need only the p-value related to regression coefficients. So we specify only the following options as shown in Fig. 14.17:

- ✓ ***Input Y Range*** as Cells B1:B41.
- ✓ ***Input X Range*** as Cells C1:F41.
- ✓ Tick the ***Labels***.
- ✓ ***Output Range*** as Cell A44.

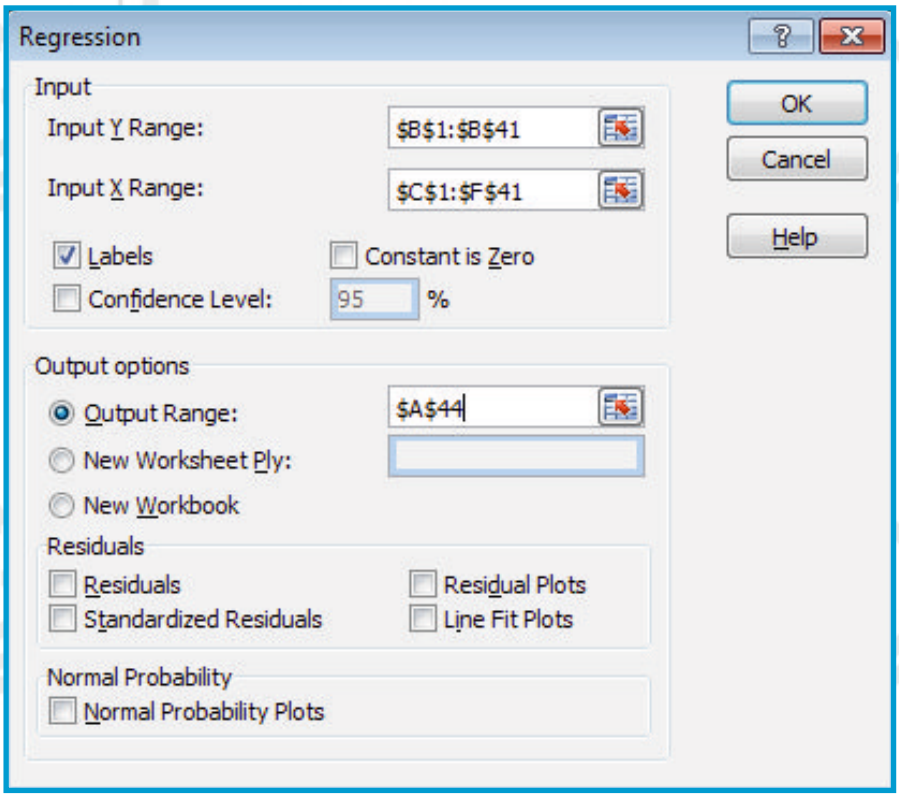


Fig. 14.17

After clicking on ***OK***, we obtain the output shown in Fig. 14.18.

	A	B	C	D	E	F	G
53	ANOVA						
54		*df*	*SS*	*MS*	*F*	*Significance F*	
55	Regression	4.0000	61059504597.6716	15264876149.4179	2092.9597	0.0000	
56	Residual	35.0000	255270402.3284	7293440.0665			
57	Total	39.0000	61314775000.0000				
58							
59		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
60	Intercept	-17069.7198	13421.7442	-1.2718	0.2118	-44317.3089	10177.8693
61	X1	60.7615	9.5184	6.3836	0.0000	41.4381	80.0850
62	X2	-579.8462	99.1325	-5.8492	0.0000	-781.0958	-378.5966
63	X3	1195.1257	693.2697	1.7239	0.0936	-212.2867	2602.5381
64	X4	3569.2517	1358.1134	2.6281	0.0127	812.1350	6326.3684
65							

Fig. 14.18

Interpretation

From Fig. 14.18, the lowest absolute value of t-statistic is 1.7239 given in Cell D63, which is associated with the regressor X_3. The p-value of t-statistic associated with the regressor X_3 is 0.0936 given in Cell E63, which is more than $\alpha_{OUT} = 0.05$. We can conclude that the regressor X_3 is insignificant and it is not contributing to the model. So we can eliminate X_3 (temperature) from the model.

Step 3: To eliminate the second regressor from the model

In Step 2, the regressor X_3 has been eliminated from the model. We now repeat the Step 2 for 3 regressors in the models that exclude X_3, i.e., (X_1, X_2, X_4) to find the second regressor to be eliminated from the model. We have explained in Lab

- We start with the model considering all regressors (Fig. 14.18):

  $\hat{y} = -17069.720 + 60.762X_1 - 579.846X_2 + 1195.1257X_3 + 3569.252X_4$
- The first regressor to be eliminated from the model is the temperature (X_3).
- The highest p-value of t test = 0.0936.
- F test p-value = 0.0000.
- After eliminating X_3, the model is (Fig. 14.21)

  $\hat{y} = -28820.1147 + 75.4046X_1 - 710.9027X_2 + 3891.4261X_4$

Session 12 that we need to put independent variables in the adjacent columns in Excel 2007 to run the regression analysis in case of more than one independent variables. So we have placed the values of (X_1, X_2, X_4) in Cells H1:J41 as shown in Fig. 14.19.

	H	I	J
1	X1	X2	X4
2	3240	85	14
3	3550	94	15
4	3890	77	16
5	3720	84	16
6	4030	75	17
7	3790	81	16
8	3950	76	16
9	3290	89	15
10	2920	79	13
11	3850	78	16
12	4030	75	17
13	3550	94	15
14	3240	85	14
15	4200	68	17
16	3470	94	15
17	3290	89	15
18	4340	64	18
19	3330	94	14
20	3790	81	16
21	4400	62	18
22	4240	67	17
23	3660	88	16
24	3720	84	16
25	4080	73	17
26	3380	94	15

Backward Selection

Fig. 14.19

To discover the second variable which can be eliminated from the model, we perform the regression analysis for the three regressors in the model as discussed in Step 2 by specifying the following:

- ✓ ***Input Y range*** as Cells B1:B41.
- ✓ ***Input X range*** as Cells H1:J41.
- ✓ Tick the ***Labels***.
- ✓ ***Output Range*** as Cell A67 as shown in Fig. 14.20.

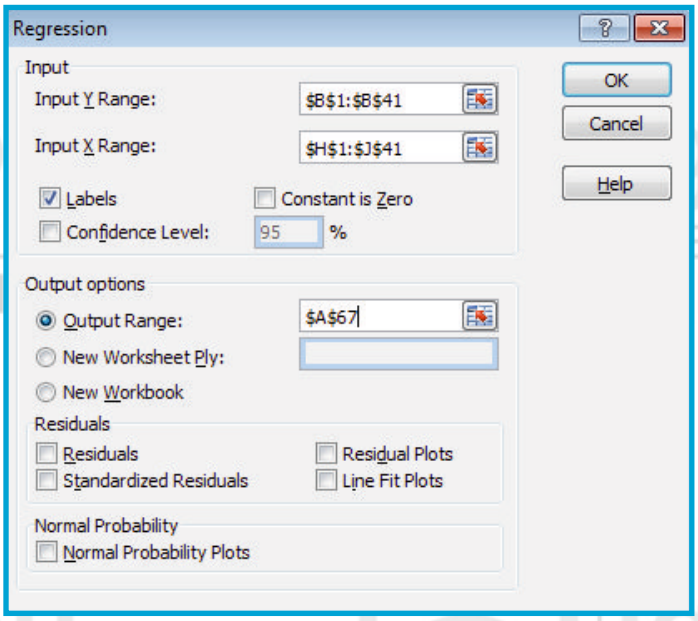


Fig. 14.20

After clicking on ***OK***, we obtain the output shown in Fig. 14.21.

	A	B	C	D	E	F	G
76	ANOVA						
77		*df*	*SS*	*MS*	*F*	*Significance F*	
78	Regression	3.0000	61037829797.5672	20345943265.8557	2644.7613	0.0000	
79	Residual	36.0000	276945202.4328	7692922.2898			
80	Total	39.0000	61314775000.0000				
81							
82		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
83	Intercept	-28820.1147	11874.5560	-2.4270	0.0204	-52902.8303	-4737.3992
84	X1	75.4046	4.4114	17.0931	0.0000	66.4578	84.3513
85	X2	-710.9027	65.3406	-10.8800	0.0000	-843.4195	-578.3859
86	X4	3891.4261	1381.5422	2.8167	0.0078	1089.5287	6693.3236
87							

Fig. 14.21

Interpretation

Fig. 14.21 reveals that the lowest absolute value of t-statistic is 2.8167 given in Cell D86, which is associated with the regressor X_4. The p-value of t-statistic associated with X_4 is 0.0078 given in Cell E86, which is less than $\alpha_{OUT} = 0.05$. So the regressor X_4 is contributing significantly to the model and we cannot eliminate X_4 from the model. Since no regressor is showing insignificant contribution, we terminate the process at this step. We have three regressors, i.e., advertising cost (X_1), price (X_2) and number of stores (X_4) in the model. Hence the resulting model is

$$\hat{Y} = -28820.1147 + 75.4046X_1 - 710.9027X_2 + 3891.4261X_4$$

Notice from Fig. 14.21 that the value of F-statistic is 2644.7613 given in Cell E78 and its p-value is 0.0000 given in Cell F78, which is less than 0.05. We conclude that the overall model is also significant.

- No more regressor is insignificant.
- The highest p-value of t test = 0.0078.
- From Fig. 14.21, the model is:

  $\hat{y} = -28820.1147 + 75.4046X_1 - 710.9027X_2 + 3891.4261X_4$
- F test p-value = 0.0000.

14.5 STEPWISE REGRESSION METHOD

You have learnt in Sec. 14.3 that the forward selection method begins without considering any regressor in the model and then we include the most significant variable one at a time, at each step. You have also learnt in Sec.14.4 that backward elimination method is just the reverse of the forward selection method. This procedure starts with all regressors in the model. Then we drop the most insignificant regressor variable one at a time, at each step. In this section, you will learn how to apply stepwise regression in Excel 2007. The stepwise regression method is a combination of both forward selection and backward elimination methods.

In this method, we use hypothesis testing approach as discussed in Secs. 14.3 and 14.4 to check the significance of the regression coefficients at each stage for possible selection and elimination.

In stepwise regression method, we start with forward selection method and also check for the possible elimination of the insignificant variable at each step. You have studied the stepwise regression method in Unit 12 of MSTE-002. We provide only a brief overview of the steps involved as follows:

Step 1: Before starting the procedure for stepwise regression, we decide the levels of significance to enter and eliminate the regressor, i.e., α_{IN} and α_{OUT}, respectively, to compare the p-values.

Step 2: We start by assuming a model with no regressor in it, i.e., we take

$$Y = B_0 + e \qquad ...(3)$$

Step 3: We perform the regression analysis for all possible one-regressor models and decide the variable to be entered in the model. We identify the regressor having the largest absolute t value of the regression coefficient (or partial F value) and also calculate its p-value. If the p-value is significant, the corresponding regressor enters in the model.

Step 4: We perform the regression analysis for all possible two-regressor models containing one regressor obtained from Step 3 and the second from the remaining regressors. We compute the p-value for the regression coefficients of the remaining regressors and include the most significant regressor in the model. Thus, we obtain 2 regressors in the model.

Step 5: Note that before checking the possibility for addition of the next regressor, we also check the significance of the regressor, which was included in the model at Step 3. If both regressors are still significant, we go to Step 7. Otherwise, we go to Step 6.

Step 6: We eliminate the insignificant regressor if its p-value is more than α_{OUT}. We repeat Step 4 with two regressors in the models, taking one regressor which was included at Step 4 and the other regressor from the remaining regressors excluding the regressor which we eliminated in this step.

Step 7: Similarly, we carry out the regression analysis for all possible three-regressor models containing two regressors obtained from Steps 5 or 6 and the third regressor from the remaining regressors. We select the most significant regressor from the remaining regressors. At every step, we check for the possiblility of elimination of the insignificant regressor.

Step 8: We repeat the procedure till all regressors are included in the model or no regressor gives significant p-value to enter into the model and all regressors in the model are significant.

Steps in Excel

We first set a significance level for deciding a regressor to be entered into the model, i.e., α_{IN} (**Alpha-to-In**) and significance level for deciding a regressor to be eliminated from the model, i.e., α_{OUT} (**Alpha-to-Out**). It is given that $\alpha_{IN} = \alpha_{OUT} = 0.05$. If the p-value is less than or equal to α_{IN} or α_{OUT}, we consider the corresponding regressor as significant, otherwise insignificant.

Step 1: We enter the data in an Excel Sheet and name it "**Stepwise Regression**" as shown in Fig. 14.22.

	A	B	C	D	E	F
1	S.No.	Y	X1	X2	X3	X4
2	1	212000	3240	85	24	14
3	2	230000	3550	94	28	15
4	3	273000	3890	77	33	16
5	4	255000	3720	84	31	16
6	5	285000	4030	75	35	17
7	6	262000	3790	81	32	16
8	7	273000	3950	76	34	16
9	8	215000	3290	89	24	15
10	9	185000	2920	79	21	13
11	10	268000	3850	78	33	16
12	11	285000	4030	75	35	17
13	12	230000	3550	94	28	15
14	13	212000	3240	85	24	14
15	14	309000	4200	68	39	17
16	15	228000	3470	94	27	15
17	16	215000	3290	89	24	15
18	17	324000	4340	64	42	18
19	18	210000	3330	94	24	14
20	19	262000	3790	81	32	16
21	20	329000	4400	62	44	18

Stepwise Regression

Fig. 14.22: Partial screenshot of the spreadsheet for the given data.

Step 2: To select the first regressor in the model

In stepwise regression method, the selection of the first regressor variable is done in the same way as discussed in Step 2 of Sec. 14.3. From Fig. 14.5, the highest absolute value of t-statistic was 56.5364 given in Cell X61, which is associated with regressor X_3. Since the p-value of t-statistic was 0.0000 (given in Cell Y61) and it is less than $\alpha_{IN} = 0.05$, the regressor X_3 is contributing significantly to the model. So we include X_3 (temperature) in the model.

- The first regressor to be entered in the model is temperature (X_3).
- The lowest p-value of t test = 0.0000.
- From Fig. 14.5, the model is:

  $\hat{Y} = 50197.432 + 6593.264X_3$
- F test p-value = 0.0000.

Step 3: To select the second regressor in the model and check for possible elimination

This step is also the same as Step 3 of Sec.14.3. So we follow Step 3 of Sec. 14.3 where the highest absolute value of t-statistic was 3.5009 given in Cell D82 and was associated with the regressor X_1. The p-value of t-statistic associated with the X_1 was 0.0012 given in Cell E82 (Fig. 14.8), which is less than $\alpha_{IN} = 0.05$. So we include X_1 (advertising cost) in the model since the regressor X_1 is contributing significantly to the model as shown in Fig. 14.8.

The only difference is that now we look again at the regressor which was included in the model at Step 2 for the possible elimination. From Fig. 14.8 of Sec. 14.3, the p-value corresponding to X_3 is 0.0000 given in Cell E81. It is still significant at 5% level of significance and so we retain the regressor X_3.

At this step, we have two regressors X_3 and X_1 in the model.

- The second regressor to be entered in the model is the advertising cost (X_1).
- The lowest p-value of t test = 0.0012.
- From Fig. 14.8, the model is:

  $\hat{y} = -4221.106 + 4436.617X_3 + 32.678X_1$
- F test p-value = 0.0000.
- Since all p-values of t tests < 0.05, no regressor will be eliminated at this step.

- The third regressor to be entered in the model is price (X_2).
- The lowest p-value of t test = 0.0000.
- From Fig.14.12, the Model is

$$\hat{Y} = 2150.574 + 1445.845X_3 + 68.213X_1 - 561.95X_2$$

- F test p-value = 0.0000.
- From Fig. 14.12, since p-value = 0.0936 > 0.05, the regressor to be eliminated from the model is the temperature (X_3).

Step 4: To select the third regressor in the model and check for possible elimination

The selection procedure of third regressor for this data set is the same as Step 4 of Sec. 14.3. You have noted there that we entered the regressor X_2 in the model as it had the highest absolute value of t-statistic, i.e., 5.2664 and the corresponding p-value is 0.0000 given in Cells D104 and E104, respectively (Fig. 14.12). It was contributing significantly to the model. So we include X_2 (price) in the model. To check for possible elimination, we go back to Fig. 14.12 of Sec. 14.3.

Notice from Fig. 14.12 that the p-value corresponding to X_3 is 0.0588 given in Cell E102. It shows insignificant contribution to the model after including the regressor X_2 at 5% level of significance. So we eliminate X_3 from the model. At this step we again have two regressors in the model, i.e., X_1 and X_2.

Step 5: To select the third regressor in the model and check for possible elimination

We now consider all possible models with 3 regressors by keeping in mind that X_1 and X_2 are already there in the model and we have eliminated X_3 from the model. So the possible combination is (X_1, X_2, X_4). For performing the regression analysis in Excel, we keep the values of (X_1, X_2, X_4) in adjacent columns, i.e., Cells Y1:AA41 as shown in Fig. 14.23.

	Y	Z	AA	AB
1	**X1**	**X2**	**X4**	
2	3240	85	14	
3	3550	94	15	
4	3890	77	16	
5	3720	84	16	
6	4030	75	17	
7	3790	81	16	
8	3950	76	16	
9	3290	89	15	
10	2920	79	13	
11	3850	78	16	
12	4030	75	17	
13	3550	94	15	
14	3240	85	14	
15	4200	68	17	
16	3470	94	15	
17	3290	89	15	
18	4340	64	18	
19	3330	94	14	
20	3790	81	16	
21	4400	62	18	

Stepwise Regression

Fig. 14.23

To discover the third regressor to be entered in the model, we perform the regression analysis for this combination of three regressors in the model.

For X_1, X_2 and X_4: We consider **Y range** as Cells B1:B41, **X range** as Cells Y1:AA41 and **output range** as Cell A107. The output is shown in Fig. 14.24.

	A	B	C	D	E	F	G
116	ANOVA						
117		*df*	*SS*	*MS*	*F*	*Significance F*	
118	Regression	3	61037829797.5672	20345943265.8557	2644.7613	0.0000	
119	Residual	36	276945202.4328	7692922.2898			
120	Total	39	61314775000				
121							
122		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
123	Intercept	-28820.1147	11874.5560	-2.4270	0.0204	-52902.8303	-4737.3992
124	X1	75.4046	4.4114	17.0931	0.0000	66.4578	84.3513
125	X2	-710.9027	65.3406	-10.8800	0.0000	-843.4195	-578.3859
126	X4	3891.4261	1381.5422	2.8167	0.0078	1089.5287	6693.3236
127							

Fig. 14.24

Note that the regressor having the highest absolute value of t-statistic, i.e., 2.8167 given in Cell D126 corresponds to X_4 (see Fig. 14.24). Its p-value is 0.0078 given in Cell E126, which is less than $\alpha_{IN} = 0.05$. Therefore, the regressor X_4 is contributing significantly to the model. We can conclude that X_4 (number of stores) enters in the model as a third regressor. We now check for possible elimination. Fig. 14.24 reveals that the p-value corresponding to both the previously added regressors X_1 and X_2 is 0.0000 given in Cells E124 and E125, respectively, which is less than α_{OUT}. Both regressors X_1 and X_2 are still contributing significantly to the model. So we will not eliminate them. In this way we have three regressors X_1, X_2 and X_4 in the model.

Step 6: To select the fourth regressor in the model and check for possible elimination

We now consider the combination of 4 regressors, i.e., (X_1, X_2, X_4, X_3). We ***Copy*** and ***Paste*** the values of these regressors in Cells AC1:AF41 to keep them in the adjacent columns as shown in Fig. 14.25.

	AC	AD	AE	AF
1	**X1**	**X2**	**X4**	**X3**
2	3240	85	14	24
3	3550	94	15	28
4	3890	77	16	33
5	3720	84	16	31
6	4030	75	17	35
7	3790	81	16	32
8	3950	76	16	34
9	3290	89	15	24
10	2920	79	13	21
11	3850	78	16	33
12	4030	75	17	35
13	3550	94	15	28
14	3240	85	14	24
15	4200	68	17	39
16	3470	94	15	27
17	3290	89	15	24
18	4340	64	18	42
19	3330	94	14	24
20	3790	81	16	32
21	4400	62	18	44

Stepwise Regression

Fig. 14.25

- The third regressor to be entered in the model is number of stores (X_4).
- The lowest p-value of t test = 0.0078.
- From Fig. 14.24, the model is (Fig. 14.24)

  $\hat{y} = -28820.1147 + 75.4046X_1 - 710.9027X_2 + 3891.4261X_4$
- F test p-value = 0.0000.
- Since all p-values of t tests < 0.05, no regressor will be eliminated at this step.

We perform the regression analysis as discussed in Step 5 of Sec. 14.3 and Step 2 of Sec. 14.4.

We consider **Y range** as Cells B1:B41, **X range** as Cells AC1:AF41 and output **range** as Cell A129. The output is shown in Fig. 14.26.

	A	B	C	D	E	F	G
138	ANOVA						
139		*df*	*SS*	*MS*	*F*	*Significance F*	
140	Regression	4.0000	61059504597.6716	15264876149.4179	2092.9597	0.0000	
141	Residual	35.0000	255270402.3284	7293440.0665			
142	Total	39.0000	61314775000.0000				
143							
144		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
145	Intercept	-17069.7198	13421.7442	-1.2718	0.2118	-44317.3089	10177.8693
146	X1	60.7615	9.5184	6.3836	0.0000	41.4381	80.0850
147	X2	-579.8462	99.1325	-5.8492	0.0000	-781.0958	-378.5966
148	X4	3569.2517	1358.1134	2.6281	0.0127	812.1350	6326.3684
149	X3	1195.1257	693.2697	1.7239	0.0936	-212.2867	2602.5381
150							

Fig. 14.26

Note from Fig. 14.26 that the absolute value of t-statistic corresponding to the regressor X_3 is 1.7239 given in Cell D149. Its p-value is 0.0936 given in Cell E149, which is greater than $\alpha_{IN} = 0.05$. Therefore, we cannot enter the last regressor X_3 (temperature) in the model as it is not contributing significantly. From Fig. 14.26, you can see that the p-values corresponding to the other regressors, i.e., X_1, X_2 and X_4 are less than $\alpha_{OUT} = 0.05$. So we can conclude that the regressors X_1, X_2 and X_4 contribute significantly to the model and we do not eliminate X_1, X_2 and X_4. In this way, we have obtained the best fitted regression model with regressors X_1, X_2 and X_4 using stepwise regression method. Hence from Fig. 14.24, the resulting model is

- No more regressor is significant.
- From Fig. 14.24, the model is:

  $\hat{y} = -28820.1147 + 75.4046X_1 - 710.9027X_2 + 3891.4261X_4$
- F test p-value = 0.0000.

$$\hat{Y} = -28820.1147 + 75.4046X_1 - 710.9027X_2 + 3891.4261X_4$$

Notice from Fig.14.24 that the value of F-statistic is 2644.7613 given in Cell E118 and its p-value is 0.0000 given in Cell F118, which is less than 0.05. We conclude that the overall model is also significant.

You should now apply the above methods on other problems for practice.

Activity

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

A1) Example 2 given in Unit 12 of MSTE-002.

A2) Exercise E2 given in Unit 12 of MSTE-002.

Match the results with the manual calculation done in Unit 12 of MSTE-002.

Continuous Assessment 14

Consider the data given in Continuous Assesment 12 of Lab Session 12 to build an appropriate linear regression model for electricity consumption of a household. Consider two more regressors, i.e., temperature and number of persons living in the house as given in Table 2.

Build a regression model by selecting appropriate regressors in the model using (i) forward selection, (ii) backward elimination and (iii) stepwise regression methods.

Table 2: Electricity consumption data

S. No.	Unit	Area	AC	Temperature	No. of Persons
1	1060	1316	5	31	2
2	1150	1420	7	36	4
3	1365	1556	12	42	5
4	1275	1488	9	32	5
5	1425	1612	13	40	3
6	1310	1516	10	37	5
7	1365	1556	12	36	2
8	1075	1352	6	34	4
9	925	1168	4	32	3
10	1340	1540	11	37	5
11	1425	1612	13	40	5
12	1150	1420	8	32	4
13	1060	1316	5	33	4
14	1545	1680	15	41	6
15	1140	1388	7	32	4
16	1075	1352	6	33	2
17	1620	1736	16	33	7
18	1050	1296	5	34	4
19	1310	1516	10	35	5
20	1645	1760	16	44	7
21	1565	1696	15	42	6
22	1215	1464	9	39	5
23	1275	1488	10	40	5
24	1465	1632	13	42	3
25	1080	1356	7	44	4
26	975	1196	4	32	3
27	1040	1256	5	33	4
28	1340	1540	11	37	5
29	865	1144	4	31	3
30	1175	1440	8	38	4
31	1080	1356	7	37	4
32	1500	1652	15	43	6
33	1175	1440	9	36	4
34	1050	1296	5	31	2
35	1365	1580	12	37	5
36	1465	1632	15	43	6
37	1215	1464	9	33	5
38	1365	1580	12	34	5
39	1140	1388	7	35	4
40	1005	1224	4	36	3

Home Work: Do It Yourself

1) Follow the steps explained in Secs. 14.3 to 14.5 to comprehend the variable selection methods in regression analysis for the data of Table 1. Take the screenshots and keep them in your record book.

2) Develop the spreadsheets for the exercise "Continuous Assessment 14" as explained in this lab session. Take screenshots of the final spreadsheets.

3) **Do not forget** to keep the screenshots in your record book as these will contribute to your continuous assessment in the Laboratory.