Simple Linear Regression

SESSION 11

Structure

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 i^{th} residual and is given by

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

Step 5: The i^{th} 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

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

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:

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

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:

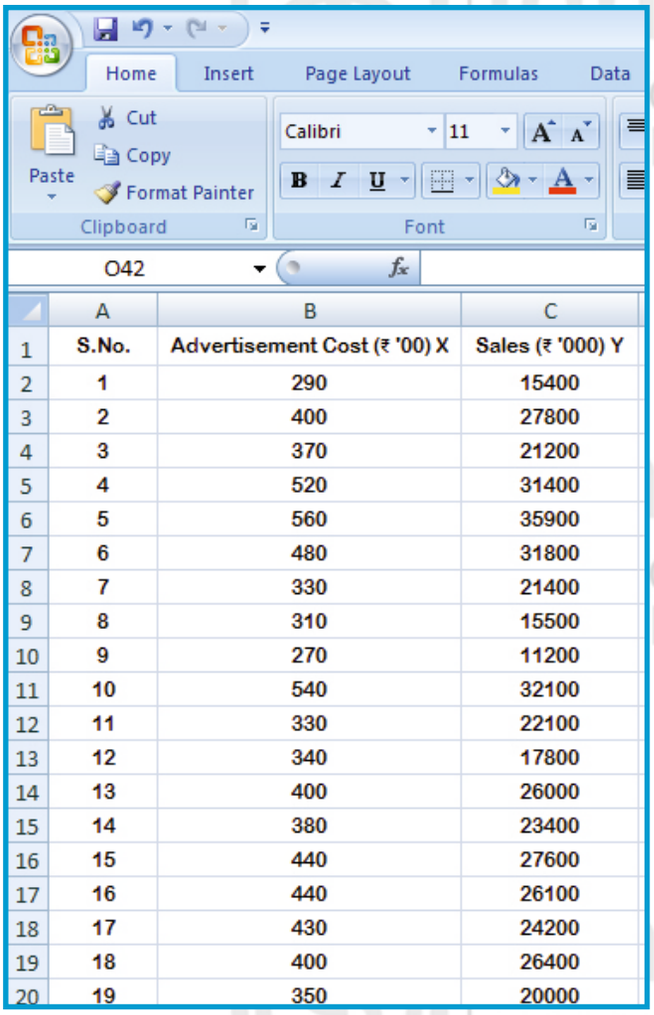
$$P_i = \frac{(i - 1/2)}{n} \times 100, \; i = 1, 2, \ldots, n \qquad \ldots(15)$$

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.

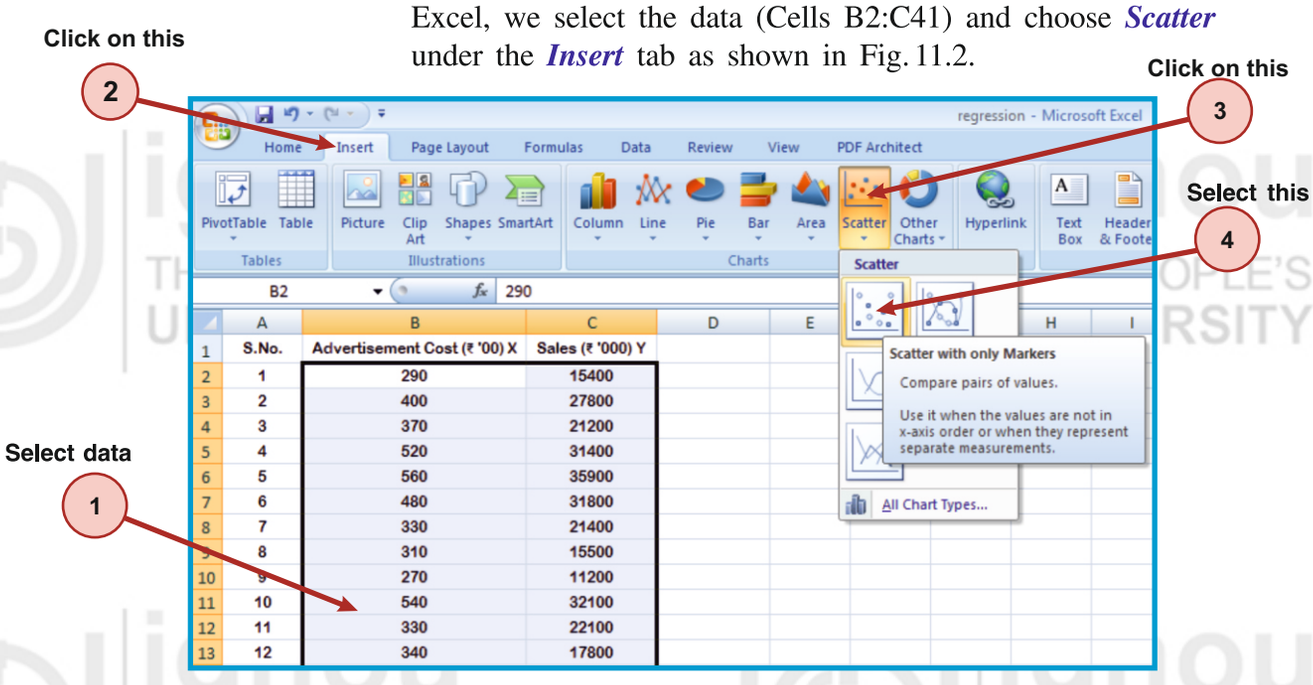


Fig. 11.2

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.

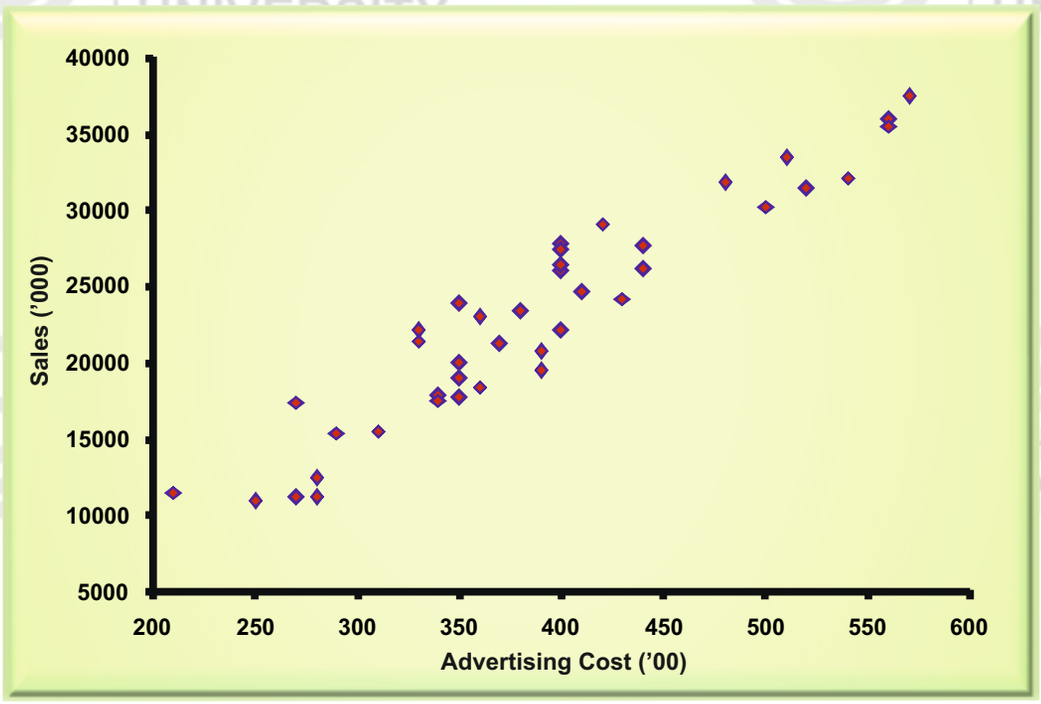


Fig. 11.3: Scatter diagram.

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.

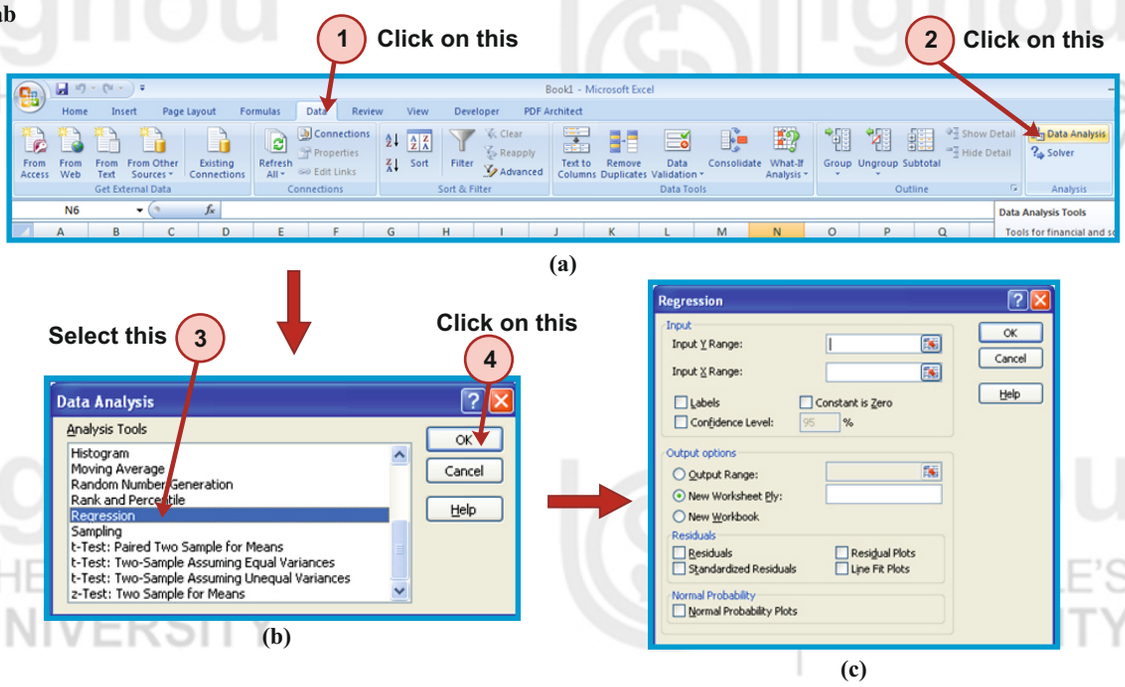


Fig. 11.4

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***.

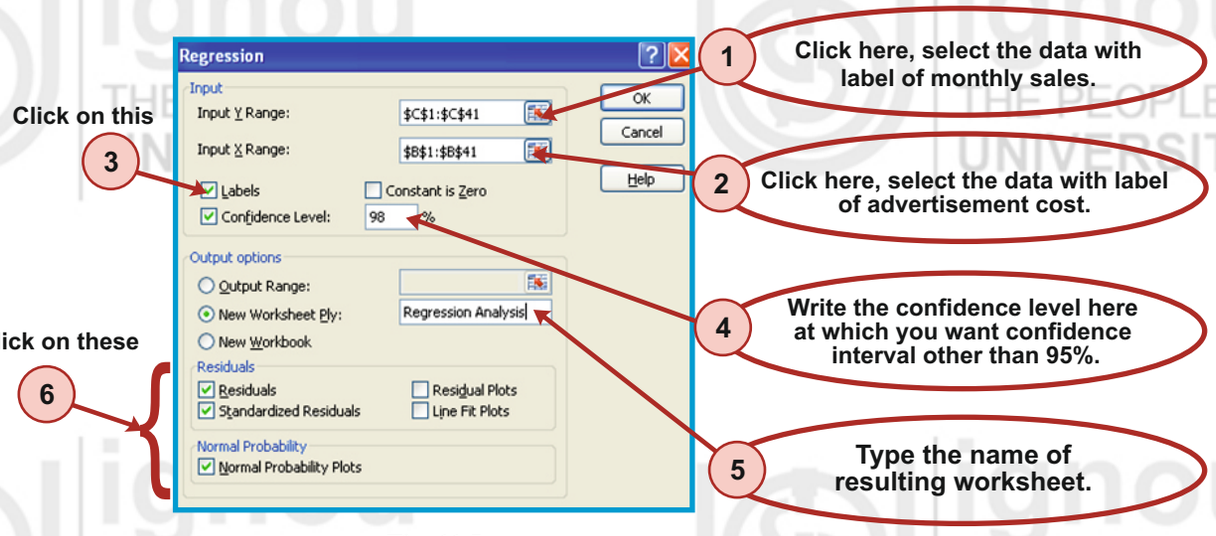
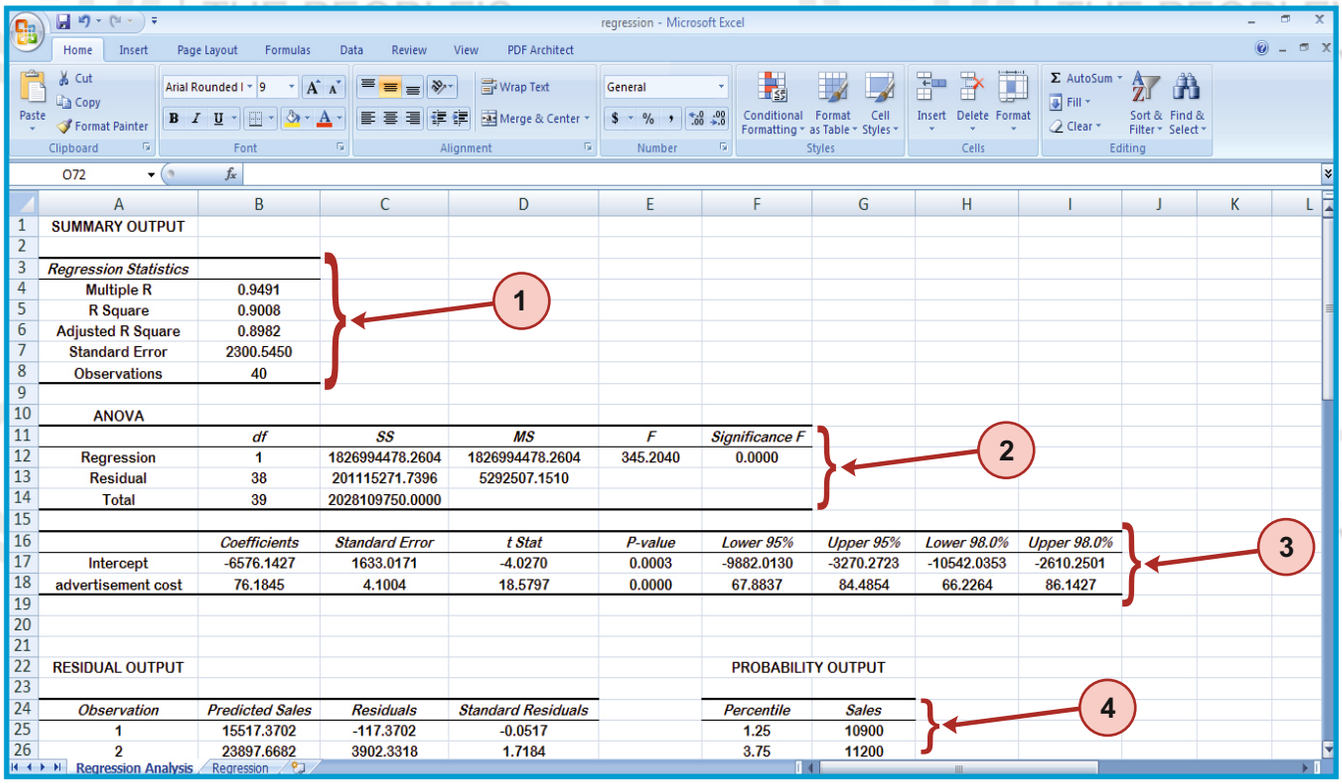


Fig. 11.5

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.

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.

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.)

	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.

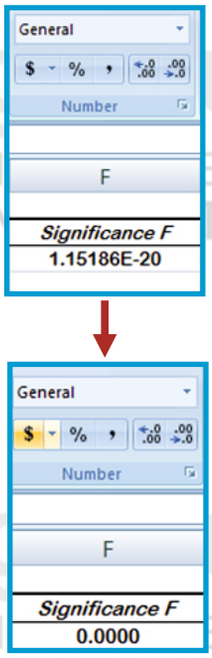
We use an analysis of variance approach to test the significance of the regression model. The analysis of variance is based on the partitioning of the total variability in the response variable Y. Here we use the usual analysis of variance (F-test) to test the null hypothesis $H_0: b = 0$.

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

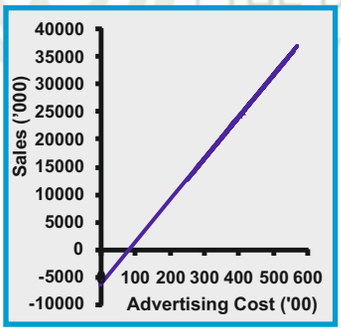


Fig. 11.10

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

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

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

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 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).

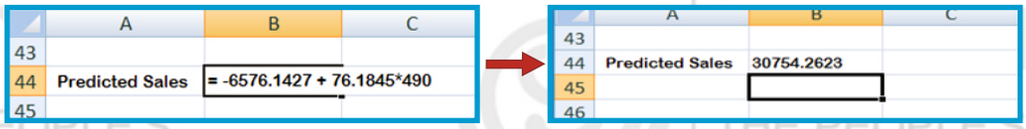


Fig. 11.11

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.,

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

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.

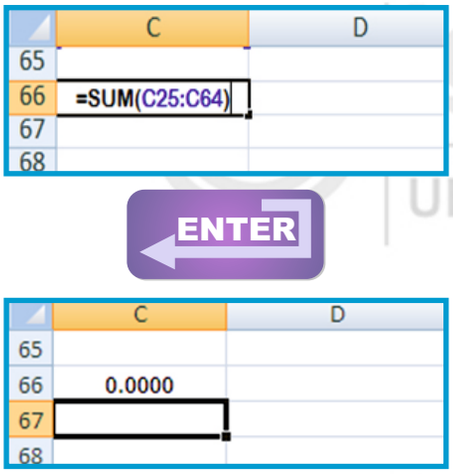


Fig. 11.14

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).

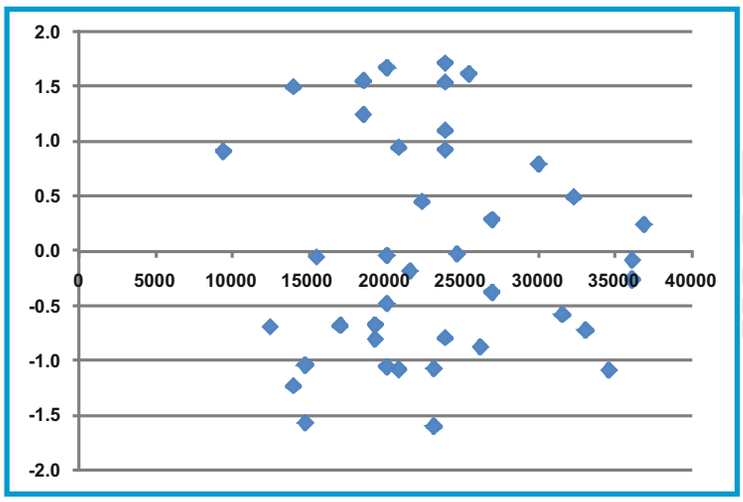


Fig. 11.16

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.

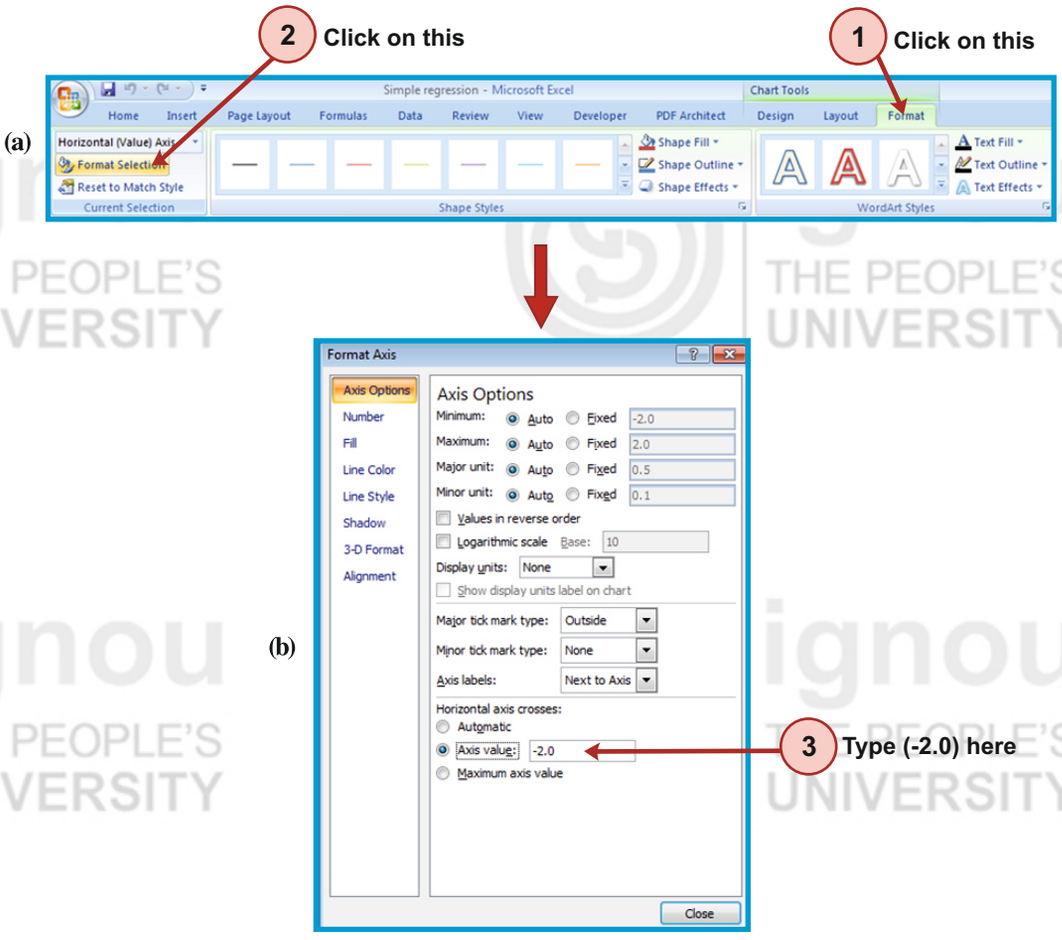


Fig. 11.17

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.

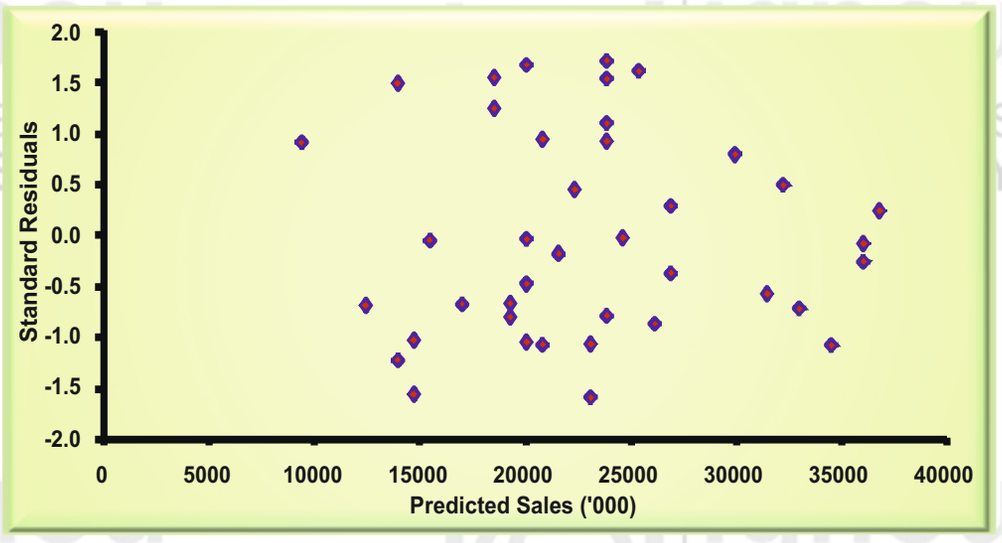


Fig. 11.18: Residual Plot.

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.

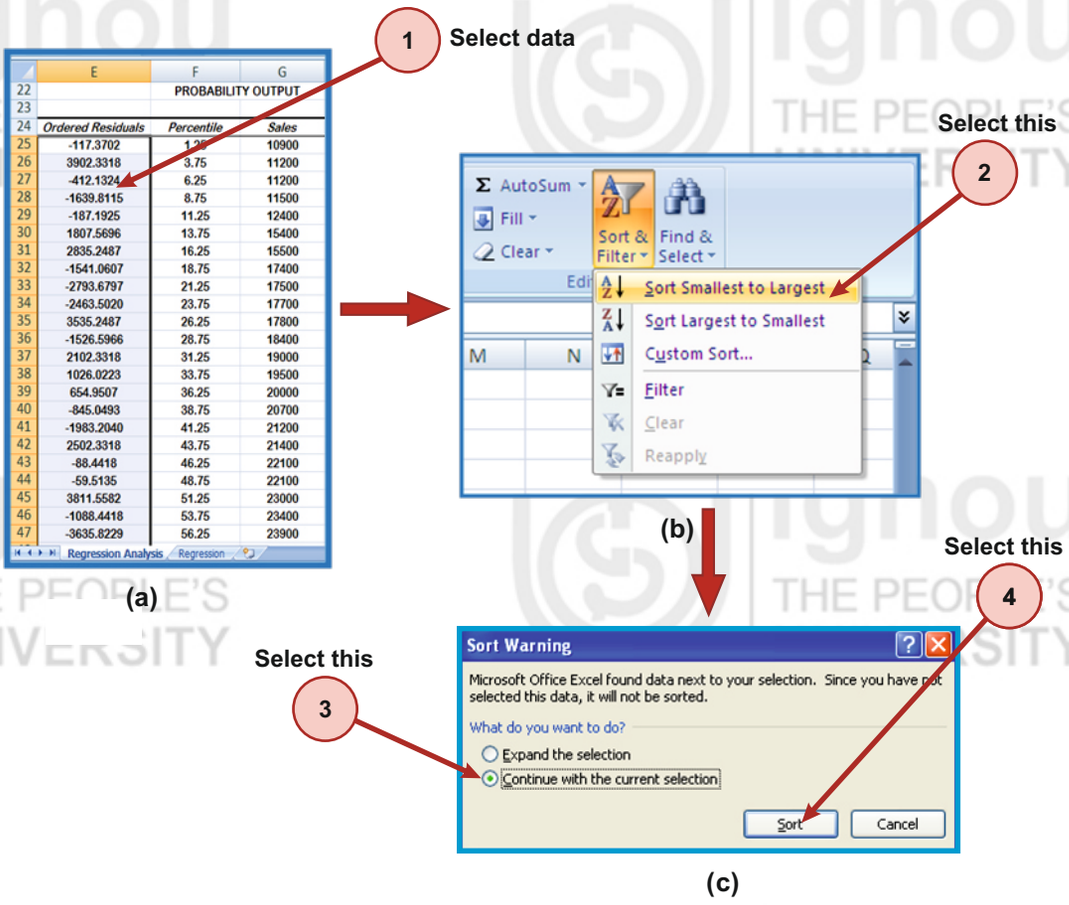


Fig. 11.20

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.

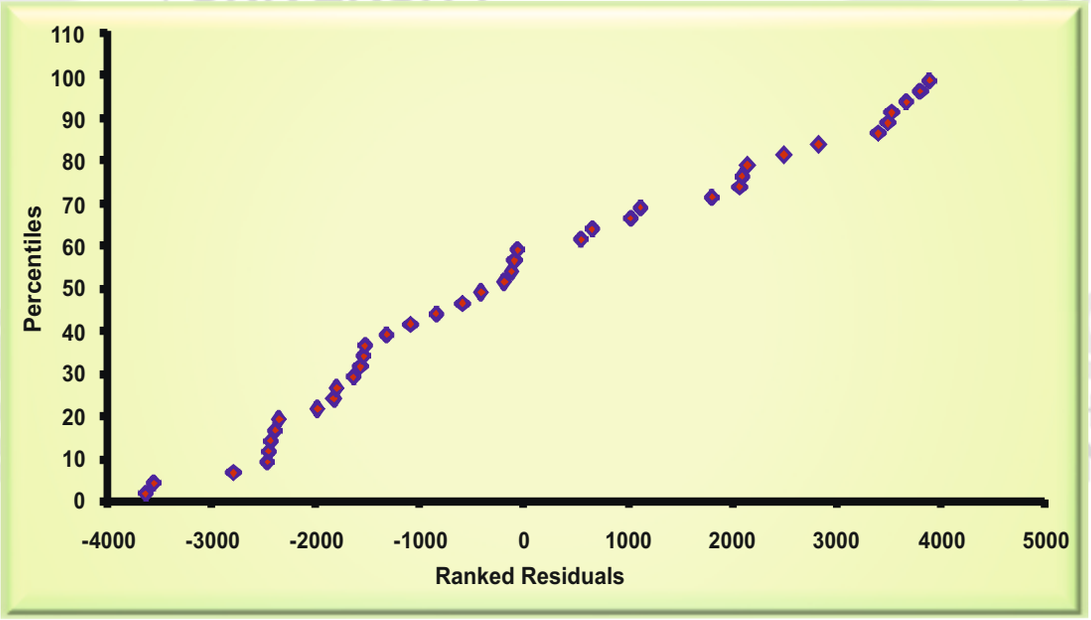


Fig. 11.22: Normal probability plot.

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.

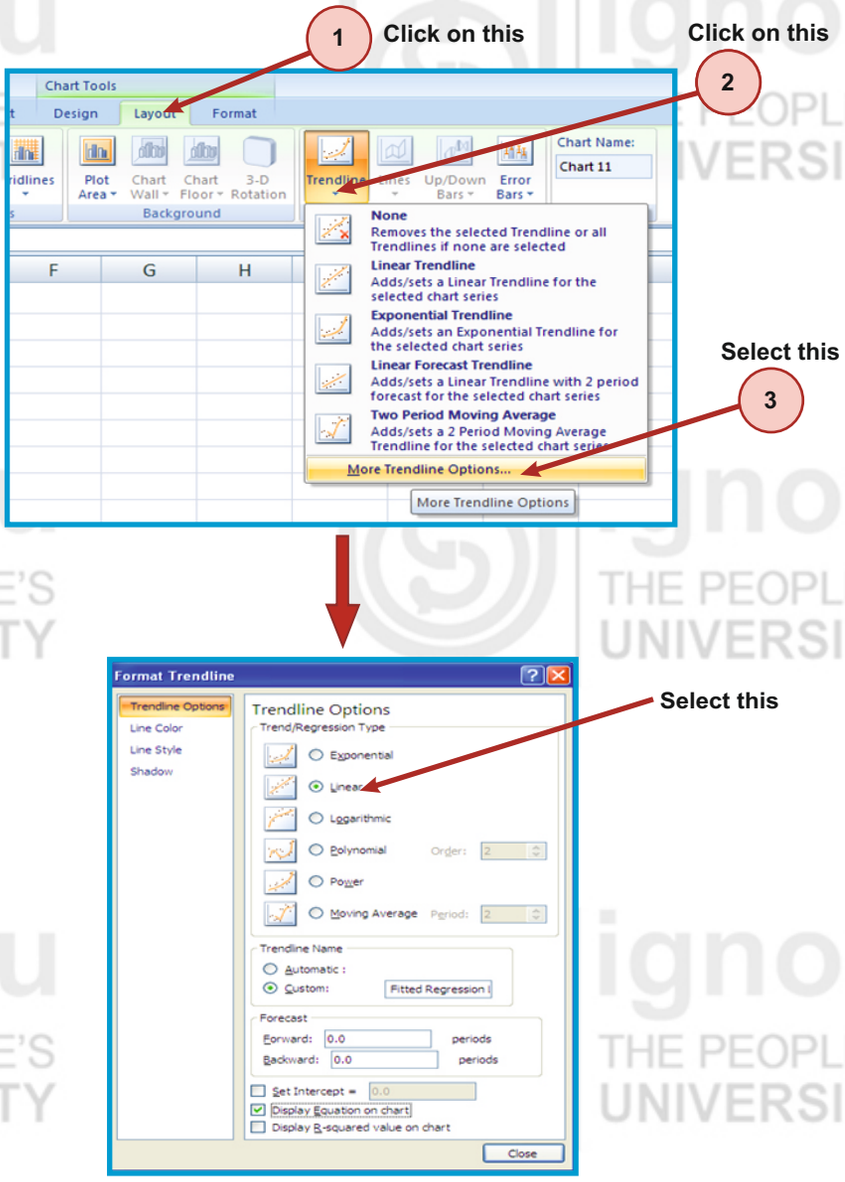


Fig. 11.23

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

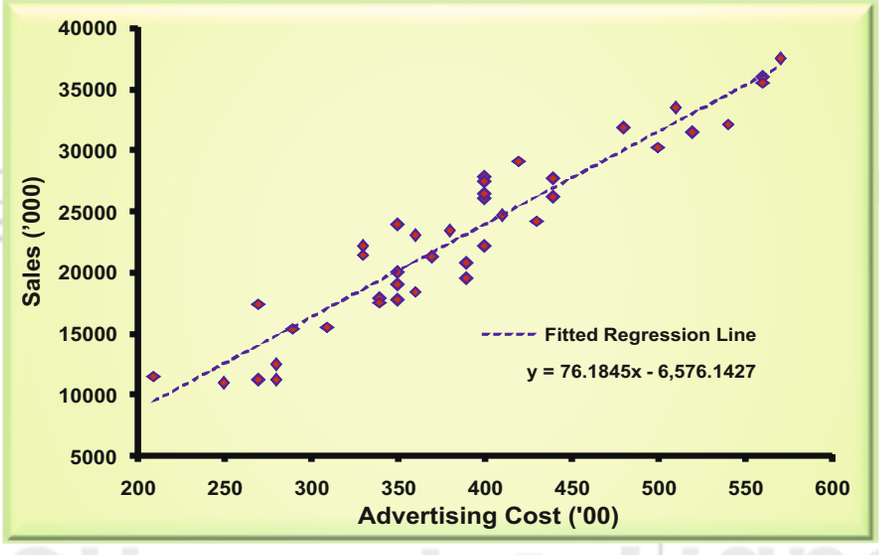


Fig. 11.24

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.