Regression with Dummy Variable

SESSION 13

Structure

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:

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

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.

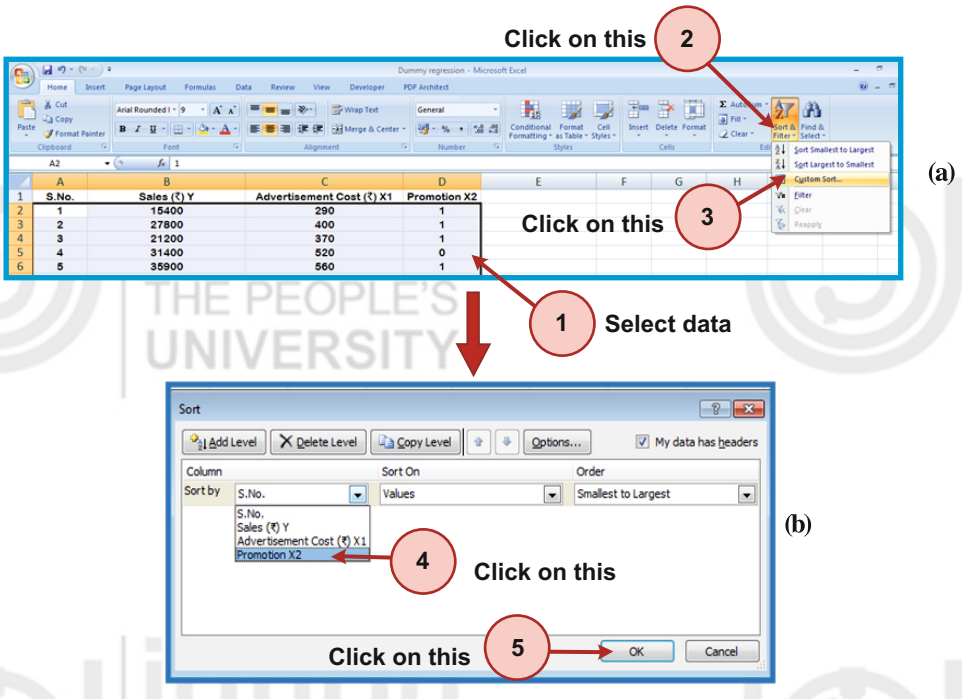


Fig. 13.2

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).

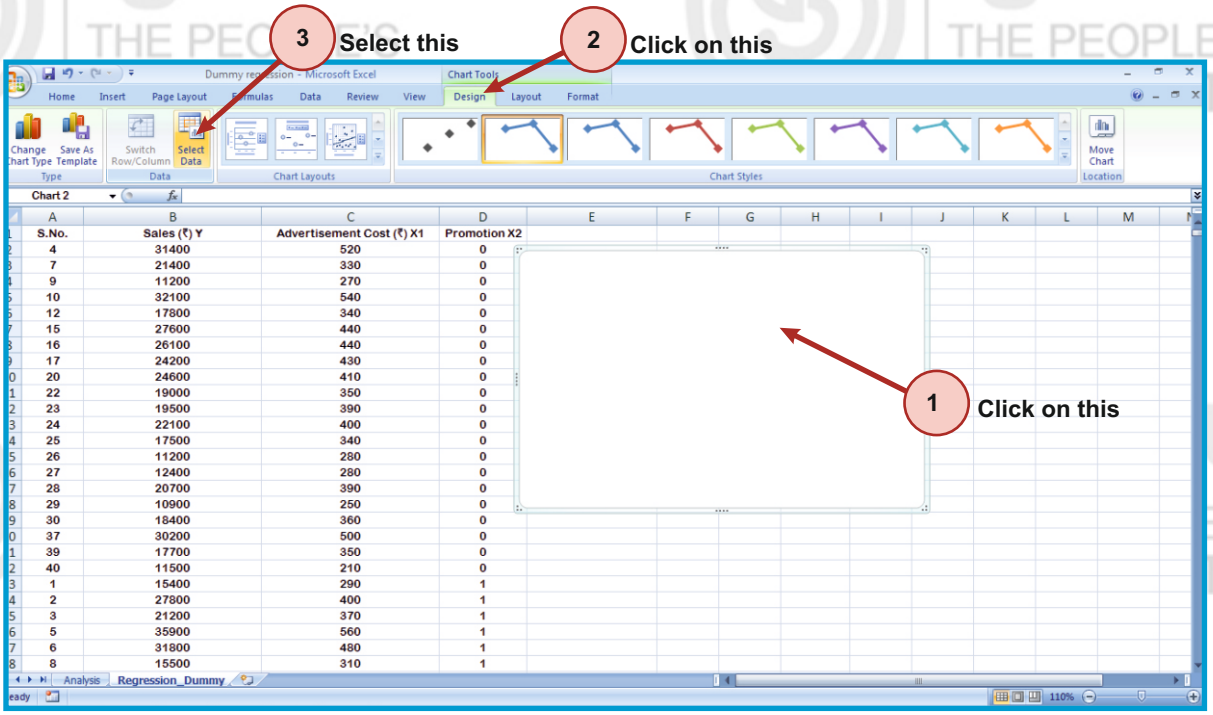


Fig. 13.4

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.

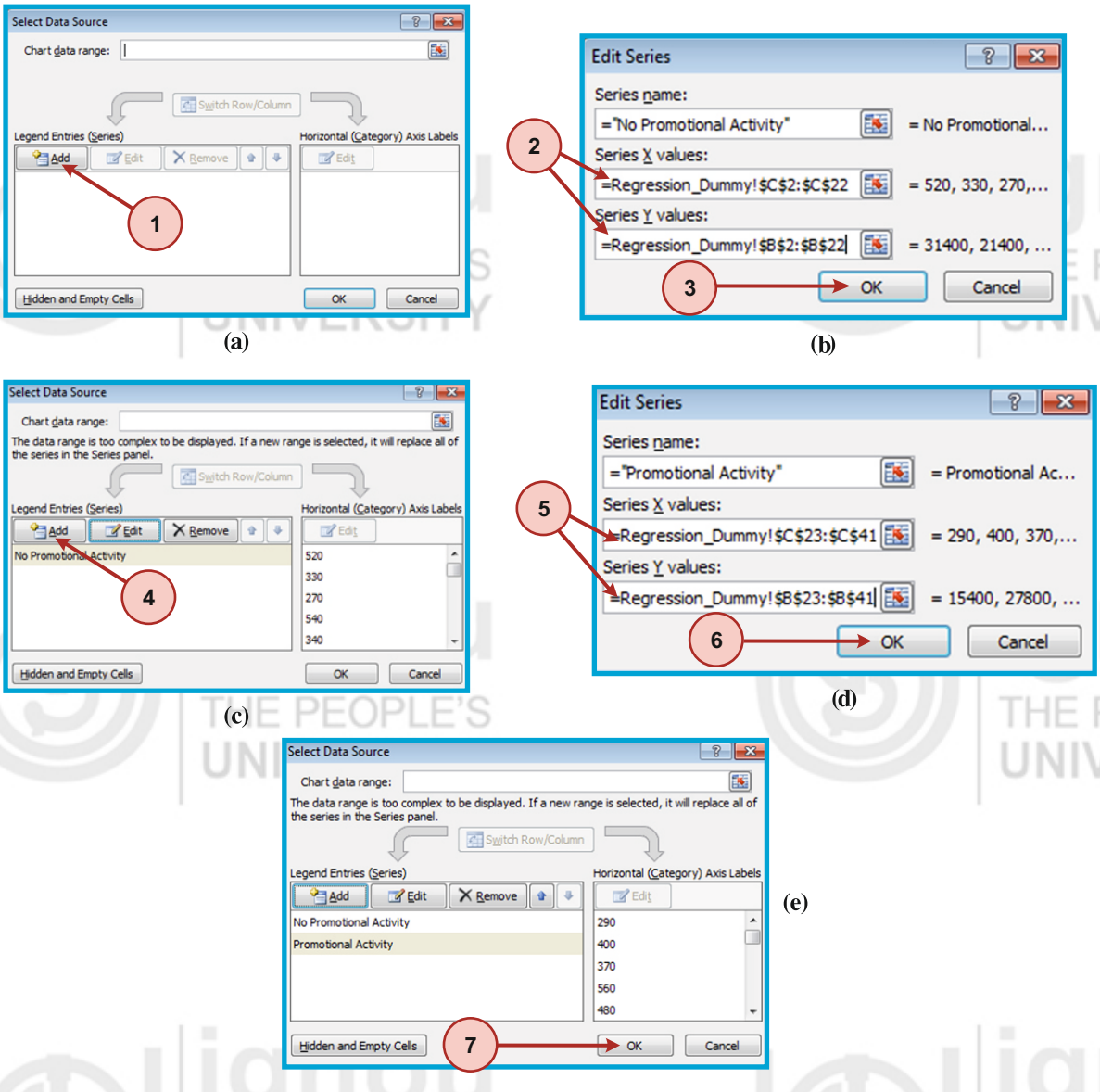


Fig. 13.5

The resulting chart is shown in Fig. 13.6.

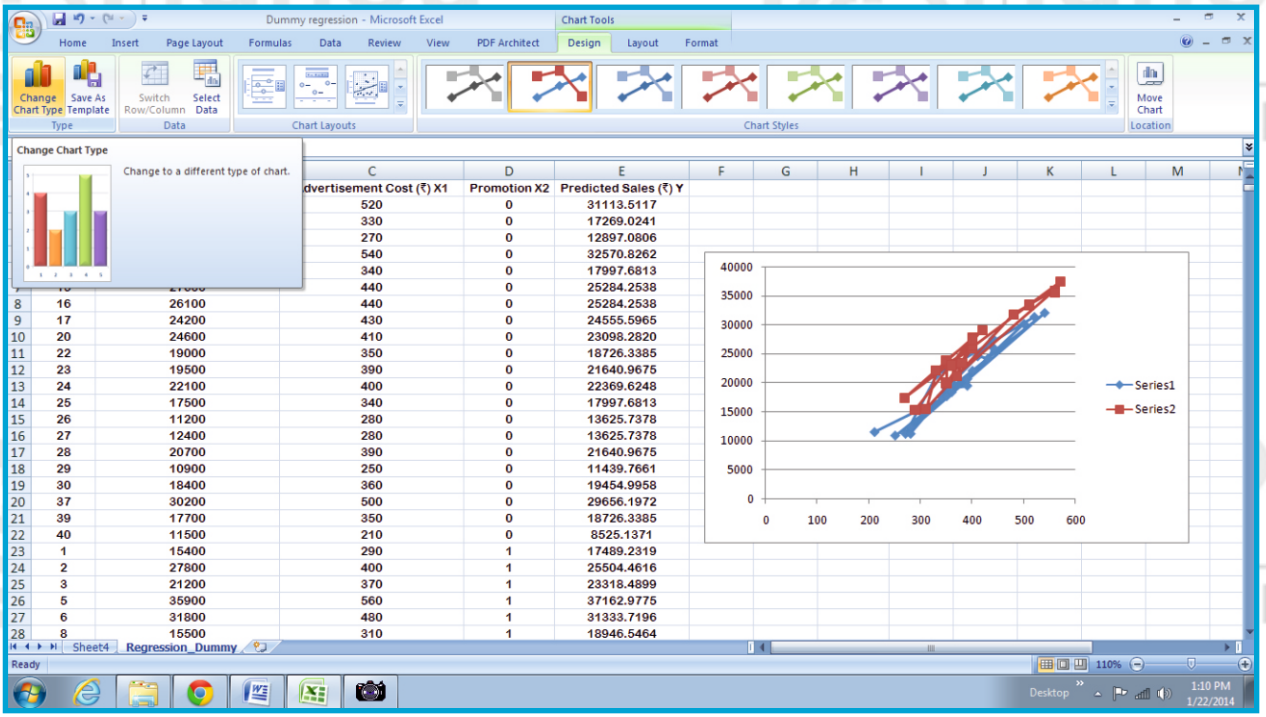


Fig. 13.6

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.

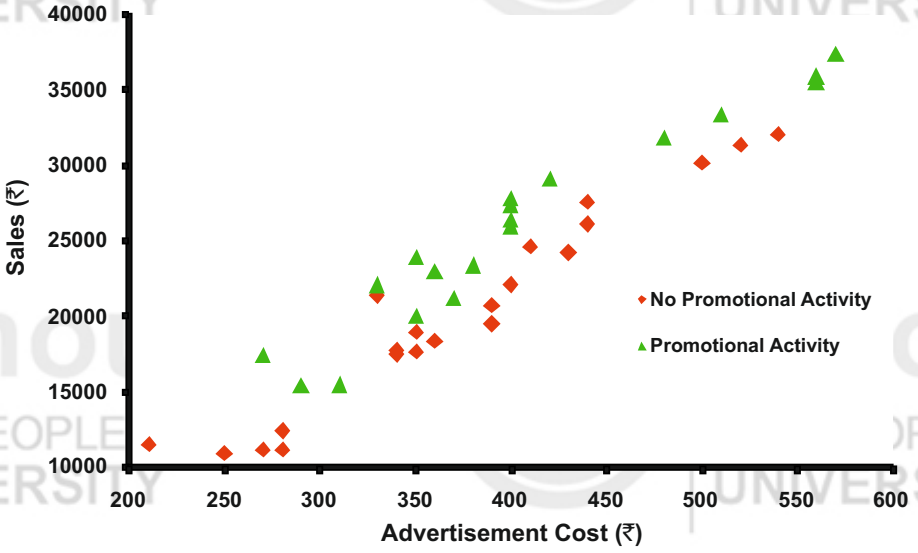


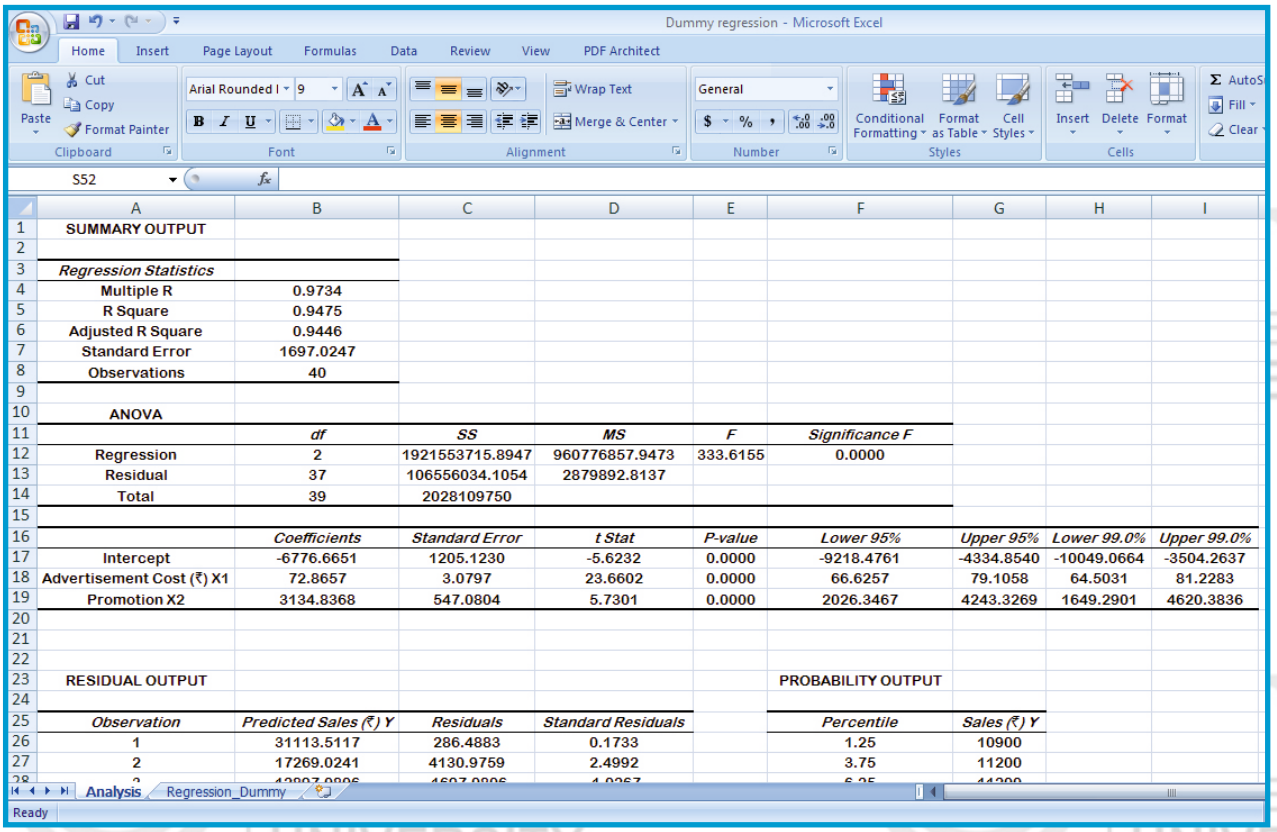
Fig. 13.7

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.

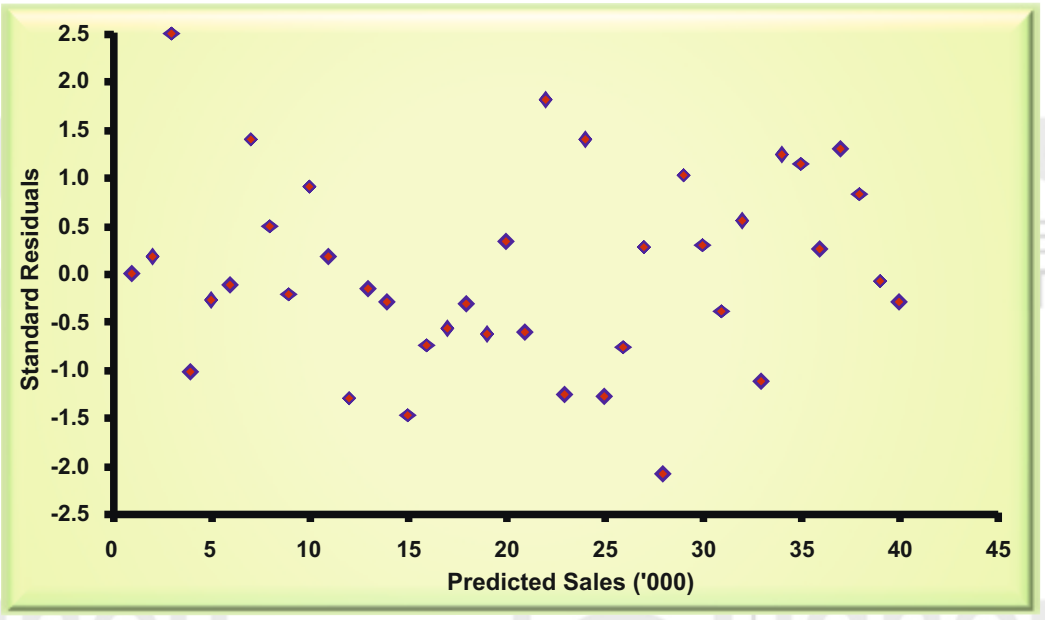


Fig. 13.9: Residual Plot.

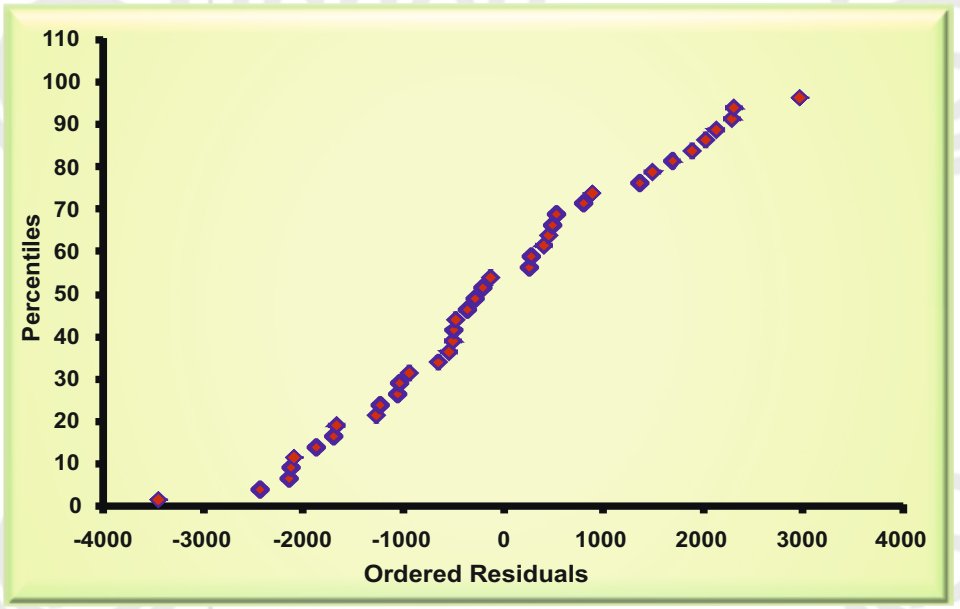


Fig. 13.10: Normal Probability Plot.

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,

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

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

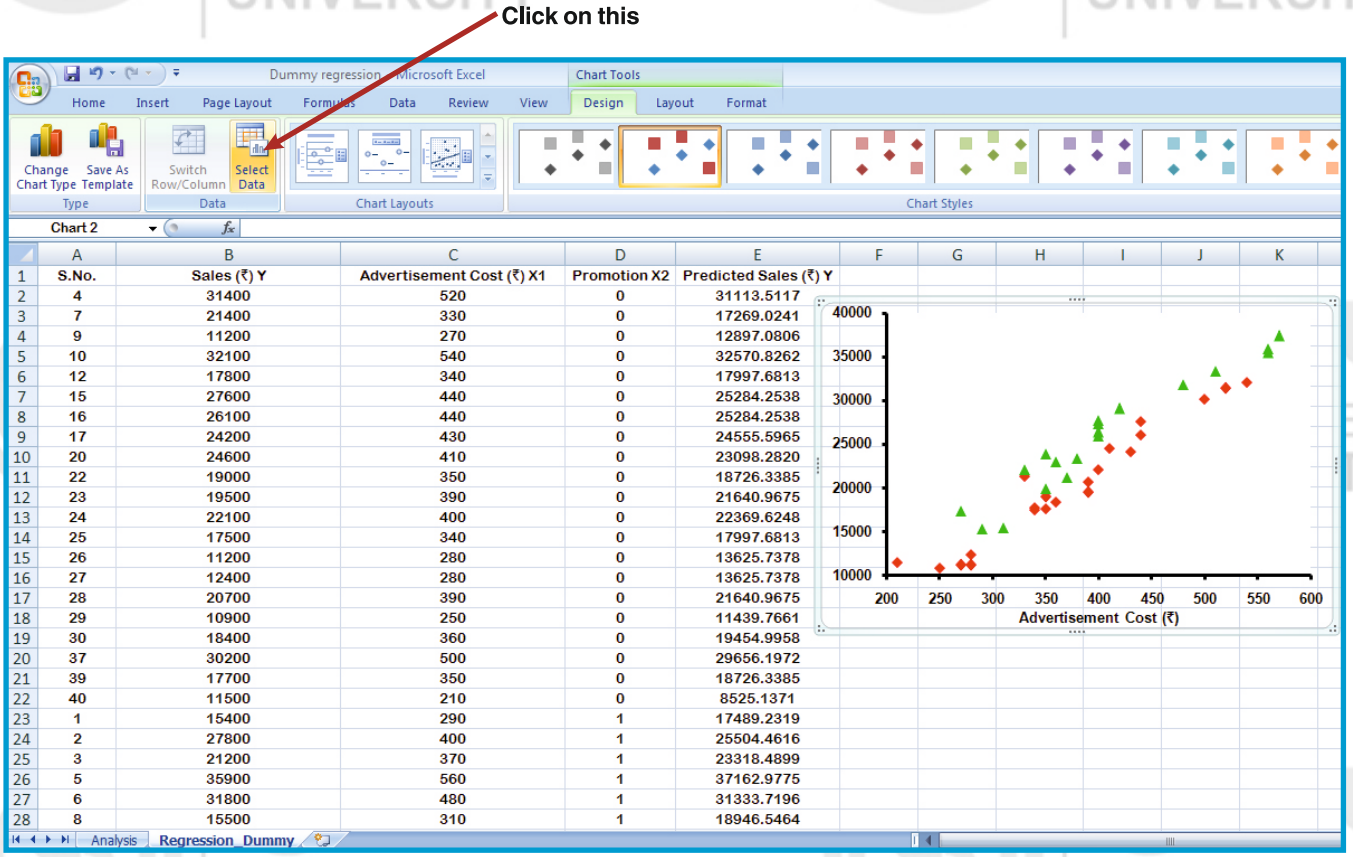
This model represents two different regression equations for different categories of promotional activities as given below:

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.

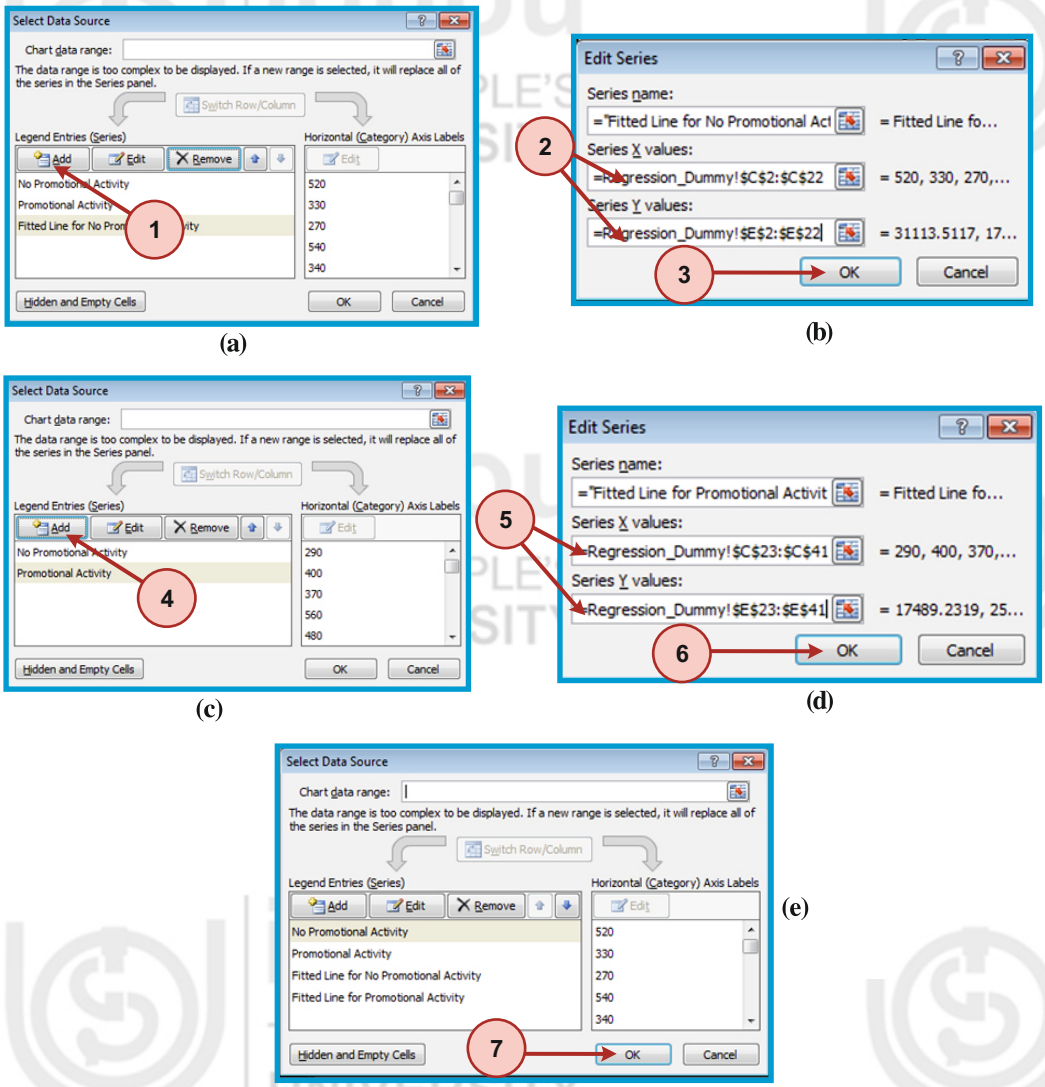


Fig. 13.12

The resulting chart is shown in Fig. 13.13.

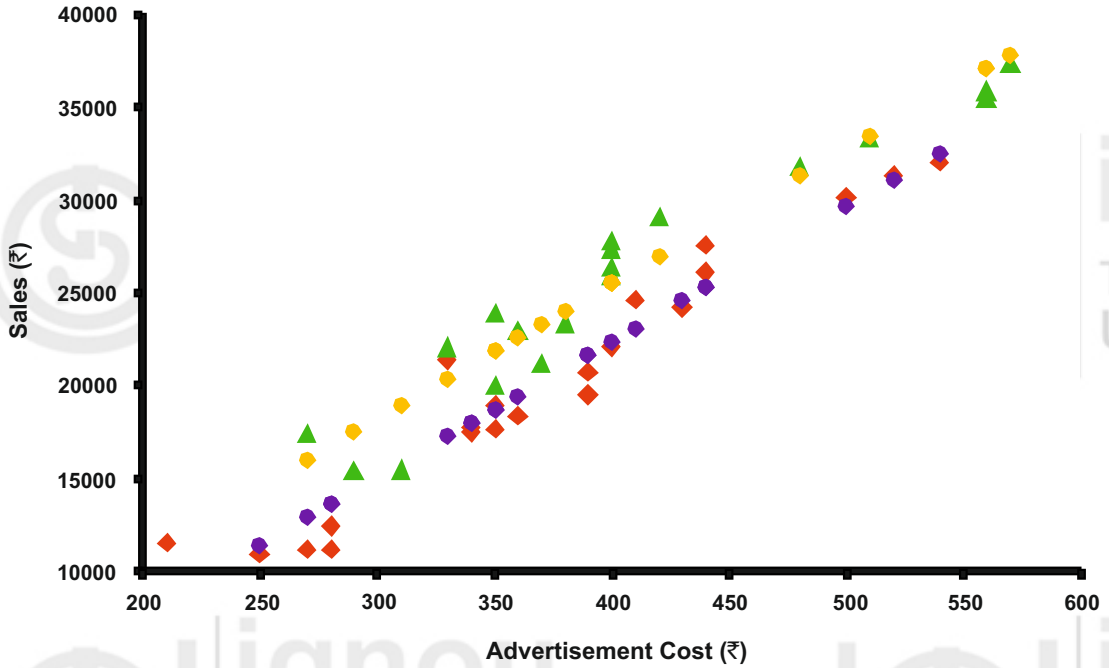
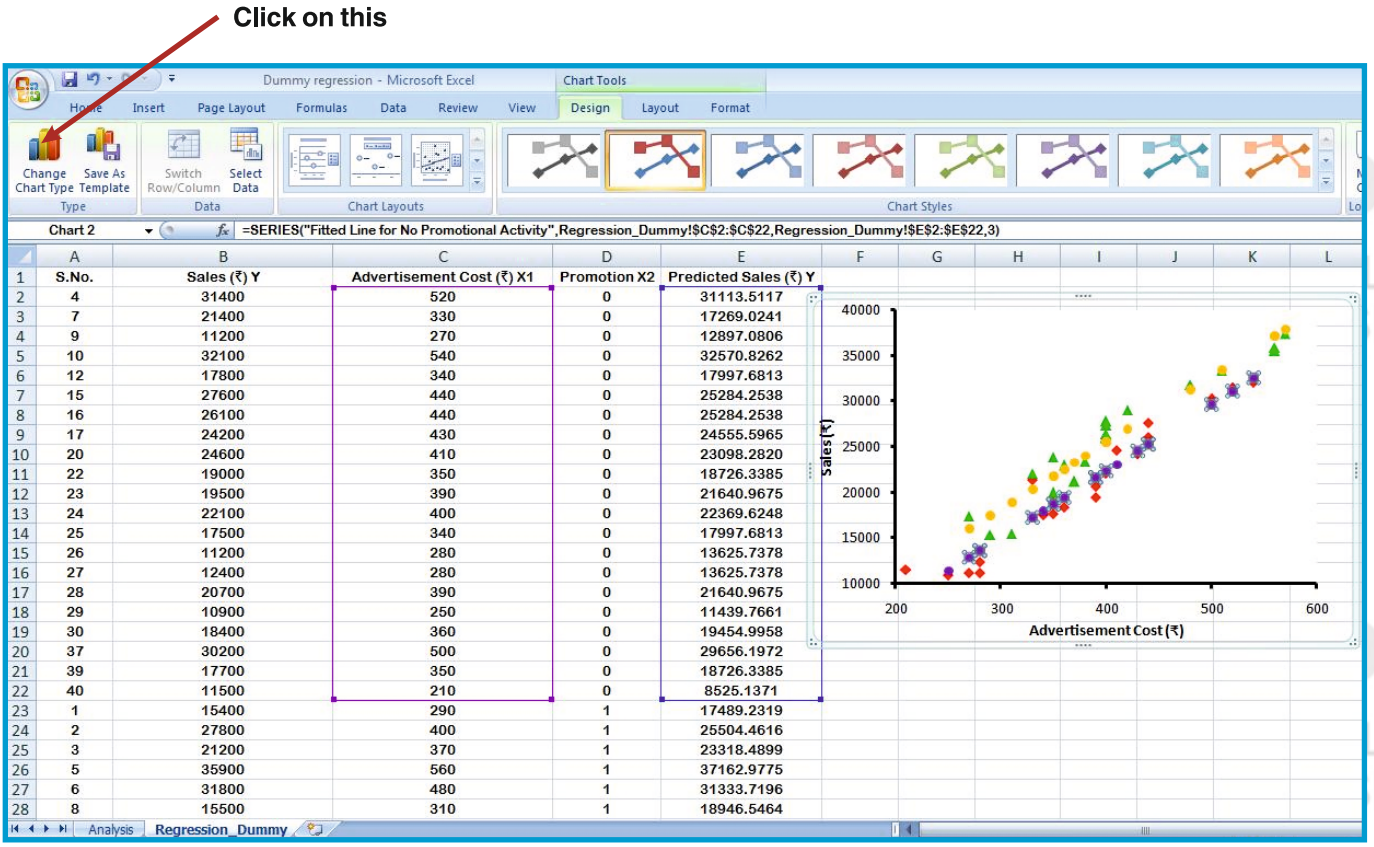


Fig. 13.13

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.



	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. 13.14

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

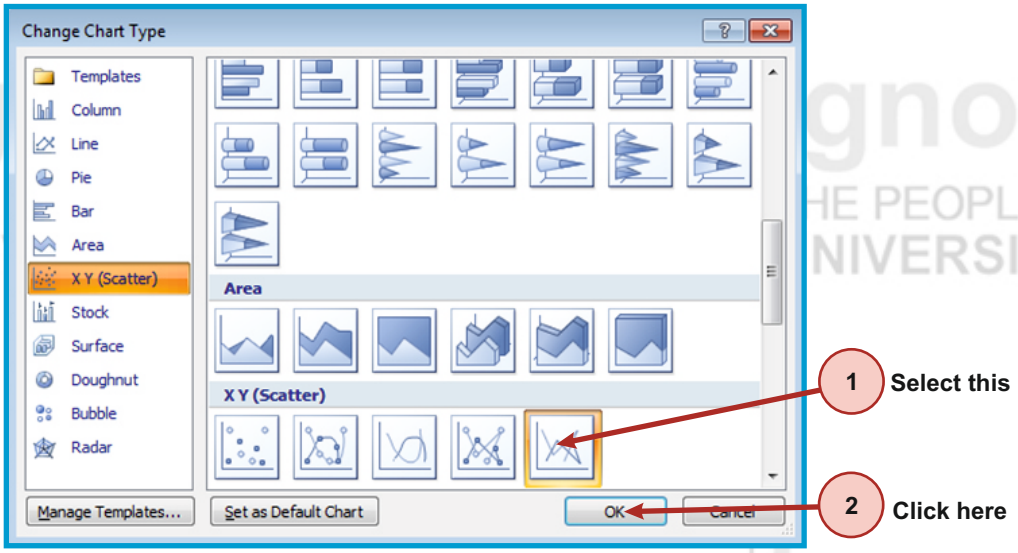


Fig. 13.15

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.

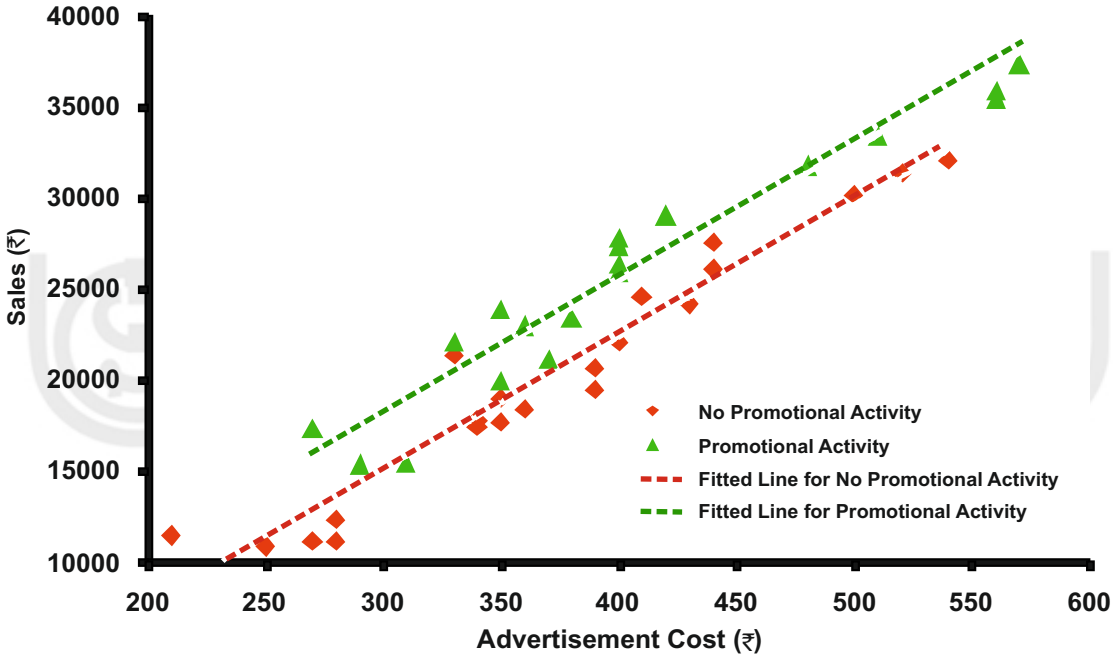


Fig. 13.16

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.