and Taiwan are the world leaders in the international bicycle market cannot be ignored. Indian players have to focus on research and design development in order to face the future challenges.

1. Suppose a leading bicycle manufacturer has divided its products into six brands. Price of these brands and unit sold for 2005 and 2006 are shown in Table 13.01. Use the techniques presented in this chapter and examine whether the distribution of unit sales has changed from 2005 –2006.

TABLE 13.01
Prices of bicycle brands and units sold by a leading bicycle manufacturer in 2005 and 2006

Brand	*Price category (in rupees)*	*2005 (in thousands)*	*2006 (in thousands)*
1	Less than 1200	110	120
2	1200–1400	95	105
3	1400–1800	105	102
4	1800–2000	102	98
5	2000–2200	90	102
6	2200–2500	80	88

2. Suppose Hero Cycles has launched three brands—Hero Premium, Hero Passion, and Hero Smart. Let us assume the Vice President (Sales) of the Hero Cycles company wants to determine whether the sales of bicycle brands are independent of age group. He has appointed a marketing researcher for this purpose. This researcher has taken a random sample of the consumers who have purchased bicycles in 2005. The market researcher has conducted a survey for analysing the consumer preference for the three brands of bicycles. The researcher has also divided the age groups into four categories; 05 to 07, 07 to 09, 09 to 12, and 12 to 17. The observations made by the researcher are given in Table 13.02:

TABLE 13.02
Consumer preference for three leading bicycle brands

Brand / *Age group*	*Hero premium*	*Hero passion*	*Hero smart*	*Row total*
05 to 07	20	25	32	77
07 to 09	10	20	22	52
09 to 12	15	12	10	37
12 to 17	25	22	23	70
Column total	70	79	87	236

Determine whether brand preference is independent of age group. Use $\alpha = 0.05$.

NOTES

1. www.indiastat.com, accessed September 2008, reproduced with permission.
2. www.herocycles.com/about.php,accessedSeptember2008.
3. www.hindubusinessline.com/catalyst/2004/05/20/stories /2004052000120100.htm, accessed September 2008.

CHAPTER **14**

Simple Linear Regression Analysis

A statistical analysis, properly conducted, is a delicate dissection of uncertainties, a surgery of suppositions.

– M. J. Moroney

LEARNING OBJECTIVES

Upon completion of this chapter, you will be able to:

- Use the simple linear regression equation
- Understand the concept of measures of variation, coefficient of determination, and standard error of the estimate
- Understand and use residual analysis for testing the assumptions of regression
- Measure autocorrelation by using the Durbin–Watson statistic
- Understand statistical inference about slope, correlation coefficient of the regression model, and testing the overall model

STATISTICS IN ACTION: TATA STEEL

Tata Steel, established in 1907, is the world's sixth-largest steel company with an existing annual crude steel capacity of 30 million tonnes. It is Asia's first integrated steel plant and India's largest integrated private-sector steel company with operations in 26 countries and commercial presence in 50 countries.[1]

In line with its vision of becoming a global company with a 50 million tonne steel capacity by 2015, the company has expanded through the acquisition route. Tracing the company's history of inorganic growth in recent years, Tata Steel acquired Natsteel in February 2005 and Millennium Steel Company renaming it as Tata Steel Thailand in April 2006. In April 2007, the company acquired Corus, the second-largest steel producer in Europe and the ninth-largest steel producer in the world for USD 13.7 billion. With the acquisition of Corus, Tata Steel has become the world's sixth-largest steel company.[2] Tata Steel made its maiden entry in the list of Global 500 Companies released by *Fortune* in 2008. Table 14.1 shows the sales volumes and marketing expenses of Tata Steel from 1995 to 2007.

The sales volume of the company has increased over the years. The increase in marketing expenses (includes commissions, rebates, discounts, sales promotional expenses on direct selling agents, and entertainment expenses) could be one of the factors that have contributed to the increasing sales. A researcher may like to analyse the relationship between sales and marketing expenses. If there is a relationship, what is

TABLE 14.1
Sales volumes and marketing expenses of Tata Steel from 1995–2007

Year	Sales (in million rupees)	Marketing expenses (in million rupees)
1995	46,274.1	576.4
1996	58,541.2	571.5
1997	63,485.0	916.8
1998	64,292.7	781.4
1999	55,160.0	747.9
2000	61,562.8	895.6
2001	71,966.3	332.2
2002	75,954.1	709.3
2003	97,884.9	871.9
2004	119,178.8	819
2005	158,676.2	861.8
2006	171,329.4	807.5
2007	197,711.9	647.1

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed September 2008, reproduced with permission.

the proportion of change in sales that can be attributed to marketing expenses? How can we develop a model to predict the relationship between sales volume and marketing expenses? This chapter focuses on the answer to all these questions. The chapter focuses on the concept of simple linear regression equation measures of variation, coefficient of determination, standard error of the estimate and the use of residual analysis for testing the assumptions of regression. The chapter also deals with the concept of autocorrelation by using the Durbin–Watson statistic and explains the understanding of statistical inference about slope, correlation coefficient of the regression model, and testing the overall model.

14.1 INTRODUCTION

In many business situations, it has been observed that decision making is based upon the understanding of the relationship between two or more variables. For example, a sales manager might be interested in knowing the impact of advertising on sales. Here, advertising can be considered as an independent variable and sales can be considered as the dependent variable. This is an example of simple linear regression where a single independent variable is used to predict a single numerical dependent variable.

The meaning of the term regression is "stepping back towards the average." The term "regression" was first used by Sir Francis Gatton in 1877. His study on the height of one thousand fathers and sons exhibited an interesting result. He found that tall fathers tend to have tall sons and short fathers tend to have short sons. However, the average height of the sons of a group of tall fathers was less than that of the fathers, and the average height of the sons of a group of short fathers was greater than that of the fathers. Gatton concluded that abnormally tall or short parents tend to "regress" or "step-back" to the average population height.

14.2 INTRODUCTION TO SIMPLE LINEAR REGRESSION

Regression analysis is the process of developing a statistical model, which is used to predict the value of a dependent variable by at least one independent variable. In simple linear regression analysis, there are two types of variables. The variable whose value is influenced or is to be predicted is called the **dependent variable** and the variable which influences the value or is used for prediction is called the **independent variable**. In regression analysis, the independent variable is also known as regressor or predictor or explanatory while the dependent variable is also known as regressed or explained variable. In a simple linear regression analysis, only a straight line relationship between two variables is examined. In fact, simple linear regression analysis is focused on developing a regression model by which the value of the dependent variable can be predicted with the help of the independent variable, based on the linear relationship between these two. This does not mean that the value of a dependent variable cannot be predicted with the help of a group of independent variables. This concept will be discussed in the next chapter (Chapter 15). In the next chapter, we will focus on non-linear relationship and regression models with more than one independent variable. Determining the impact of advertisement on sales is an example of simple linear regression. Determining the impact of other variables such as personal selling, distribution support and advertisement on sales in an example of multiple regression.

Regression analysis is the process of developing a statistical model, which is used to predict the value of a dependent variable by at least one independent variable. In simple linear regression analysis, there are two types of variables. The variable whose value is influenced or to be predicted is called dependent variable and the variable which influences the value or is used for prediction is called independent variable.

In regression analysis, independent variable is also known as regressor or predictor, or explanatory while the dependent variable is also known as regressed or explained variable. In a simple linear regression analysis, only a straight line relationship between two variables is examined.

14.3 DETERMINING THE EQUATION OF A REGRESSION LINE

Simple linear regression is based on the slope–intercept equation of a line. This equation is given as

$$y = ax + b$$

where a is the slope of the line and b the y intercept of the line.

The straight line regression model with respect to population parameters β_0 and β_1 can be given as

$$y = \beta_0 + \beta_1 x$$

where β_0 is the population y intercept which represents the average value of the dependent variable when $x = 0$ and β_1 the slope of the regression line which indicates expected change in the value of y for per unit change in the value of x.

In case of specific dependent variable y_i

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

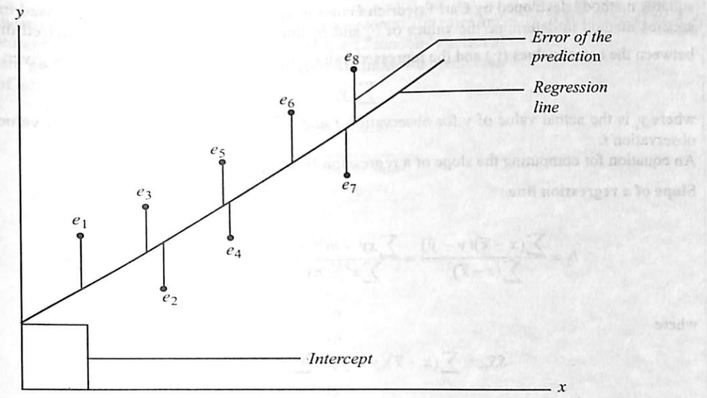


FIGURE 14.1
Error in simple regression

where β_0 is the population y intercept, β_1 the slope of the regression line, y_i the value of the dependent variable for ith value, x_i the value of the independent variable for ith value, and ε_i the random error in y for observation i (ε is the Greek letter *epsilon*).

ε is the error of the regression line in fitting the points of the regression equation. If a point is on the regression line, the corresponding value of ε is equal to zero. If the point is not on the regression line, the value of ε measures the error. This concept leads to two models in regression; deterministic model and probabilistic model.

ε is the error of the regression line in fitting the points of the regression equation. If a point is on the regression line, the corresponding value of ε is equal to zero. If the point is not on the regression line, the value of ε measures the error.

A deterministic model is given as

$$y = \beta_0 + \beta_1 x$$

A probabilistic model is given as

$$y = \beta_0 + \beta_1 x + \varepsilon$$

It can be noticed that in the deterministic model, all the points are assumed to be on the regression line and hence, in all the cases random error ε is equal to zero. Probabilistic model includes an error term which allows the value of y to vary for any given value of x. Figure 14.1 presents error in simple regression.

It can be noticed that in the deterministic model, all the points are assumed to be on the regression line and hence, in all the cases random error ε is equal to zero. Probabilistic model includes an error term which allows the value of y to vary for any given value of x.

In order to predict the value of y, a researcher has to calculate the value of β_0 and β_1. In this process, difficulty occurs in terms of observing the entire population. This difficulty can be handled by taking a sample data and ultimately developing a sample regression model. This sample regression model can be used to make predictions about population parameters. So, β_0 and β_1 (population parameters) are estimated on the basis of the sample statistics b_0 and b_1. Thus, the simple regression equation (based on samples) is used to estimate the linear regression model.

The equation of the simple regression line is given as

$$\hat{y} = b_0 + b_1 x$$

where b_0 is the sample y intercept which represent the average value of the dependent variable when $x = 0$ and b_1 the slope of the sample regression line, which indicates expected change in the value of y for per unit change in the value of x.

For determining the equation of the simple regression line, values of b_0 (sample y intercept) and b_1 (slope of the sample regression line) must be determined. Once b_0 and b_1 are determined, a researcher can plot a straight line and the comparison of this straight line with the original data can be performed very easily. The main focus of simple regression analysis is on finding the straight line that fits the data best. In other words, we need to minimize the difference between the actual values (y_i) and the regressed values ($\hat{y}_i$). This difference between the actual values (y_i) and the regressed values ($\hat{y}_i$) is referred to as residual (ε). In order to minimize this difference, a mathematical technique "least-

The main focus of the simple regression analysis is on finding the straight line that fits the data best. In other words, we need to minimize the difference between the actual values (y_i) and the regressed values ($\hat{y}_i$). This difference between the actual values (y_i) and the regressed values ($\hat{y}_i$) is referred to as residual (ε).

The sample data are used in the least squares method to determine the values of b_0 and b_1 that minimizes the sum of squared differences between the actual values (y_i) and the regressed values ($\hat{y}_i$).

squares method" developed by Carl Friedrich Gauss is applied. The sample data are used in the least squares method to determine the values of b_0 and b_1 that minimizes the sum of squared differences between the actual values (y_i) and the regressed values ($\hat{y}_i$). Least squares criterion is given by

$$\sum (y_i - \hat{y}_i)^2$$

where y_i is the actual value of y for observation i and ($\hat{y}_i$) the regressed (predicted) value of y for observation i.

An equation for computing the slope of a regression line is given below:

Slope of a regression line

$$b_1 = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} = \frac{\sum xy - n(\bar{x} \times \bar{y})}{\sum x^2 - n\bar{x}^2} = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

where

$$SS_{xy} = \sum (x-\bar{x})(y-\bar{y}) = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

and

$$SS_{xx} = \sum (x-\bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$b_1 = \frac{SS_{xy}}{SS_{xx}}$$

The sample y intercept of the regression line is given as

$$b_0 = \bar{y} - b_1\bar{x} = \frac{\sum y}{n} - b_1\frac{(\sum x)}{n}$$

It has already been discussed that in the estimation process through a simple linear regression, unknown population parameters, β_0 and β_1, are estimated by sample statistics b_0 and b_1. Figure 14.2 exhibits the summary of the estimation process for simple linear regression.

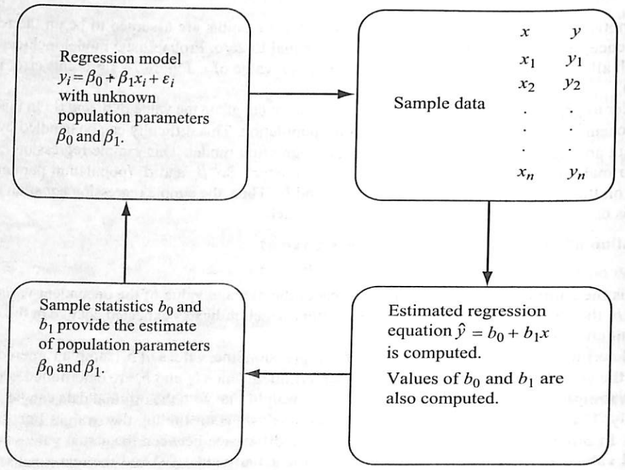


FIGURE 14.2
Summary of the estimation process for simple linear regression.

Example 14.1

A cable wire company has spent heavily on advertisements. The sales and advertisement expenses (in thousand rupees) for the 12 randomly selected months are given in Table 14.2. Develop a regression model to predict the impact of advertisement on sales.

TABLE 14.2
Sales and advertisement expenses (in thousand rupees) of a cable wire company

Months	*Advertisement (in thousand rupees)*	*Sales (in thousand rupees)*
Jan	92	930
Feb	94	900
Mar	97	1020
Apr	98	990
May	100	1100
Jun	102	1050
Jul	104	1150
Aug	105	1120
Sep	105	1130
Oct	107	1200
Nov	107	1250
Dec	110	1220

Solution

The first step is to determine whether the relationship between two variables is linear. For doing this, a scatter plot, drawn by any of the statistical software programs (MS Excel, Minitab, or SPSS) can be used. Figure 14.3 is the scatter plot produced using Minitab.

Scatter plot (Figure 14.3) exhibits the linear relationship between sales and advertisement. After this linear relationship is confirmed, further steps for developing a linear regression model can be adopted. For computing the regression coefficient, b_0 and b_1, the values of Σx, Σy, Σx^2, and Σxy must be determined. Sales is a dependent variable and advertisement is an independent variable.

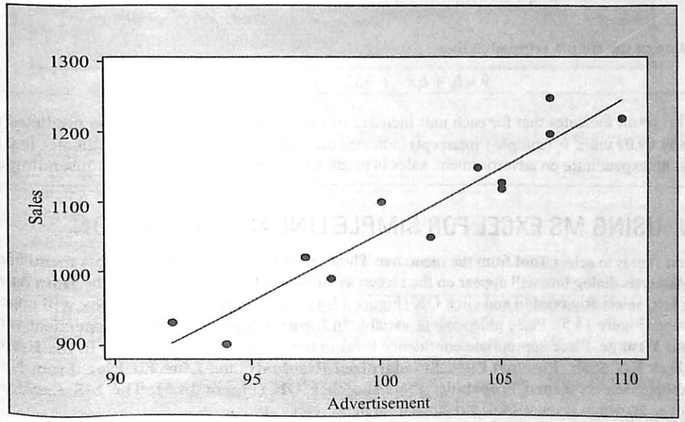


FIGURE 14.3
Scatter plot between sales and advertisement produced using Minitab

Computation of Σx, Σy, Σx^2, and Σxy for Example 14.1

Months	*Advertisement (in thousand rupees): x*	*Sales (in thousand rupees): y*	x^2	xy
Jan	92	930	8464	85,560
Feb	94	900	8836	84,600
Mar	97	1020	9409	98,940
Apr	98	990	9604	97,020
May	100	1100	10,000	110,000
Jun	102	1050	10,404	107,100
Jul	104	1150	10,816	119,600
Aug	105	1120	11,025	117,600
Sep	105	1130	11,025	118,650
Oct	107	1200	11,449	128,400
Nov	107	1250	11,449	133,750
Dec	110	1220	12,100	134,200
	$\Sigma x = 1221$	$\Sigma y = 13,060$	$\Sigma x^2 = 124,581$	$\Sigma xy = 1,335,420$

$$SS_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} = 1,335,420 - \frac{(1221) \times (13,060)}{12} = 6565$$

$$SS_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 124,581 - \frac{(1221)^2}{12} = 344.25$$

$$b_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{6565}{344.25} = 19.0704$$

$$b_0 = \frac{\sum y}{n} - b_1 \frac{(\sum x)}{n} = \frac{13,060}{12} - (19.0704) \times \frac{1221}{12} = -852.08$$

Equation of the simple regression line

$$\hat{y} = b_0 + b_1 x = (-852.08) + (19.07)x$$

This result indicates that for each unit increase in x (advertisement), y (sales) is predicted to increase by 19.07 units. b_0 (sample y intercept) indicates the value of y when $x = 0$. It indicates that when there is no expenditure on advertisement, sales is predicted to decrease by 852.08 thousand rupees .

14.4 USING MS EXCEL FOR SIMPLE LINEAR REGRESSION

The first step is to select **Tool** from the menu bar. Then select **Data Analysis** from this menu bar. The **Data Analysis** dialog box will appear on the screen as shown in Figure 14.4. From the **Data Analysis** dialog box, select **Regression** and click **OK** (Figure 14.4). The **Regression** dialog box will appear on the screen (Figure 14.5). Place independent variable in **Input X Range** and place dependent variable in **Input Y range**. Place appropriate confidence level in the **Confidence level** box. In the **Residuals** box, check **Residuals**, **Residual Plots**, **Standardized Residuals**, and **Line Fit Plot**. From **Normal Probability**, select **Normal Probability Plots** and click **OK** (Figure 14.5). The MS Excel output (partial) as shown in (Figure 14.6) will appear on the screen.

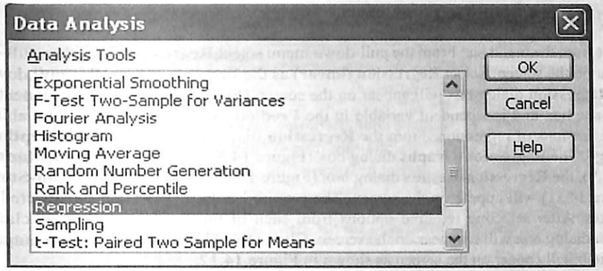


FIGURE 14.4
MS Excel Data Analysis dialog box

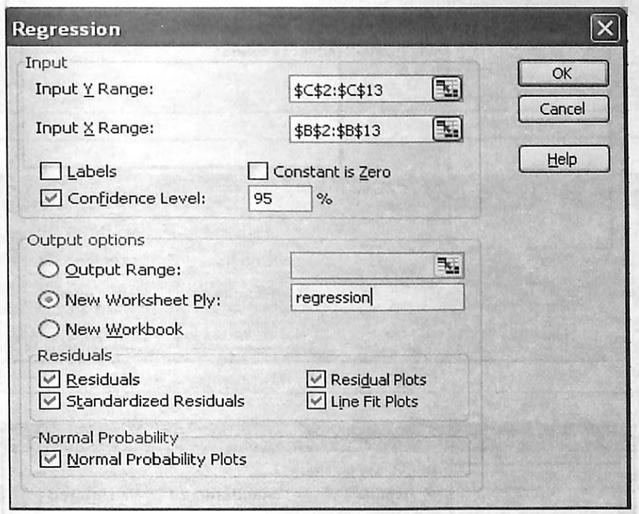


FIGURE 14.5
MS Excel Regression dialog box

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.949166574					
5	R Square	0.900917166					
6	Adjusted R Square	0.891008904					
7	Standard Error	37.10688403					
8	Observations	12					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	125197.4582	125197.4582	90.92568	2.45382E-06	
13	Residual	10	13769.20842	1376.920842			
14	Total	11	138966.6667				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	-852.0842411	203.7758887	-4.181477243	0.001883	-1306.125214	-398.04327
18	X Variable 1	19.07044299	1.999942514	9.535495577	2.45E-06	14.61429339	23.5265926

FIGURE 14.6
MS Excel output (partial) for Example 14.1

14.5 USING MINITAB FOR SIMPLE LINEAR REGRESSION

Select **Stat** from the menu bar. From the pull-down menu select **Regression.** Another pull-down menu will appear on the screen. Select **Regression (linear)** as the first option from this pull down menu.

The **Regression** dialog box will appear on the screen (Figure 14.7). Place dependent variable in the **Response** box and independent variable in the **Predictors** box. Minitab has the ability to open various dimensions of regression. From the **Regression** dialog box, click **Graph**, **Options**, **Result**, and **Storage**. The **Regression-Graphs** dialog box (Figure 14.8), the **Regression-Options** dialog box (Figure 14.9), the **Regression-Results** dialog box (Figure 14.10), and the **Regression-Storage** dialog box (Figure 14.11) will appear on the screen. The required output range can be selected from these dialog boxes. After selecting required options from each of the four dialog boxes, click **OK**. The **Regression** dialog box will reappear on the screen. Click **OK**. The partial regression output produced using Minitab will appear on the screen as shown in Figure 14.12.

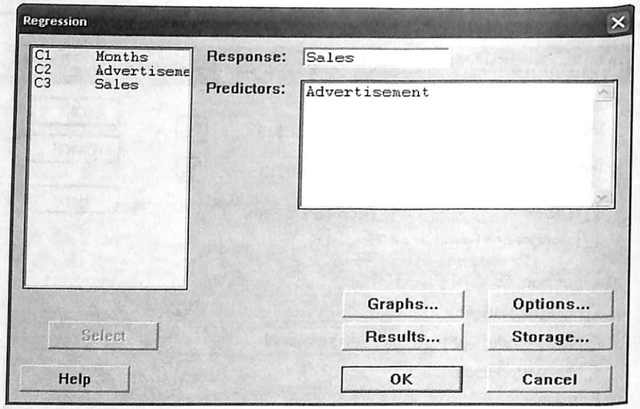


FIGURE 14.7
Minitab Regression dialog box

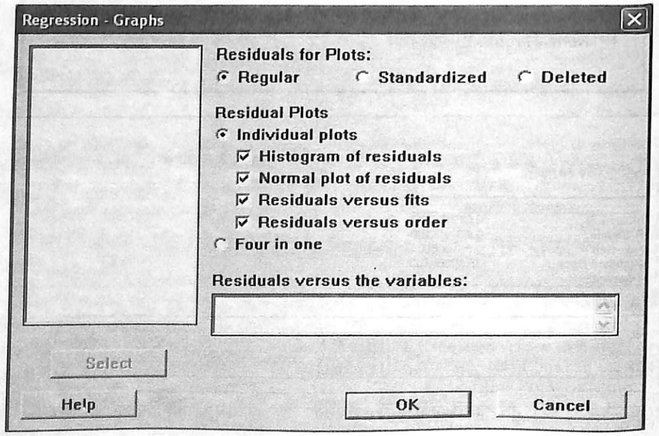


FIGURE 14.8
Minitab Regression-Graphs dialog box

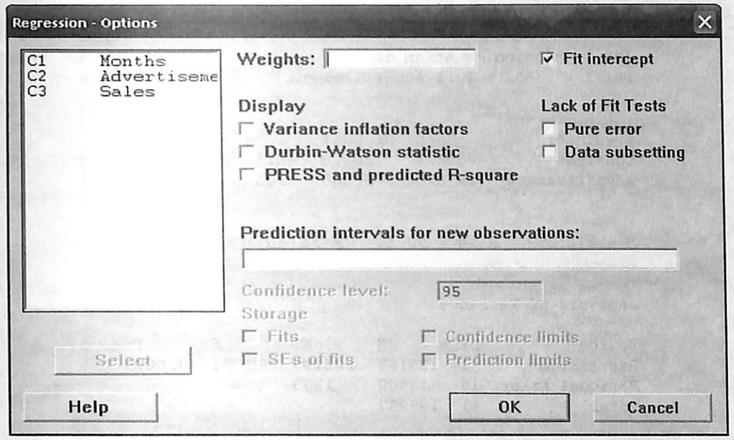


FIGURE 14.9
Minitab Regression-Options dialog box

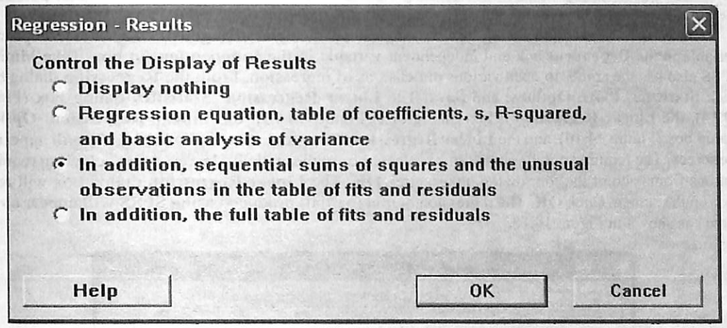


FIGURE 14.10
Minitab Regression-Results dialog box

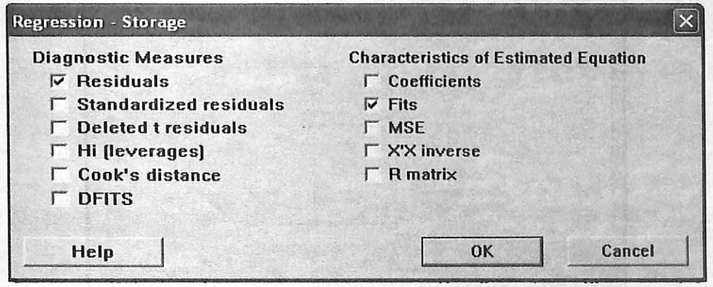


FIGURE 14.11
Minitab Regression-Storage dialog box

Regression Analysis: Sales versus Advertisement

```
The regression equation is
Sales = - 852 + 19.1 Advertisement


Predictor         Coef  SE Coef      T      P
Constant       -852.1    203.8  -4.18  0.002
Advertisement  19.070    2.000   9.54  0.000


S = 37.1069   R-Sq = 90.1%   R-Sq(adj) = 89.1%


Analysis of Variance

Source          DF      SS      MS      F      P
Regression       1  125197  125197  90.93  0.000
Residual Error  10   13769    1377
Total           11  138967
```

FIGURE 14.12
Minitab output (partial) for Example 14.1

14.6 USING SPSS FOR SIMPLE LINEAR REGRESSION

Select **Analyze** from the menu bar. Select **Regression** from the pull-down menu. Another pull-down menu will appear on the screen. Select **Linear** from this menu.

The **Linear Regression** dialog box will appear on the screen (Figure 14.13). Place dependent variable in the **Dependent** box and independent variable in the **Independent(s)** box. Like Minitab, SPSS also has the ability to open various dimensions of regression. From the **Regression** dialog box, click **Statistics**, **Plots**, **Options**, and **Save**. The **Linear Regression**: **Statistics** dialog box (Figure 14.14), the **Linear Regression**: **Plots** dialog box (Figure 14.15), the **Linear Regression**: **Options** dialog box (Figure 14.16), and the **Linear Regression**: **Save** dialog box (Figure 14.17) will appear on the screen. The required output range can be selected from these dialog boxes. After selecting required options from each of the four dialog boxes, click **OK**. The **Linear Regression** dialog box will reappear on the screen. Click **OK**. The regression output (partial) produced using SPSS will appear on the screen as shown in Figure 14.18.

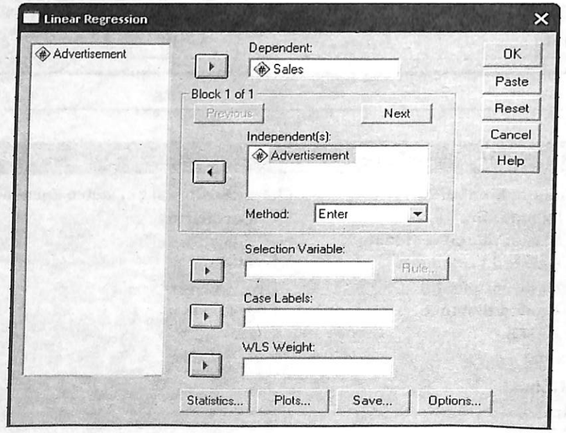


FIGURE 14.13
SPSS Linear Regression dialog box

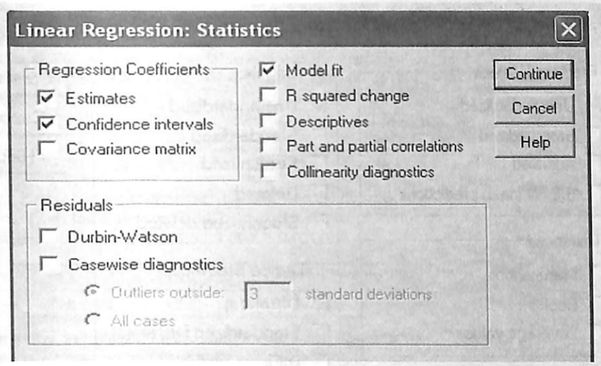


FIGURE 14.14
SPSS Linear Regression: Statistics dialog box

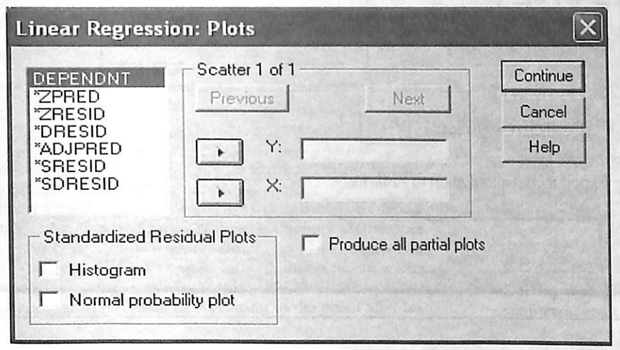


FIGURE 14.15
SPSS Linear Regression: Plots dialog box

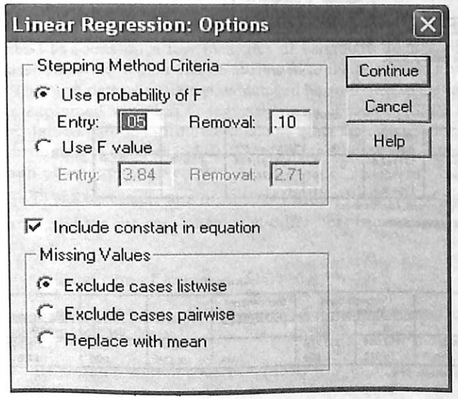


FIGURE 14.16
SPSS Linear Regression: Options dialog box

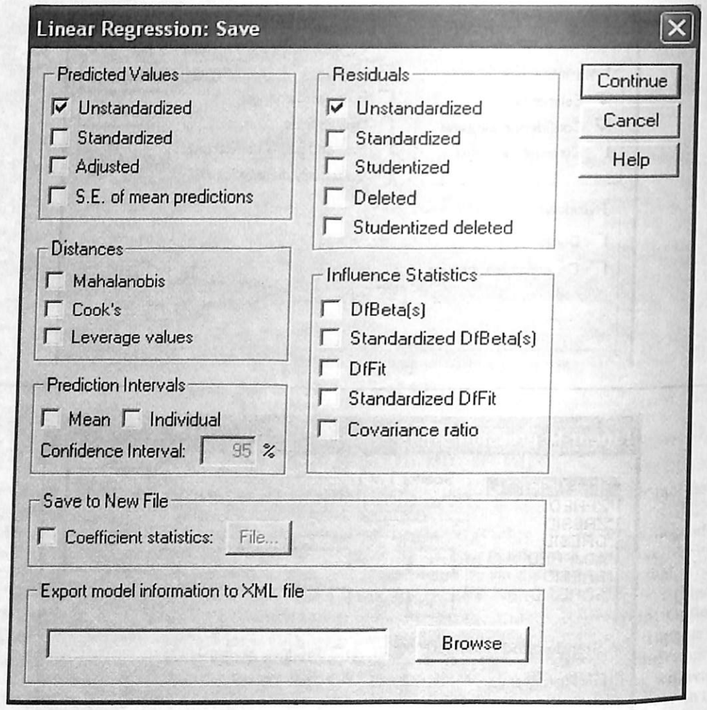


FIGURE 14.17
SPSS Linear Regression: Save dialog box

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.949[a]	.901	.891	37.10688

a. Predictors: (Constant), Advertisement
b. Dependent Variable: Sales

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	125197.5	1	125197.459	90.926	.000[a]
	Residual	13769.208	10	1376.921		
	Total	138966.7	11			

a. Predictors: (Constant), Advertisement
b. Dependent Variable: Sales

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-852.084	203.776		-4.181	.002	-1306.125	-398.043
	Advertisement	19.070	2.000	.949	9.535	.000	14.614	23.527

a. Dependent Variable: Sales

FIGURE 14.18
SPSS output (partial) for Example 14.1

SELF-PRACTICE PROBLEMS

14A1. Taking x as the independent variable and y as the dependent variable from the following data, determine the line of regression. Let $\alpha = 0.05$.

x	12	21	28	25	32	42	43	39	55
y	14	22	12	28	35	37	32	44	49

14A2. Taking x as the independent variable and y as the dependent variable from the following data, construct a scatter plot and determine the line of regression. Let $\alpha = 0.05$.

x	13	18	25	30	22	24	40
y	14	16	17	18	15	22	38

14A3. A company believes that the number of salespersons employed is a good predictor of sales. The following table exhibits sales (in thousand rupees) and number of salespersons employed for different years.

Sales (in thousand rupees)	120	125	118	115	100	130	140	135	130	123
Number of salespersons employed	10	15	12	18	20	21	22	20	15	19

Develop a simple regression model to predict sales based on the number of salespersons employed.

14A4. Cadbury India Ltd, incorporated in 1948, is the wholly owned Indian subsidiary of the UK-based Cadbury Schweppes Plc., which is a global confectionary and beverages company. Cadbury India Ltd operates in India in the segments of chocolates, sugar confectionary, and food drinks.[2] The following table provides data relating to the profit after tax and advertisement of Cadbury India Ltd from 1989–1990 to 2006–2007.

Year	*Advertisement (in million rupees)*	*Profit after tax (in million rupees)*
Mar 1990	73.4	55.5
Mar 1991	101.8	55.1
Mar 1992	99	37.1
Mar 1993	110.9	13.6
Mar 1994	145.3	86.8
Mar 1995	127.7	95.9
Mar 1996	190.3	200.8
Mar 1997	255.9	196.3
Mar 1998	296.2	185.7
Mar 1999	394.1	262.1
Mar 2000	532.8	367
Mar 2001	577.8	520.2
Mar 2002	731.6	574
Mar 2003	876.7	749.1
Mar 2004	904.4	456.5
Mar 2005	910.2	462.1
Mar 2006	958.2	459.6
Mar 2007	1218.5	688.1

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed December 2008, reproduced with permission.

Develop a simple regression line to predict the profit after tax from advertisement.

14.7 MEASURES OF VARIATION

While developing a regression model to predict the dependent variable with the help of the independent variable, we need to focus on a few measures of variations. Total variation (SST) can be partitioned into two parts: variation which can be attributed to the relationship between x and y and unexplained variation. The first part of variation, which can be attributed to the relationship between x and y is referred to as explained variation or regression sum of squares (SSR). The second part of variation, which is unexplained can be attributed to factors other than the relationship between x and y, and is referred to as error sum of squares (SSE). So, in a simple linear regression model, total variation, that is, the total sum of squares is given as:

Total sum of squares (SST) = Regression sum of squares (SSR) + Error sum of squares (SSE)

Total sum of squares (SST) is the sum of squared differences between each observed value (y_i) and the average value of y.

$$\text{Total sum of squares} = (\text{SST}) = \sum (y_i - \bar{y}_i)^2$$

Regression sum of squares (SSR) is the sum of squared differences between regressed (predicted) values and the average value of y.

$$\text{Regression sum of squares} = (\text{SSR}) = \sum (\hat{y}_i - \bar{y})^2$$

While developing a regression model to predict the dependent variable with the help of the independent variable, we need to focus on a few measures of variation. Total variation (SST) can be partitioned into two parts: variation which can be attributed to the relationship between x and y and unexplained variation.

The first part of variation, which can be attributed to the relationship between x and y, is referred to as explained variation or regression sum of squares (SSR). The second part of the variation, which is unexplained can be attributed to factors other than the relationship between x and y, and is referred to as error sum of squares (SSE).

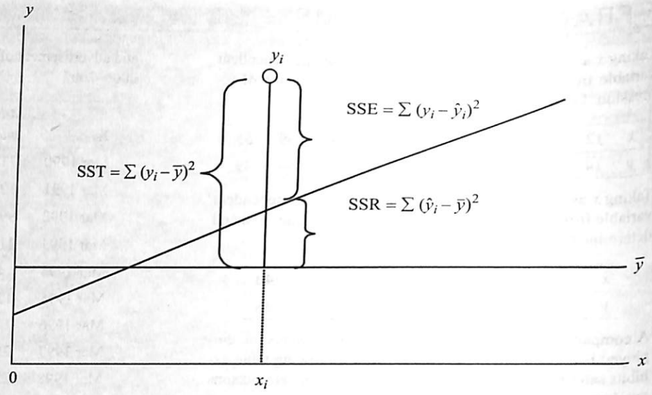
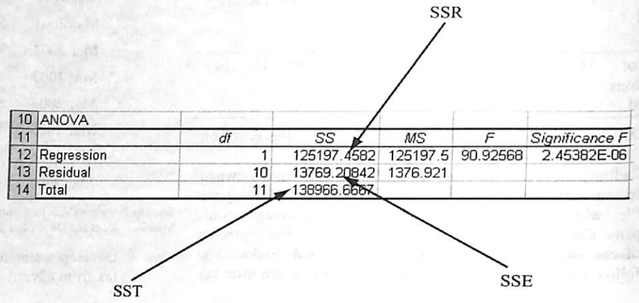


FIGURE 14.19
Measures of variation in simple linear regression



10	ANOVA					
11		*df*	*SS*	*MS*	*F*	*Significance F*
12	Regression	1	125197.4582	125197.5	90.92568	2.45382E-06
13	Residual	10	13769.20842	1376.921		
14	Total	11	138966.6667			

FIGURE 14.20
Values of SST, SSR and SSE for Example 14.1 produced using MS Excel

Error sum of squares (SSE) is the sum of squared differences between each observed value (y_i) and regressed (predicted) value of y.

$$\text{Error sum of squares} = (\text{SSE}) = \sum (y_i - \hat{y}_i)^2$$

Figure 14.19 exhibits the measures of variation in simple linear regression. It can be seen easily that Total sum of squares (SST) = regression sum of squares (SSR) + error sum of squares (SSE), that is, 138,966.6667(SST) = 125,197.4582 (SSR) +13,769.20842 (SSE)

Figure 14.20 is the ANOVA table produced using MS Excel exhibiting values of SST, SSR and SSE and other values for Example 14.1. The same ANOVA table as shown in Figure 14.20 can be obtained using Minitab and SPSS. Figures 14.12 and 14.18 exhibit this ANOVA table containing SST, SSR, and SSE values obtained from Minitab and SPSS, respectively.

14.7.1 Coefficient of Determination

The ratio of regression sum of squares (SSR) to total sum of squares (SST) leads to a very important result which is referred to as coefficient of determination. The values of coefficient of determination ranges from 0 to 1.

Coefficient of determination is a very commonly used measure of fit for regression models and is denoted by r^2. The utility of SST, SSR, and SSE is limited in terms of direct interpretation. The ratio of regression sum of squares (SSR) to total sum of squares (SST) leads to a very important result, which is referred to as coefficient of determination. In a regression model, the coefficient of determination measures the proportion of variation in y that can be attributed to the independent variable x. The values of coefficient of determination range from 0 to 1. Coefficient of determination can be defined as

$$r^2 = \frac{\text{Regression sum of squares}}{\text{Total sum of squares}} = \frac{\text{SSR}}{\text{SST}}$$

In Example 14.1, coefficient of determination r^2 can be calculated as

$$r^2 = \frac{\text{Regression sum of squares}}{\text{Total sum of squares}} = \frac{\text{SSR}}{\text{SST}} = \frac{125{,}197.4582}{138{,}966.6667} = 0.9009$$

As discussed, the coefficient of determination leads to an important interpretation of the regression model. In Example 14.1, r^2 is calculated as 0.9009. This indicates that 90.09% of the variation in sales can be explained by the independent variable, that is, advertisement. This result also explains that 9.91% of the variation in sales is explained by factors other than advertisement.

Figures 14.21, 14.22, and 14.23, are the partial regression outputs from MS Excel, Minitab, and SPSS respectively, exhibiting coefficient of determination and other important results.

14.7.2 Standard Error of the Estimate

A residual is the difference between actual values (y_i) and the regressed values ($\hat{y}_i$), determined by the regression equation for a given value of the independent variable x.

It has already been discussed that sample data are used in the least squares method to determine the values of b_0 and b_1 that minimize the sum of squared differences between the actual values (y_i) and the regressed values ($\hat{y}_i$). Variability in actual values (y_i) and the regressed values ($\hat{y}_i$) is measured in terms of residuals. A residual is the difference between the actual values (y_i) and the regressed values ($\hat{y}_i$), determined by the regression equation for a given value of the independent variable x. The residual around the regression line is given as

$$\text{Residual } (e_i) = \text{actual values } (y_i) - \text{regressed values } (\hat{y}_i)$$

3	*Regression Statistics*		
4	Multiple R	0.949166574	
5	R Square	0.900917186	← r^2(coefficient of determination)
6	Adjusted R Square	0.891008904	
7	Standard Error	37.10688403	← S_{yx}(Standard error)
8	Observations	12	

FIGURE 14.21
Partial regression output from MS Excel showing coefficient of determination and other important results

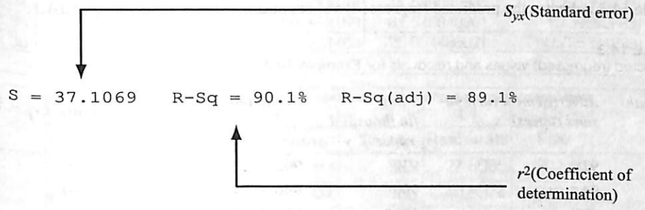


FIGURE 14.22
Partial regression output from Minitab showing coefficient of determination and other important results

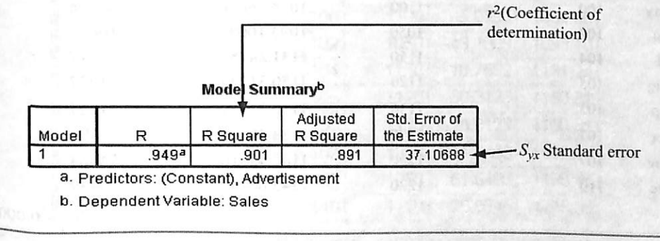


FIGURE 14.23
Partial regression output from SPSS showing coefficient of determination and other important results

Standard deviation measures the deviation of data around the arithmetic mean; similarly, standard error can be understood as the standard deviation around the regression line.

Variation of the dots around the regression line represents the degree of relationship between two variables x and y. Though the least squares method results in a regression line that fits the data best, all the observed data points do not fall exactly on the regression line. There is an obvious variation of the observed data points around the regression line. So, there is a need to develop a statistic which can measure the differences between the actual values (y_i) and the regressed values ($\hat{y}_i$). Standard error fulfils this need. Standard error measures the amount by which the regressed values ($\hat{y}_i$) are away from the actual values (y_i). This is the same as the concept of standard deviation that we developed in Chapter 4. Standard deviation measures the deviation of data around the arithmetic mean; similarly, standard error can be understood as the standard deviation around the regression line. Standard error of the estimate can be defined as

Standard error of the estimate

$$S_{yx} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$$

A large standard error indicates a large amount of variation or scatter around the regression line and a small standard error indicates small amount of variation or scatter around the regression line. A standard error equal to zero indicates that all the observed data points fall exactly on the regression line.

where y_i is the actual value of y, for observation i and $\hat{y}_i$ the regressed (predicted) value of y, for observation i.

In the above formula, the numerator is the error sum of squares and the denominator is degrees of freedom determined by subtracting the number of parameters, β_0 and β_1, that is, 2 from sample size n. Hence, the degrees of freedom is $n - 2$. In Example 14.1, the sample size is 12 and there are two parameters. Therefore, the degrees of freedom can be computed as $12 - 2 = 10$. A large standard error indicates a large amount of variation or scatter around the regression line and a small standard error indicates small amount of variation or scatter around the regression line. A standard error equal to zero indicates that all the observed data points fall exactly on the regression line.

For Example 14.1, standard error of the estimate can be computed as

$$S_{yx} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{13769.20842}{12-2}} = 37.1068$$

Figures 14.21, 14.22, and 14.23 exhibit the computation of standard error from MS Excel, Minitab, and SPSS, respectively. Figure 14.24 is the scatter plot exhibiting actual values and the regression line for Example 14.1.

Table 14.3 indicates the predicted (regressed) values and residuals for Example 14.1.

TABLE 14.3
Predicted (regressed) values and residuals for Example 14.1

Months	*Advertisement (in thousand rupees): x*	*Sales (in thousand rupees): y*	*Predicted values: $\hat{y}$*	*Residuals $(y_i - \hat{y}_i)$*
Jan	92	930	902.39651	27.60349
Feb	94	900	940.53740	–40.53740
Mar	97	1020	997.74873	22.25127
Apr	98	990	1016.81917	–26.81917
May	100	1100	1054.96006	45.03994
Jun	102	1050	1093.10094	–43.10094
Jul	104	1150	1131.24183	18.75817
Aug	105	1120	1150.31227	–30.31227
Sep	105	1130	1150.31227	–20.31227
Oct	107	1200	1188.45316	11.54684
Nov	107	1250	1188.45316	61.54684
Dec	110	1220	1245.66449	–25.66449
				$\Sigma(y_i - \hat{y}_i) = 0.000$

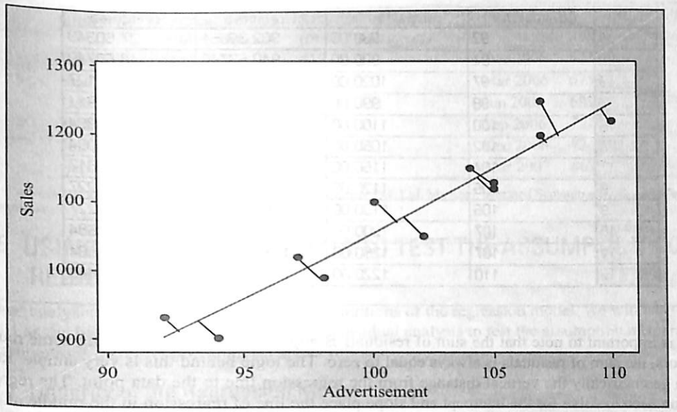


FIGURE 14.24
Scatter plot exhibiting actual values and the regression line for Example 14.1

Figures 14.25, 4.26, and 14.27 exhibit the computation of predicted values (fits) and residuals, and are the part of the regression outputs obtained from MS Excel, Minitab, and SPSS, respectively.

24	Observation	Predicted Y	Residuals	Standard Residuals
25	1	902.3965142	27.60348584	0.780199711
26	2	940.5374001	-40.53740015	-1.145770793
27	3	997.7487291	22.25127088	0.62892184
28	4	1016.819172	-26.81917211	-0.758031447
29	5	1054.960058	45.0399419	1.273033046
30	6	1093.100944	-43.10094408	-1.218228172
31	7	1131.24183	18.75816993	0.530190964
32	8	1150.312273	-30.31227306	-0.856762324
33	9	1150.312273	-20.31227306	-0.574116967
34	10	1188.453159	11.54684096	0.326366098
35	11	1188.453159	61.54684096	1.739592883
36	12	1245.664488	-25.66448802	-0.725394838

FIGURE 14.25
MS Excel output (partial) exhibiting the computation of predicted values, residuals, and standardized residuals for Example 14.1

↓	C1-D	C2	C3	C4	C5
	Months	Advertisement	Sales	Residuals	Fits
1	Jan	92	930	27.6035	902.40
2	Feb	94	900	-40.5374	940.54
3	Mar	97	1020	22.2513	997.75
4	Apr	98	990	-26.8192	1016.82
5	May	100	1100	45.0399	1054.96
6	Jun	102	1050	-43.1009	1093.10
7	Jul	104	1150	18.7582	1131.24
8	Aug	105	1120	-30.3123	1150.31
9	Sep	105	1130	-20.3123	1150.31
10	Oct	107	1200	11.5468	1188.45
11	Nov	107	1250	61.5468	1188.45
12	Dec	110	1220	-25.6645	1245.66

FIGURE 14.26
Minitab output (partial) exhibiting the computation of residuals and predicted values (fits) for Example 14.1

	Advertisement	Sales	Predicted	Residuals
1	92	930.00	902.39651	27.60349
2	94	900.00	940.53740	-40.53740
3	97	1020.00	997.74873	22.25127
4	98	990.00	1016.81917	-26.81917
5	100	1100.00	1054.96006	45.03994
6	102	1050.00	1093.10094	-43.10094
7	104	1150.00	1131.24183	18.75817
8	105	1120.00	1150.31227	-30.31227
9	105	1130.00	1150.31227	-20.31227
10	107	1200.00	1188.45316	11.54684
11	107	1250.00	1188.45316	61.54684
12	110	1220.00	1245.66449	-25.66449

FIGURE 14.27
SPSS output (partial) exhibiting the computation of predicted values (fits) and residuals for Example 14.1

It is important to note that the sum of residuals is approximately zero. The logic behind this is very simple. In fact, residuals are geometrically the vertical distance from the regression line to data point. The regression equation which we solve for intercept and slope, place the line of regression in the middle of all the data points. So, the vertical distance from the line to data points cancel each other and lead to a sum that is approximately equal to zero.

It is important to note that the sum of residuals is approximately zero. Ignoring some rounding off errors, the sum of residuals is always equal to zero. The logic behind this is very simple. Residuals are geometrically the vertical distance from the regression line to the data point. The regression equation used to solve for the intercept and slope place the line of regression in the middle of all the data points. So, the vertical distance from the line to data points cancel each other and lead to a sum that is approximately equal to zero. Figure 14.24 is the scatter plot with residuals (distance between actual values and predicted values) for Example 14.1. This figure clearly exhibits that that the line of regression is geometrically in the middle of all the data points. This also exhibits that the residuals with (+) sign fall above the regression line and residuals with (–) sign fall below the regression line. Table 14.3 clearly exhibits that the sum of residuals is approximately equal to zero. Residuals are also used to find out outliers in the data set. This can be done by examining the scatter plot. Outliers can produce residuals with large magnitudes. These outliers may be due to misreported or miscoded data. These outliers sometimes pull the regression line towards them and hence put undue influence on the regression line. A researcher after identifying the origin of the outlier can decide whether the outlier should be retained in the regression equation or regression line should be computed without it.

SELF-PRACTICE PROBLEMS

14B1. Compute the value of r^2 and standard error for Problem 14A1. Discuss the meaning of the value of r^2 and standard error in developing a regression model.

14B2. Compute the value of r^2 and standard error for Problem 14A2. Discuss the meaning of the value of r^2 and standard error in developing a regression model.

14B3. Nestle India Ltd, incorporated in 1959, is one of the largest dairy product companies in India. The company has a broad product portfolio comprising of milk products, beverages, prepared dishes, cooking aids, chocolate, and confectionary. The following table shows the net sales (in million rupees) and salaries and wages (in million rupees) of the company for different quarters.

Develop a simple regression line to predict net sales from salaries and wages. Discuss the meaning of the value of r^2 and standard error in developing a regression model.

Quarters	*Net sales (in million rupees)*	*Salaries and wages (in million rupees)*	*Quarters*	*Net sales (in million rupees)*	*Salaries and wages (in million rupees)*
Jun 1999	3639	220	Dec 2001	4681	369
Sep 1999	4169	211	Mar 2002	5300.1	321.9
Dec 1999	4230	277	Jun 2002	5114.8	336.9
Mar 2000	3478	243	Sep 2002	5235	500.3
Jun 2000	4198	259	Dec 2002	4827.1	303
Sep 2000	4694	264	Mar 2003	5981	388.3
Dec 2000	4403	284	Jun 2003	5460.7	380.7
Mar 2001	4516	308	Sep 2003	5326.1	390.7
Jun 2001	4683	314	Dec 2003	5305	424.4
Sep 2001	5329.6	329.7	Mar 2004	6200.7	413.1
Dec 2001	5143.9	412	Dec 2005	6227.9	440.7
Sep 2004	5600.2	390.3	Mar 2006	6759.2	542.8
Dec 2004	5719.8	427.1	Jun 2006	6811.8	565.1
Mar 2005	6135.3	443.9	Sep 2006	7226.6	566.1
Jun 2005	6157.7	475	Dec 2006	7362.9	569.4
Sep 2005	6248.1	473.3	Mar 2007	8630.8	1399.8

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed September 2008, reproduced with permission.

14.8 USING RESIDUAL ANALYSIS TO TEST THE ASSUMPTIONS OF REGRESSION

Residual analysis is mainly used to test the assumptions of the regression model. We will take Example 14.1 as the base example for understanding residual analysis to test the assumption of regression. The assumptions of regression analysis are as follows:

14.8.1 Linearity of the Regression Model

Linearity of the regression model can be obtained by plotting the residuals on the vertical axis against the corresponding x_i values of the independent variable on the horizontal axis. There should not be any apparent pattern in the plot for a fit regression model. Any deviation from linear residual plot (plot with appearent pattern) indicates that there is a non-linear relationship between the independent variable and the dependent variable.

Linearity of the regression model can be obtained by plotting the residuals on the vertical axis against the corresponding x_i values of the independent variable on the horizontal axis. There should not be any apparent pattern in the plot for a fit regression model.

Figure 14.28 (MS Excel plot of residuals and x_i values for Example 14.1) clearly exhibits no apparent pattern in the plot between residuals and x_i values of the independent variable. It is important to note that for meaningful interpretation of the residual plot, large sample size is required. Residual analysis can lead to over interpretation for small sample size. Figure 14.29 (MS Excel plot of residuals and x_i values for a large sample size) exhibits the non-linearity in the plot between residuals and x_i values of the independent variable for a large sample size. Similarly, Figure 14.31 exhibits the non-linearity in the Minitab produced plot between residuals and x_i values of the independent variable for a large sample size. Figure 14.30 is a part of Minitab regression analysis output for Example 14.1 and does not indicate an apparent pattern in the plot between residuals and x_i values of the independent variable.

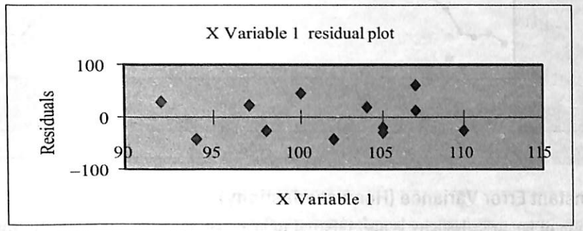


FIGURE 14.28
MS Excel plot of residuals for Example 14.1 exhibiting linearity

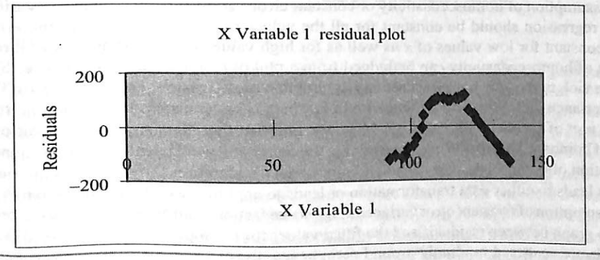


FIGURE 14.29
MS Excel plot of residuals showing non-linearity for a large sample size

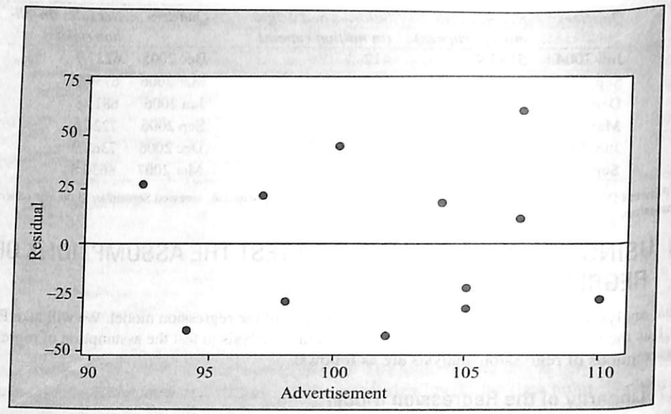


FIGURE 14.30
Minitab plot of residuals versus independent variable (advertisement) for Example 14.1 showing linearity

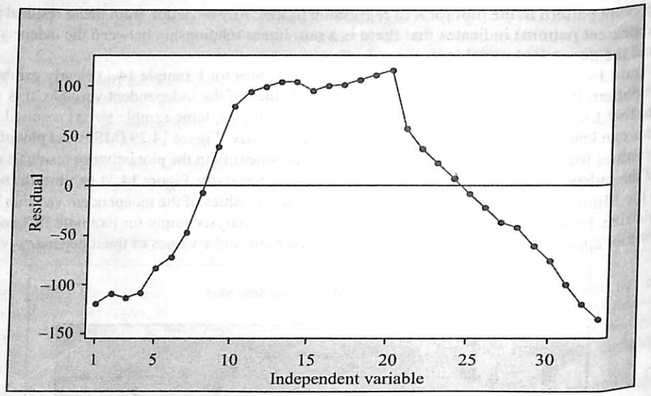


FIGURE 14.31
Minitab plot of residuals showing non-linearity for a large sample size

The assumption of homoscedasticity is also referred to as constant error variance. As the name suggests, the assumption of homoscedasticity or constant error variance requires that the variance around the line of regression should be constant for all the values of x_i.

14.8.2 Constant Error Variance (Homoscedasticity)

The assumption of homoscedasticity is also referred to as constant error variance. As the name suggests, the assumption of homoscedasticity or constant error variance requires that the variance around the line of regression should be constant for all the values of x_i. This means that the error variance should be constant for low values of x as well as for high values of x. As shown in Figure 14.32, the assumption of homoscedasticity can be judged from a plot of residuals and values of x_i. Figure 14.32 exhibits the violation of the homoscedasticity assumption of regression. From Figure 14.32, it is clear that error variance increases with the increase in x, which is not constant. If we examine Figure 14.28 (MS Excel plot of residuals for Example 14.1), we find that there is no apparent violation of the assumption of homoscedasticity. While determining the regression coefficient from least squares method, the assumption of homoscedasticity is a very important consideration. Any serious violation from this assumption leads to either data transformation or leads to applying weighted least squares method.

The assumption of constant error variance or homoscedasticity can also be understood by examining the Minitab graph between residuals and the fitted values for Example 14.1 (Figure 14.33). In this plot the residuals are scattered randomly around zero, hence, the errors have constant variance or do not

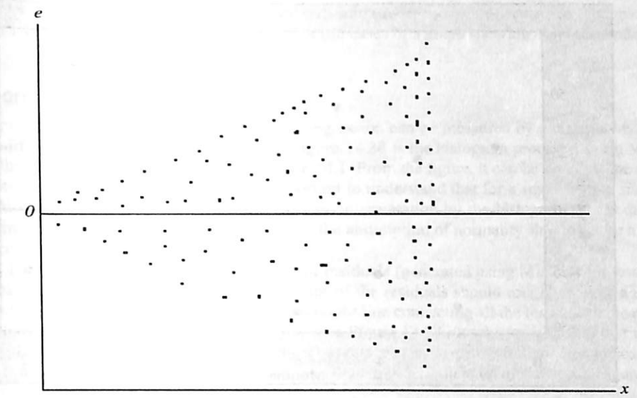


FIGURE 14.32
Violation of the homoscedasticity assumption of regression

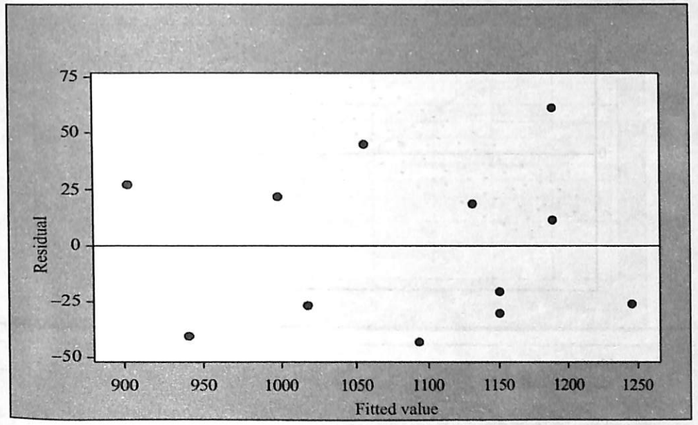


FIGURE 14.33
Minitab worksheet showing constant error variance (homoscedasticity) for Example 14.1

violate the assumption of homoscedasticity. If the residuals increase or decrease with fitted value in a funnel pattern (shown in Figure 14.32), errors may not have constant variance.

14.8.3 Independence of Error

The assumption of independence of error indicates that the value of error ε, for any particular value of independent variable x, should not be related to the value of error ε, for any other value of independent variable x. This means that the errors around the line of regression should be independent for each value of the independent variable x. This assumption is particularly important when a researcher collects the data over a period of time. In this situation, there is a possibility that the errors for a specific time period may correlate with the errors of another time period. In other words, we can say that the data collected over a specific period of time may exhibit autocorrelation effect with the data collected over another specific period of time. In this situation, there exists a relationship between consecutive residuals. The effect of autocorrelation can be measured by the Durbin–Watson statistic, which we will discuss later in this chapter. Residual versus time graph can be plotted to ascertain the assumption of independence of error.

The assumption of independence of error indicates that the value of error ε, for any particular value of independent variable x, should not be related to the value of error ε, for any other value of independent variable x. This means that the errors around the line of regression should be independent for each value of x.

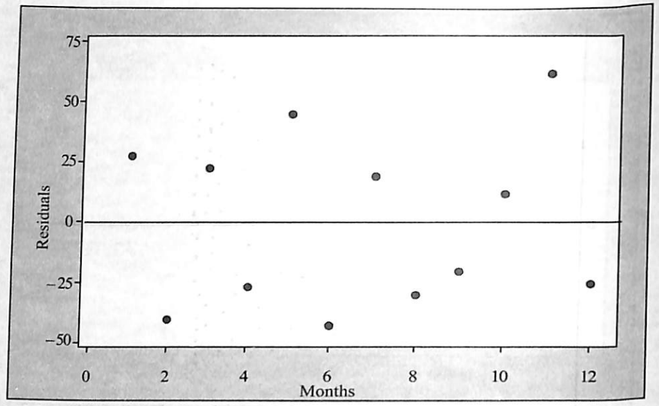


FIGURE 14.34
Minitab sheet showing independence of error for Example 14.1

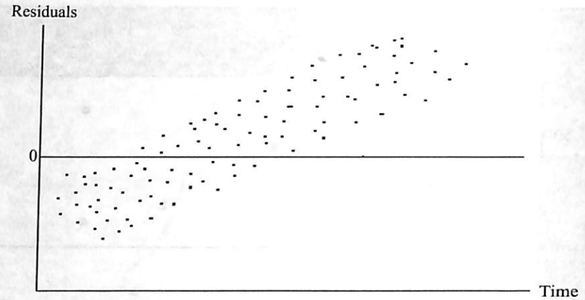


FIGURE 14.35
Graph of non-independence of error (Case 1)

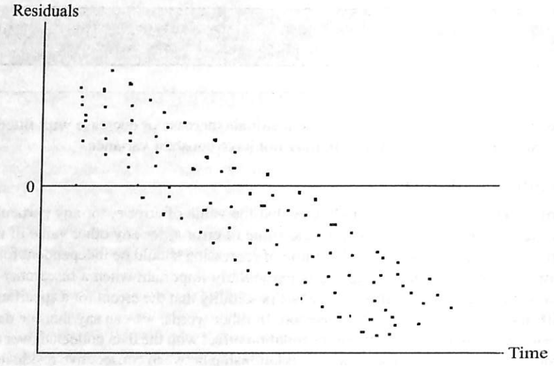


FIGURE 14.36
Graph of non-independence of error (Case 2)

Figure 14.34 shows the Minitab worksheet indicating independence of error (for Example 14.1) and Figures 14.35 and 14.36 illustrate the two specific cases of a graph showing non-independence of error.

14.8.4 Normality of Error

The assumption of normality around the line of regression can be measured by plotting a histogram between residuals and frequency distribution. Figure 14.38 is the histogram produced using Minitab for testing the normality assumption for Example 14.1. From the figure, it can be seen that the residuals are right-skewed distributed. Here, it is important to understand that for a small sample size such as 12, meeting the assumption of normality and its interpretation by the histogram plot is difficult. With this kind of sample size, any deviation from the assumption of normality should not be a matter of serious concern.

The assumption of normality around the line of regression can be measured by plotting a histogram between residuals and frequency distribution.

Figure 14.37 is the normal probability plot of residuals (generated using Minitab) for testing the normality assumption. The normal probability plot of the residuals should roughly follow a straight line for meeting the assumption of normality. A straight line connecting all the residuals indicates that the residuals are normally distributed. If we observe Figure 14.37 closely, we will find that the line connecting all the residuals is not exactly straight but rather close to a straight line. This indicates that the residuals are nearly normal in shape. A curve in the tail is an indication of skewness. Figure 14.38

The normal probability plot of the residuals should roughly follow a straight line for meeting the assumption of normality. A straight line connecting all the residuals indicates that the residuals are normally distributed.

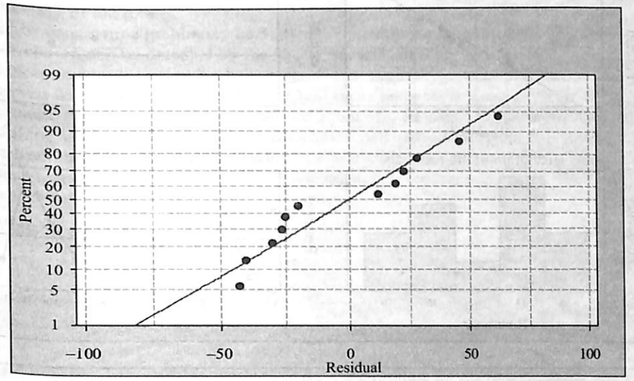


FIGURE 14.37
Normal probability plot of residuals for testing the normality assumption for Example 14.1 produced using Minitab

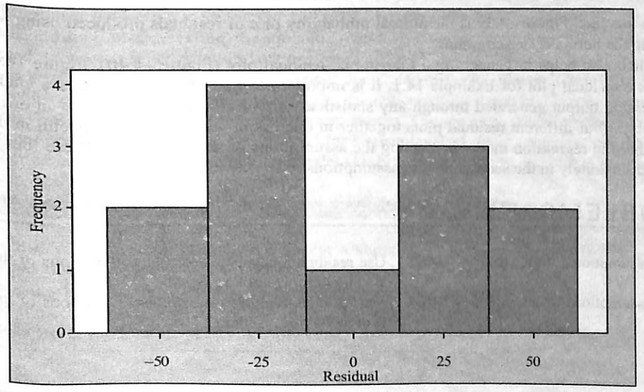


FIGURE 14.38
Histogram of residuals for testing the normality assumption for Example 14.1 produced using Minitab

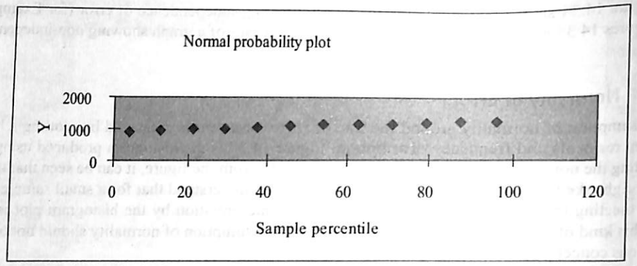


FIGURE 14.39
MS Excel normal probability plot of residuals for testing the normality assumption for Example 14.1

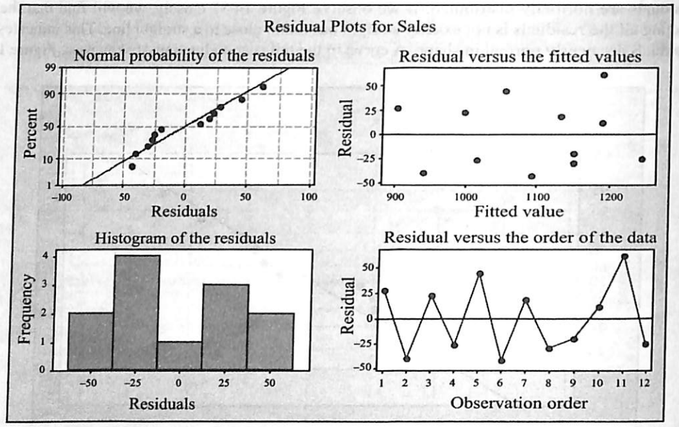


FIGURE 14.40
Minitab generated four-in-one-residual plot for Example 14.1

confirms this fact. Figure 14.39 is the normal probability plot of residuals produced using MS Excel for testing the normality assumption.

Minitab also helps in generating a four-in-one residual plot (Figure 14.40). Figure 14.40 is the four-in-one residual plot for Example 14.1. It is important to note that these plots are vital parts of the regression output generated through any statistical software program. This four-in-one-residual plot displays four different residual plots together in one graph window. This is useful in determining whether the regression model is meeting the assumptions of the regression. These four plots are explained separately in the section on the assumptions of regression.

SELF-PRACTICE PROBLEMS

14C1. Use residual analysis to test the assumptions of the regression model for problem 14A1.

14C2. Use residual analysis to test the assumptions of the regression model for problem 14A2.

14C3. Use residual analysis to test the assumptions of the regression model for problem 14A4.

14C4. Use residual analysis to test the assumptions of the regression model for problem 14B3.

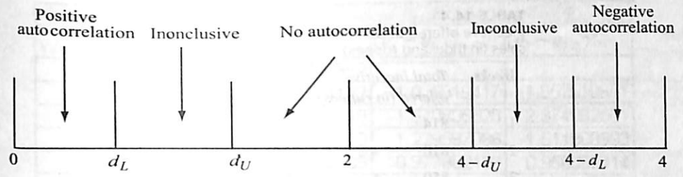


FIGURE 14.41
Using Durbin–Watson statistic for detecting autocorrelation

14.9 MEASURING AUTOCORRELATION: THE DURBIN–WATSON STATISTIC

As discussed, in the previous section, independence of errors is one of the basic assumptions of regression analysis. When a researcher collects data over a period of time, there is a possibility that the errors for a specific time period may be correlated with the errors of another time period because residuals at any given time period tend to be similar to residuals at another period of time. This is termed as autocorrelation and the presence of autocorrelation in a regression model raises questions about the validity of the model.

When a researcher collects data over a period of time, there is a possibility that the errors for a specific time period may be correlated with the errors of another time period because residuals at any given time period may tend to be similar to residuals at another period of time. This is called autocorrelation and the presence of autocorrelation in any regression model raises questions about the validity of the model.

A residual versus time graph may be plotted for determining autocorrelation (Figure 14.34). Positive autocorrelation can be detected by the cluster of residuals with the same sign. In case of negative autocorrelation, residuals tend to vary from positive to negative to positive and so on. This pattern is rarely observed in regression analysis, so we will focus on positive autocorrelation. It has also been discussed earlier that the pattern of residual–time plot may be observed for determining autocorrelation. In addition to this, the status of autocorrelation in regression analysis may also be determined through the Durbin–Watson statistic. The Durbin–Watson statistic measures the degree of correlation between each residual and the residual of the immediately preceding time period. The Durbin–Watson statistic can be defined as

Durbin–Watson statistic

$$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

The Durbin–Watson statistic measures the degree of correlation between each residual and the residual of the immediately preceding time period.

where e_i is the residual for the time period i and e_{i-1} the residual for the time period $i-1$.

Here, it is important to note that the numerator of the Durbin–Watson statistic is the sum of squared differences between two successive residuals from the second observation to the nth observation because for the first observation, the squared differences between two successive residuals cannot be computed. If there is no correlation between residuals, the value of D will be close to 2. In case of negative correlation, the value of D will be greater than 2 and can reach its maximum value 4.

If there is no correlation between residuals, the value of D will be close to 2. In case of negative correlation, the value of D will be greater than 2 and can reach its maximum value 4.

The values of the lower-critical value (d_L) and the upper-critical value (d_U) can be obtained from the Durbin–Watson statistical table given in the appendices. The values of the lower critical value (d_L) and the upper critical value (d_U) can be obtained for a given level of significance (α); sample size (n), and number of independent variables in the model (k). Figure 14.41, shows how the Durbin–Watson statistic can be used for detecting autocorrelation.

Example 14.2 explains the concept of positive autocorrelation clearly.

Example 14.2

A retail outlet of a footwear company is facing a slump in sales. The company has adopted a policy of giving incentives to its salesmen for additional sales in order to boost the sales volume. The total incentives offered by the company and the sales volumes for 15 weeks (in thousand rupees) selected at random are given in Table 14.4.

TABLE 14.4
Incentive offered to salesmen (in rupees) and sales (in thousand rupees)

Weeks	*Total incentive offered (in rupees)*	*Sales (in thousand rupees)*
1	814	10.5
2	810	9.4
3	850	8.6
4	870	10.2
5	855	10.9
6	845	11.1
7	865	12.1
8	880	12.45
9	890	13.05
10	930	13.55
11	905	12.9
12	865	11.4
13	945	11.75
14	995	12.15
15	845	9.65

Fit a line of regression and also determine whether autocorrelation is present.

Solution

It is clear from the example that the data are collected over a period of 15 randomly selected weeks from the same retail store. So, apart from verifying the assumptions of homoscedasticity and normality, verification of independence of error in terms of using Durbin–Watson statistic is also very important. The first step in determining autocorrelation is the examination of residual versus time graph. The MS Excel plot between residuals versus time is shown in Figure 14.42.

It is clear from Figure 14.43,14.44, and 14.45 that the Durbin–Watson statistic is calculated as 0.51. From the Durbin–Watson statistic table, for a given level of significance (0.05); sample size (15) and number of independent variables in the model (1), lower critical value (d_L) and the upper critical value (d_U) are observed

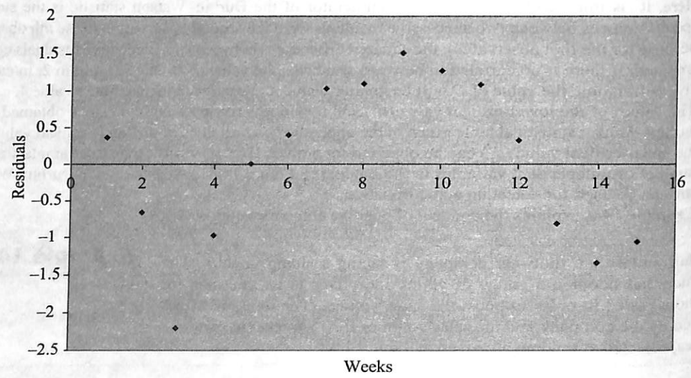


FIGURE 14.42
MS Excel produced residuals versus time plot for Example 14.2

E	F	G	H	I
Weeks	Residuals	e^2	$(e_i - e_{i-1})$	$(e_i - e_{i-1})^2$
1	0.3645479	0.1328952		
2	-0.6613715	0.4374122	-1.025919417	1.052510651
3	-2.2021773	4.8495849	-1.540805828	2.374082601
4	-0.9725802	0.9459123	1.229597086	1.511908993
5	0.005222	2.727E-05	0.977802186	0.956097114
6	0.3904234	0.1524304	0.385201457	0.148380163
7	1.0200205	1.0404418	0.629597086	0.39639249
8	1.0922183	1.1929409	0.07219781 4	0.005212524
9	1.5070169	2.2710998	0.414798543	0.172057831
10	1.266211	1.6032904	-0.240805828	0.057987447
11	1.0792147	1.1647043	-0.186996357	0.034967638
12	0.3200205	0.1024131	-0.759194172	0.57637579
13	-0.8115912	0.6586802	-1.131611657	1.280544942
14	-1.3375984	1.7891696	-0.526007286	0.276683664
15	-1.0595766	1.1227025	0.278021857	0.077296153
Sum=		17.463705		8.920498001
		D=	0.510802147	

FIGURE 14.43
MS Excel worksheet showing computation of the Durbin–Watson statistic for Example 14.2

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.635[a]	.403	.357	1.15903	.511

a. Predictors: (Constant), Incentive
b. Dependent Variable: Sales

Durbin-watson statistic

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	11.789	1	11.789	8.775	.011[a]
	Residual	17.464	13	1.343		
	Total	29.252	14			

a. Predictors: (Constant), Incentive
b. Dependent Variable: Sales

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-4.940	5.495		-.899	.385
	Incentive	.019	.006	.635	2.962	.011

a. Dependent Variable: Sales

FIGURE 14.44
SPSS regression output for Example 14.2

Regression Analysis: Sales versus Incentive

The regression equation is
Sales = - 4.94 + 0.0185 Incentive

Predictor	Coef	SE Coef	T	P
Constant	-4.940	5.495	-0.90	0.385
Incentive	0.018520	0.006252	2.96	0.011

S = 1.15903 R-Sq = 40.3% R-Sq(adj) = 35.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	11.789	11.789	8.78	0.011
Residual Error	13	17.464	1.343		
Total	14	29.252			

Durbin-Watson statistic = 0.510802

FIGURE 14.45
Minitab regression output for Example 14.2

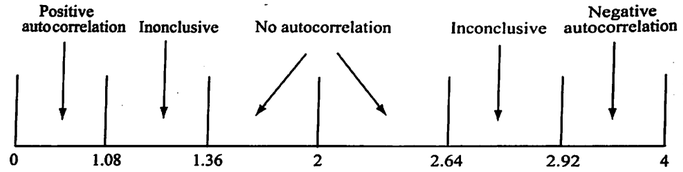


FIGURE 14.46
Durbin–Watson statistic range for Example 14.2

as 1.08 and 1.36, respectively. By substituting the values of the lower critical value (d_L) and the upper critical value (d_U) in the range presented in Figure 14.41, the acceptance and rejection range can be determined easily. After placing the values of the lower critical value (d_L) and the upper critical value (d_U) in the range presented in Figure 14.41, the Durbin–Watson static range for Example 14.2 is constructed as shown in Figure 14.46. The Durbin–Watson statistic for Example 14.2 is calculated as 0.51. This value (0.51) is less than the lower critical value (d_L = 1.08). Hence, it can be concluded that a significant positive autocorrelation exists between the residuals. So, the outputs (Figure 14.43, Figure 14.44, and Figure 14.45) based on least squares method are inappropriate. There is a need to focus on alternative approaches.

14.10 STATISTICAL INFERENCE ABOUT SLOPE, CORRELATION COEFFICIENT OF THE REGRESSION MODEL, AND TESTING THE OVERALL MODEL

If there is no serious violation of the assumption of linear regression and residual analysis has confirmed that the straight line regression model is appropriate, an inference about the linear relationship between variables can be obtained on the basis of sample results.

14.10.1 *t* Test for the Slope of the Regression Line

After verifying the assumptions of linear regression, a researcher has to determine whether a significant linear relationship exists between the independent variable x and the dependent variable y. This is determined by performing a hypothesis test to check whether the population slope (β_1) is zero. The hypotheses for the test can be stated as below:

$$H_0: \beta_1 = 0 \text{ (There is no linear relationship)}$$
$$H_1: \beta_1 \neq 0 \text{ (There is a linear relationship)}$$

Any negative or positive value of the slope will lead to the rejection of the null hypothesis and acceptance of the alternative hypothesis (as the above hypothesis test is two-tailed). A negative value of the slope indicates the inverse relationship between the independent variable x and the dependent variable y. This means that larger values of the independent variable x are related to smaller values of variable y. This means that larger values of the independent variable x are related to smaller values of the dependent variable y and vice versa. In order to test the significant positive relationship between the two variables, the null and alternative hypotheses can be stated as below:

$$H_0: \beta_1 = 0 \text{ (There is no linear relationship)}$$
$$H_1: \beta_1 > 0 \text{ (There is a positive relationship)}$$

To test the significant negative relationship between the two variables, the null and alternative hypotheses can be stated as below:

$$H_0: \beta_1 = 0 \text{ (There is no linear relationship)}$$
$$H_1: \beta_1 < 0 \text{ (There is a negative relationship)}$$

The test statistic t can be defined as below:

$$t = \frac{b_1 - \beta_1}{S_b}$$

where

$$S_b = \frac{S_{yx}}{\sqrt{SS_{xx}}}$$

$$S_{yx} = \sqrt{\frac{\text{SSE}}{n-2}}$$

$$SS_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n}$$

The test statistic t follows a t distribution with $n-2$ degrees of freedom and β_1 as the hypothesized population slope.
On the basis of above formula, the t statistic for Example 14.1 can be computed as

$$t = \frac{b_1 - \beta_1}{S_b} = \frac{19.07 - 0}{\frac{37.1068}{\sqrt{344.25}}} = 9.53$$

where

$$SS_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 124{,}581 - \frac{(1221)^2}{12} = 344.25$$

$$S_{yx} = \sqrt{\frac{\text{SSE}}{n-2}} = \sqrt{\frac{13{,}769.20842}{12\text{-}2}} = 37.1068$$

Figures 14.47(A), 14.47(B), and 14.47(C) show the computation of the t statistic using MS Excel, Minitab, and SPSS, respectively.

Using the p value from the above outputs, the null hypothesis is rejected and the alternative hypothesis is accepted at 5% level of significance. In light of the positive value of b_1 and p value = 0.000, it can be concluded that a significant positive linear relationship exists between the independent variable x and the dependent variable y.

16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	-852.0842411	203.7758887	-4.18148	0.001883	-1306.125214	-398.043
18	X Variable 1	19.07044299	1.999942514	9.535496	2.45E-06	14.61429339	23.52659

t statistic

FIGURE 14.47(A)
Computation of the *t* statistic for Example 14.1 using MS Excel

```
Predictor          Coef  SE Coef       T      P
Constant         -852.1    203.8   -4.18  0.002
Adevertisement   19.070    2.000    9.54  0.000
```

t statistic

FIGURE 14.47(B)
Computation of *t* statistic for Example 14.1 using Minitab

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	-852.084	203.776		-4.181	.002	-1306.125	-398.043
	advertisement	19.070	2.000	.949	9.535	.000	14.614	23.527

t statistic

FIGURE 14.47(C)
Computation of the *t* statistic for Example 14.1 using SPSS

14.10.2 Testing the Overall Model

The *F* test is used to determine the significance of overall regression model in regression analysis. More specifically, in case of a multiple regression model, the *F* test determines that at least one of the regression coefficients is different from zero. In case of simple regression, where there is only one predictor the *F* test for overall significance tests the same phenomenon as the *t*-statistic test in simple regression. The *F* statistic can be defined as the ratio of regression mean square (MSR) and error mean square (MSE).

***F* statistic for testing the slope**

$$F = \frac{MSR}{MSE}$$

where $MSR = \frac{SSR}{k}$, $MSE = \frac{SSE}{n-k-1}$, and k is the number of independent (explanatory) variables in regression model (In case of simple regression $k = 1$).

The *F* statistic follows the *F* distribution with degrees of freedom k and $n - k - 1$.

Figures 14.48(A), 14.48(B), and 14.48(C) illustrate the computation of *F* statistic using MS Excel, Minitab, and SPSS, respectively. On the basis of the *p* value obtained from the outputs, it can be

10	ANOVA					
11		df	SS	MS	F	Significance F
12	Regression	1	125197.4582	125197.5	90.92568	2.45382E-06
13	Residual	10	13769.20842	1376.921		
14	Total	11	138966.6667			

F statistic

FIGURE 14.48(A)
Computation of the *F* statistic from MS Excel for Example 14.1


```
Analysis of Variance

Source           DF      SS      MS      F      P
Regression        1  125197  125197  90.93  0.000
Residual Error   10   13769    1377
Total            11  138967
```

F statistic

FIGURE 14.48(B)
Computation of *F* statistic for Example 14.1 using Minitab

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	125197.5	1	125197.458	90.926	.000[a]
	Residual	13769.208	10	1376.921		
	Total	138966.7	11			

a. Predictors: (Constant), advertisement

b. Dependent Variable: sales

F statistic

FIGURE 14.48(C)
Computation of *F* statistic for Example 14.1 using SPSS

concluded that expenses on advertisement is significantly (at 5% level of significance) related to sales. If we compare the *p* value obtained from Figures 14.47 and 14.48, we find that the *p* values are the same in both the cases.

14.10.3 Estimate of Confidence Interval for the Population Slope (β_1)

Estimate of confidence interval for the population slope (β_1) provides an alternative approach to test the linear relationship between the independent variable *x* and the dependent variable *y*. This can be done by determining whether the hypothesized value of β_1 ($\beta_1 = 0$) is within the interval or outside the interval. For understanding the concept, we will take Example 14.1 again. Confidence interval for the population slope (β_1) is defined as

Estimate of confidence interval for the population slope (β_1)

$$b_1 \pm t_{n-2} S_b$$

From the outputs given in Figures 14.6, 14.12, and 14.18, the following values can be obtained

$$b_1 = 19.0704 \qquad n = 12, \qquad \text{and} \qquad S_b = 1.9999$$

From the table, for $\alpha = 0.05 \left(\frac{\alpha}{2} = 0.025\right)$ and degrees of freedom $= n - 2 = 10$, the value of t is 2.2281. By substituting all these values in the formula of confidence interval estimate for the population slope, we get

$$b_1 \pm t_{n-2} S_b = 19.0704 \pm 2.2281\ (1.9999) = 19.0704 \pm (4.4559)$$

So, the upper limit is 23.5263 (19.0704 + 4.4559) and the lower limit is 14.6145 (19.0704 − 4.4559).

So, population slope β_1 is estimated with 95% confidence to be in the interval of 14.6145 and 23.5263. Hence,

$14.6145 \le \beta_1 \le 23.5263$

The upper limit as well as the lower limit is greater than 0 and population slope lies in between these two limits. So, it can be concluded with 95% confidence that there exists a significant linear relationship between advertisement and sales. If the interval would have included 0, the inference would have been different. In this situation, the existence of a significant linear relationship between the two variables could not have been concluded. This confidence interval also indicates that for each thousand rupee increase in the advertisement expenditure, sales will increase by at least Rs 14,614.50 but less than Rs 23,526.30 (with 95% confidence).

14.10.4 Statistical Inference about Correlation Coefficient of the Regression Model

Correlation coefficient (r) measures the strength of the relationship between two variables.

From Figures 14.6, 14.12, and 14.18, it can be seen that the value of correlation coefficient is a part of the output. Correlation coefficient (r) measures the strength of the relationship between two variables. Correlation coefficient (r) specifies whether there is a statistically significant relationship between two variables. The t test can be applied to check this. The population correlation coefficient (ρ) can be hypothesized as equal to zero. In this case, the null and the alternative hypotheses can be stated as follows:

$$H_0: \rho = 0$$
$$H_1: \rho \neq 0$$

In order to test the significant relationship between two numerical variables statistically, the t statistic can be defined as

The t statistic for testing the statistical significant correlation coefficient

$$t = \frac{r - \rho}{\sqrt{\frac{1-r^2}{n-2}}}$$

where

$$r = +\sqrt{r^2}, \text{ if } b_1 \geq 0$$

$$r = -\sqrt{r^2}, \text{ if } b_1 < 0$$

The t statistic follows the t distribution with $n - 2$ degrees of freedom. From Figures 14.6, 14.12, and 14.18, the following values can be obtained:
$r = 0.9491$ and $b_1 = 19.0704$
By substituting these values in the above formula, we get

$$t = \frac{0.9491 - 0}{\sqrt{\frac{1-0.9009}{10}}} = 9.53$$

From the table, for $\alpha = 0.05\left(\frac{\alpha}{2} = 0.025\right)$ and degress of freedom $= n - 2 = 10$, the value of t is 2.2281. The calculated value of t is 9.53. The calculated value of t (= 9.53) > tabular value of t (= 2.2281). Hence, the null hypothesis is rejected and the alternative hypothesis is accepted. So, it can be concluded there is a significant relationship between two variables. It is important to note that the value of t is the same as calculated in Figures 14.6, 14.12, and 14.18.

The statistical significance of correlation coefficient can be directly inferred using Minitab and SPSS.

14.10.5 Using SPSS for Calculating Statistical Significant Correlation Coefficient for Example 14.1

Select **Analyze** from the menu bar and select **Correlate** from the pull-down menu. Another pull-down menu will appear on the screen, select **Bivariate** from this pull-down menu. The **Bivariate Correlations** dialog box will appear on the screen (Figure 14.49). Place both the variables in the **Variables** box, select **Pearson Correlation Coefficient** and **Two-tailed test of significance**. Select **Flag significant correlations** and click **OK**. SPSS will compute the **Pearson Correlation Coefficient** as shown in Figure 14.50.

14.10.6 Using Minitab for Calculating Statistical Significant Correlation Coefficient for Example 14.1

Select **Stat** from the menu bar. Select **Basic Statistics** from the pull-down menu. Another pull-down menu will appear on the screen, from this pull-down menu, select **Correlation**. The **Correlation** dialog box will appear on the screen (Figure 14.51). Place both the variables in the **Variables** box, select **Display *p*-values** and click **OK**. The Minitab output will appear on the screen as shown in Figure 14.52.

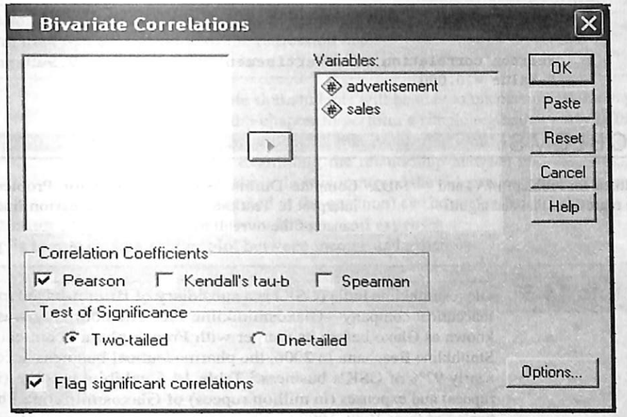


FIGURE 14.49
SPSS Bivariate correlation dialog box

Correlations

		Advertisement	Sales
Advertisement	Pearson Correlation	1	.949**
	Sig. (2-tailed)	.	.000
	N	12	12
Sales	Pearson Correlation	.949**	1
	Sig. (2-tailed)	.000	.
	N	12	12

**. Correlation is significant at the 0.01 level (2-tailed).

FIGURE 14.50
Calculation of Pearson correlation coefficient using SPSS

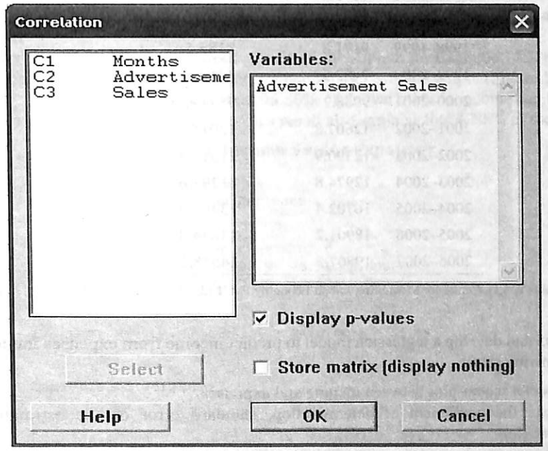


FIGURE 14.51
Minitab Correlation dialog box

FIGURE 14.52
Calculation of Pearson correlation coefficient using Minitab

Correlations: Advertisement, Sales

```
Pearson correlation of Advertisement and Sales = 0.949
P-Value = 0.000
```

SELF-PRACTICE PROBLEMS

14D1. Compute the Durbin–Watson statistic for Problem 14A4 and interpret it. Test the slope of the regression line and significance of the overall model.

14D2. Compute Durbin–Watson statistic for Problem 14B3 and interpret it. Test the slope of the regression line and significance of the overall model.

Example 14.3

Glaxosmithkline India (GSK) is a subsidiary of Britain-based major pharmaceutical company—Glaxosmithkline Plc. The company was formally known as Glaxo before its merger with French pharmaceutical company Smithkline Beecham. In 2006, the pharmaceutical business accounted for nearly 92% of GSK's business.[2] Table 14.5 exhibits income (in million rupees) and expenses (in million rupees) of Glaxosmithkline Pharmaceuticals Ltd from 1989–1990 to 2006–2007 (except 1993–1994).

TABLE 14.5
Income (in million rupees) and expenses (in million rupees) of Glaxosmithkline Pharmaceuticals Ltd from 1989–1990 to 2006–2007 (except 1993–1994)

Year	*Income (in million rupees)*	*Expenses (in million rupees)*
1989–1990	3566.4	3441.8
1990–1991	4232	4241.5
1991–1992	5024.8	5052.3
1992–1993	5650.8	5666.3
1994–1995	8076.4	7641.2
1995–1996	11478.9	9678.5
1996–1997	7315.3	6881.9
1997–1998	7883.5	7695.7
1998–1999	9171.8	8185.5
1999–2000	9482.5	8789.8
2000–2001	9958.2	9571.6
2001–2002	12607.8	12015.7
2002–2003	12390.9	11513.6
2003–2004	12974.8	11297.6
2004–2005	16702.4	13403.4
2005–2006	18901.2	13874.9
2006–2007	19807.5	14578.3

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, December 2008, reproduced with permission.

Use $\alpha = 0.05$ and develop a regression model to predict income from expenses incurred by performing the following steps:

1. Construct a scatter plot between income and expenses.
2. Calculate the coefficient of determination, standard error of the estimate, and state its interpretation.
3. Predict income when expenses are 20,000 million rupees.
4. Use residual analysis to test the assumptions of the regression model.
5. Perform the t test for the slope of the regression line.
6. Test the overall model.

Solution

It is important to note that students will be able to understand all the important points discussed in the chapter to perform a simple regression analysis from the step-wise solution provided for this problem. As discussed earlier, regression analysis starts with examining the relationship between two variables. In this case, the dependent variable is income and the independent variable is expenses. The six steps (mentioned in the question) can be performed as below:

1. Construction of a scatter plot between income and expenses

The first step is to construct a scatter plot between income and expenses

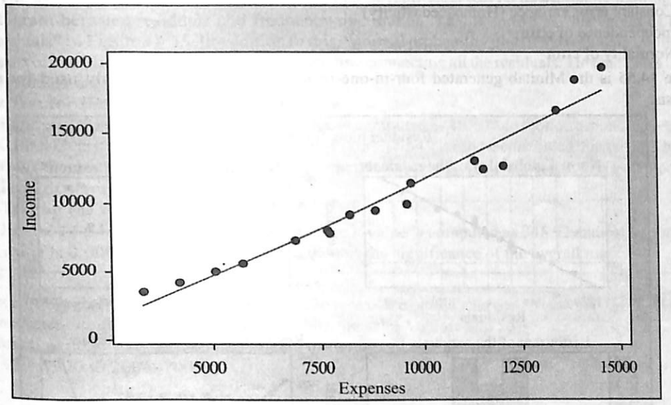


FIGURE 14.53
Scatter plot between income and expenses for Example 14.3

The scatter plot shown in Figure 14.53 (produced using Minitab) clearly exhibits a linear relationship between income and expenses. We can proceed further for regression analysis after confirming the linear relationship.

2. Calculation of coefficient of determination, standard error of the estimate, and its interpretation

Figure 14.54 is the regression analysis output generated by Minitab for Example 14.3. As discussed earlier in the chapter, r^2 is the coefficient of determination. The Minitab output (Figure 14.54) shows that the value of r^2 is 95.8%. This indicates that 95.80% of the variation in income can be explained by the independent variable, that is, expenses. This result also explains that 4.20 % of the variation in

Regression Analysis: Income versus Expenses

```
The regression equation is
Income = - 2323 + 1.40 Expenses


Predictor      Coef  SE Coef      T      P
Constant   -2323.3    722.9  -3.21  0.006
Expenses   1.39857  0.07520  18.60  0.000


S = 1021.97   R-Sq = 95.8%   R-Sq(adj) = 95.6%


Analysis of Variance

Source          DF         SS         MS      F      P
Regression       1  361293779  361293779  345.92  0.000
Residual Error  15   15666447    1044430
Total           16  376960226
```

FIGURE 14.54
Regression analysis output for Example 14.3 generated using Minitab

income is explained by factors other than expenses. The standard error is computed as 1021.97, which is relatively low and is an indication of a strong predictor regression model. The high value of r^2 and the low value of standard error provides a foundation for a good estimator model.

3. Predicting income when expenses are 20,000 million rupees

As exhibited in the Minitab output, regression equation is given as:

Income = –2323 + 1.40 (Expenses)

The predicted income when expenses are 20,000 million rupees can be computed as

Income = –2323 + 1.40 × (20,000) = 25,677

Hence, when expenses are Rs 20,000 million, the predicted income will be Rs 25,677 million.

4. Using residual analysis to test the assumptions of the regression model

As discussed in the chapter, we need to test the following four assumptions of the regression model:

(i) Linearity of the regression model
(ii) Constant error variance (Homoscedasticity)
(iii) Independence of error
(iv) Normality of error

Figure 14.55 is the Minitab generated four-in-one-residual plot, which is mainly used for residual analysis.

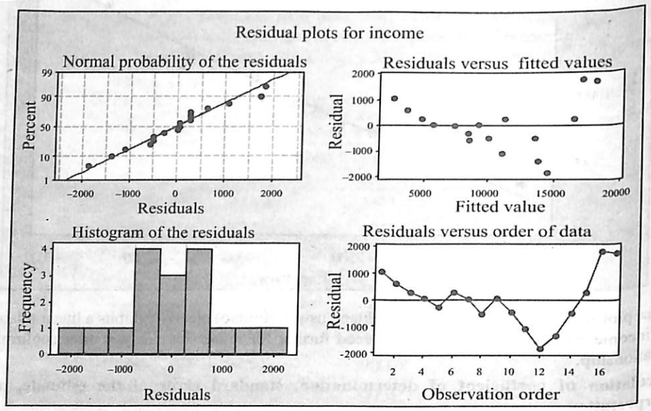


FIGURE 14.55
Minitab generated four-in-one-residual plot for Example 14.3

(i) Linearity of the regression model

As discussed in the chapter, for testing the assumption of linearity we have to construct a plot between residuals and the independent variable. Figure 14.56 shows the plot between residuals and independent variable expenses produced using Minitab.

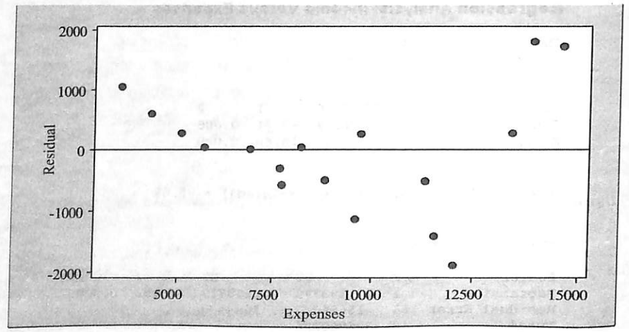


FIGURE 14.56
Minitab output exhibiting a plot between residuals and independent variable (expenses) for Example 14.3

Figure 14.56 clearly exhibits that there is no apparent pattern in the plot between residuals and x_i values of the independent variable (expenses). Hence, the assumption of linearity is not violated.

(ii) Constant error variance (Homoscedasticity)

The assumption of constant error variance or homoscedasticity can also be examined by the second part of the Minitab graph titled "residuals versus the fitted values" (Figure 14.55). In this plot, residuals are scattered randomly around zero. Hence, errors have constant variance or there is no violation of the assumption of homoscedasticity.

(iii) Independence of error

Residuals versus time graph can be plotted to ascertain the assumption of independence of error. This is shown as "residuals versus the order of the data" in the Minitab output (Figure 14.55). No apparent pattern again indicates independence of error.

(iv) Normality of error

The assumption of normality around the line of regression can be measured by plotting a histogram between residuals and frequency distribution. This is shown as "histogram of the residuals" in Figure 14.55. In addition to this, "normal probability plot of the residuals", which is a part of the Minitab output shows a straight line connecting all the residuals. This indicates that the residuals are normally distributed.

5. *t* Test for the slope of the regression line

Figure 14.54 clearly shows that the t value is computed as 18.60 and the corresponding p value is 0.000. Using the p value from the output (Figure 14.54), it can be concluded that the null hypothesis (slope is zero) is rejected and the alternative hypothesis (slope is not zero) is accepted at 5% level of significance.

6. Testing the overall model

Figure 14.54 includes a ANOVA table. The F value is computed as 345.92 and corresponding p value is 0.000. The p value (0.000) indicates the significance of the overall model.

Example 14.4

Ranbaxy Laboratories Ltd, incorporated in 1961, is one of India's largest pharmaceutical companies. Table 14.6 exhibits the sales volume and advertisement expenditure (in million rupees) of Ranbaxy Laboratories Ltd from 1989–1990 to 2006–2007.

TABLE 14.6
Sales and advertisement expenditure of Ranbaxy Laboratories Ltd from 1989–1990 to 2006–2007

Year	*Sales (in million rupees)*	*Advertisement (in million rupees)*
1989–1990	2064.5	30.4
1990–1991	2587.8	51
1991–1992	3396.9	59.1
1992–1993	4622.2	79.5
1993–1994	5944.7	50.8
1994–1995	7139.2	98.2
1995–1996	8940.1	112.7
1996–1997	10,427.3	141.6
1997–1998	12,421.3	224.8
1998–1999	11,296.5	169.8
1999–2000	16,670.3	409.3
2000–2001	17,757.1	560.2
2001–2002	19,597.8	863.5
2002–2003	31,317.6	1306.5
2003–2004	38,889.8	1822.6
2004–2005	38,658.7	2017.2
2005–2006	32,840.3	2008.1
2006–2007	35,991.5	1487.1

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed December 2008, reproduced with permission.

Use $\alpha = 0.05$ and develop a regression model to predict sales from advertisement expenses incurred by performing the following steps:

1. Construct a scatter plot between sales and advertisement.
2. Calculate the coefficient of determination, standard error of the estimate, and state its interpretation.
3. Predict sales when advertisement is 3000 million rupees.
4. Use residual analysis to test the assumptions of the regression model.
5. Perform the t test for the slope of the regression line.
6. Test the overall model.

Solution The first step in developing a regression model is to construct a scatter plot between sales and advertisement to ascertain the type of relationship between sales and advertisement.

1. Construction of a scatter plot between sales and advertisement expenditure

Figure 14.57 is the scatter plot between sales and adverticement of Ranbaxy Laboratories Ltd produced using Minitab. Since the scatter plot between sales and advertisement exhibits a linear relationship as shown in the figure, the further steps of performing a regression analysis can be carried out.

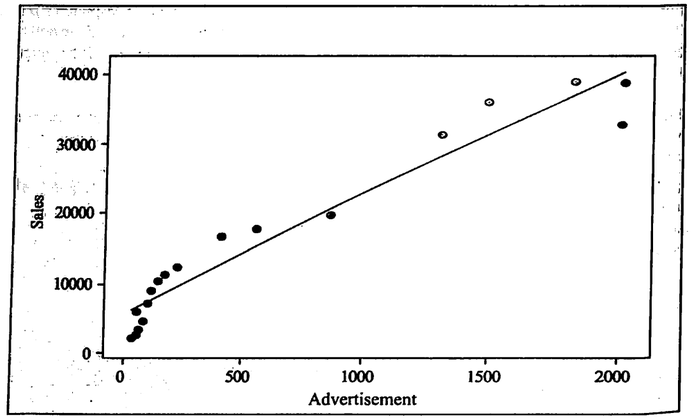


FIGURE 14.57
Scatter plot between sales and advertisement of Ranbaxy Laboratories Ltd for Example 14.4 produced using Minitab

2. Calculation of coefficient of determination, standard error of the estimate, and its interpretation

Figure 14.58 is the regression analysis output generated using MS Excel for Example 14.4. From the regression statistics part of the figure, it can be seen that the value of R^2 is 0.9355 (93.55%). This clearly explains that 93.55% of the variation in sales can be explained by the variation

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.9672348					
5	R Square	0.9355432					
6	Adjusted R Square	0.9315146					
7	Standard Error	3430.2371					
8	Observations	18					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	2732519348	2732519348	232.2282	6.02655E-11	
13	Residual	16	188264427.6	11766526.72			
14	Total	17	2920783775				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	5794.2888	1079.65324	5.366805359	6.295E-05	3505.526186	8083.05141
18	X Variable 1	17.07793	1.120670001	15.23903548	6.027E-11	14.70221585	19.45364421

FIGURE 14.58
Regression analysis output generated using MS Excel for Example 14.4

in the explanatory variable (advertisement). The standard error is computed as 3430.23. The value of R^2 is an indication of a good predictor regression model.

3. Predicting sales when advertisement is 3000 million rupees

As exhibited in the MS Excel output, the regression equation can be written as:

$$\text{Sales} = 5794.28 + 17.07\ (\text{Advertisement})$$

The predicted sales when advertisement is Rs 3000 million can be computed as

$$\text{Sales} = 5794.28 + 17.07 \times (3000) = 57{,}004.28 \text{ Rs.}$$

Hence, the predicted income is Rs 57,004.28 million, when the advertisement expenditure is Rs 3000 million.

4. Using residual analysis to test the assumptions of the regression model

In order to use residual analysis to test the assumptions of the regression model, we have to test the following four assumptions:

(i) Linearity of the regression model
(ii) Constant error variance (Homoscedasticity)
(iii) Independence of error
(iv) Normality of error

Figure 14.59 is the Minitab generated four-in-one-residual plot, which is mainly used for residual analysis.

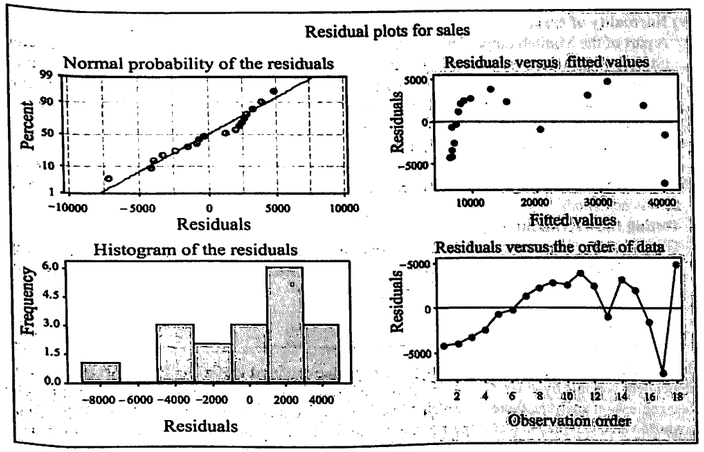


FIGURE 14.59
Four-in-one-residual plot for Example 14.4 generated using Minitab

(i) Linearity of the regression model

Figure 14.60 clearly exhibits that there is no apparent pattern in the plot between residuals and x_i values of the independent variable (advertisement). Hence, the assumption of linearity is not violated.

(ii) Constant error variance (Homoscedasticity)

The assumption of constant error variance or homoscedasticity can be investigated by "residuals versus the fitted values" part of the Minitab graph (Figure 14.59). In this plot, residuals are scattered randomly around zero. Hence, errors have constant variance or there is no violation of the assumption of homoscedasticity.

(iii) Independence of error

For verifying the assumption of independence of error, residuals versus time graph can be plotted. This is shown as "residuals versus the order of the data" in the Minitab output (Figure 14.59). No apparent pattern indicates independence of error.

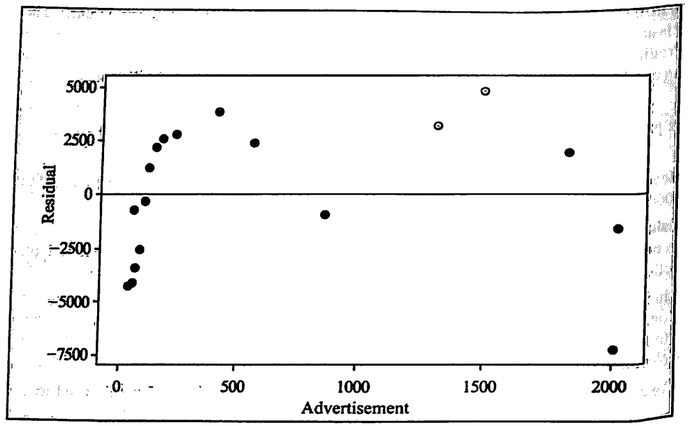


FIGURE 14.60
Minitab output exhibiting a plot between residuals and independent variable (advertisement) for Example 14.4

(iv) Normality of error

A part of the Minitab output "histogram of the residuals" in Figure 14.59 shows a left-skewed normal distribution. By observing "normal probability plot of the residuals" in Figure 14.59 closely, we find that the line connecting all the residuals is not exactly straight but rather close to a straight line. This indicates that the residuals are nearly normal in shape. A curve around the upper part of the line is an indication of skewness.

5. **t Test for the slope of the regression line**
Figure 14.58 shows that the t value is computed as 15.23. The corresponding p-value test (0.000) indicates that this is significant. Hence, the alternative hypothesis that the slope is not equal to zero is accepted.

6. **Testing the overall model**
The ANOVA table is a part of the MS Excel output as shown in Figure 14.58. The computed F value is 232.22. The corresponding p value is 0.0000, which is significant. This p value indicates the significance of the overall model.

SUMMARY

Regression analysis is the process of developing a statistical model which is used to predict the value of a dependent variable by at least one independent variable. In simple linear regression analysis, there are two types of variables. The variable whose value is influenced or is to be predicted is called dependent variable and the variable which influences the value or is used for prediction is called independent variable. Simple linear regression is based on the slope–intercept equation of a line. In regression analysis, sample regression model can be used to make predictions about population parameters. So, β_0 and β_1 (population parameters) are estimated on the basis of sample statistics b_0 and b_1. For this purpose, least squares method is used. Least-squares method use the sample data to determine the values of b_0 and b_1 that minimizes the sum of squared differences between actual values (y_i) and the regressed values ($\hat{y}_i$). Once line of regression is developed, by substituting the required variable values and values of regression coefficient, regressed values, or predicted values can be obtained.

While developing a regression model to predict the dependent variable with the help of independent variable, we need to focus on a few measures of variations. Total variation (SST) can be partitioned in two parts: variation which can be attributed to the relationship between x and y and unexplained variation. First part of variation which can be attributed to the relationship between x and y is referred to as explained variation or regression sum of squares (SSR). The second part of the variation, which is unexplained can be attributed to factors other than the relationship between x and y is referred to as error sum of squares (SSE). Coefficient of determination is also a very important phenomenon in regression analysis. Coefficient of determination measures the proportion of variation in y that can be attributed to independent variable x. A residual is the difference between actual values (y_i) and the regressed values ($\hat{y}_i$) and is used to examine the magnitude of the errors produced by the regression model. In addition, residual analysis can be used to verify the assumptions of regression analysis. These assumptions are (1) linearity of the regression model (2) constant error variance (homoscedasticity) (3) independence of error (4) normality of error.

After verifying the assumptions of linear regression, a researcher determines whether a significant linear relationship between independent variable x and dependent variable y exists. This can be done by performing a hypothesis test to check whether the population slope (β_1) is zero or not. The t test is applied for this purpose. A significant p value for the t statistic establishes the linear relationship between the independent variable x and the dependent variable y. In regression analysis, the F test is used to determine the significance of the overall regression model. More specifically, in case of a multiple regression model, the F test determines that at least one of the regression coefficients is different from zero. In case of simple regression, where predictor is only one, the F test for overall significance tests the same phenomenon as the t-statistic test in simple regression. Apart from coefficient of determination (r^2), regression analysis also provides the correlation coefficient (r), which measures the strength of the relationship between two variables. Correlation coefficient (r) specifies whether there is a significant relationship between two variables. Again t statistic is used to determine the significant relationship between two variables.

KEY TERMS

Autocorrelation, 481
Coefficient of determination (r^2), 470
Correlation coefficient (r), 488
Dependent variable, 458
Durbin–Watson statistic, 481
Error sum of squares (SSE), 470
Homoscedasticity, 476
Independence of error, 477
Independent variable, 458
Least-squares method, 460
Regression sum of squares (SSR), 469
Residual, 471
Standard error, 472
Total sum of squares (SST), 469

NOTES

1. www.tatasteel.com/Company/profile.asp,accessed September 2008.
2. Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed September 2008, reproduced with permission.

DISCUSSION QUESTIONS

1. What is the conceptual framework of simple linear regression and how can we use it for business decision making?
2. Regression analysis is an important tool for forecasting. Explain this statement.
3. What are the assumptions of regression analysis?
4. Write short notes on:
   - Linearity of the regression model
   - Constant error variance (Homoscedasticity)
   - Independence of error
   - Normality of error
5. Explain the concept of regression sum of squares (SSR) and error sum of squares (SSE) in a regression model.
6. Explain the concept of coefficient of determination and standard error of the estimate in a regression model.
7. What is autocorrelation? How can we use Durbin–Watson statistic in detecting autocorrelation.
8. How can we use the t test for determining the statistical significance of the slope of the regression line?
9. How can we test the significance of the overall regression model?
10. How can we use correlation coefficient (r) for determining the statistical significance of the relationship between two variables in a regression model?

NUMERICAL PROBLEMS

1. A large supermarket has adopted a new strategy to increase its sales. It has adopted a few consumer friendly policies and is using video clips of 15 minutes to propagate the new policies. The following table provides data about the number of video clips shown in a randomly selected day and the sales turnover of the supermarket in the corresponding day.

Days	*No. of video clips shown*	*Sales (in thousand rupees)*
1	25	150
2	25	210
3	25	140
4	35	180
5	35	230
6	35	270
7	40	310
8	40	330
9	40	300
10	50	270
11	50	310
12	50	340

(1) Develop a regression model to predict sales from the number of video clips shown.
(2) Calculate the coefficient of determination and interpret it.
(3) Calculate the standard error of the estimate.

2. The HR manager of a multinational company wants to determine the relationship between experience and income of employees. The following data are collected from 14 randomly selected employees.

Employees	*Experience (in years)*	*Income (in thousand rupees)*
1	2	30
2	4	40
3	5	45
4	6	35
5	7	50
6	8	60
7	9	70
8	10	65
9	12	60
10	13	55

11	14	75
12	15	80
13	16	85
14	18	75

(1) Develop a regression model to predict income based on the years of experience.
(2) Calculate the coefficient of determination and interpret it.
(3) Calculate the standard error of the estimate.
(4) Predict the income of an employee who has 22 years of experience.

3. A dealer of a motorcycle company believes that there is a positive relationship between the number of salespeople employed and the increase in the sales of bikes. Data for 14 randomly selected weeks are given in the following table.

Weeks	*No. of salespeople employed*	*Sales (in units)*
1	17	34
2	14	39
3	25	60
4	40	80
5	15	38
6	18	50
7	13	35
8	11	25
9	27	51
10	12	29
11	38	89
12	36	85
13	41	90
14	28	63

(1) Develop a regression model to predict sales from the number of salespeople employed.
(2) Calculate the coefficient of determination and interpret it.
(3) Calculate the standard error of the estimate.
(4) Predict sales when number of salespeople employed are 100.

4. For Problem 3, use residual analysis to verify the following assumption of linear regression:
(1) Linearity of the regression model
(2) Constant error variance (Homoscedasticity)
(3) Normality of error

5. For Problem 3, estimate the following:
(1) t Test for the slope of the regression line
(2) Testing the overall model
(3) Statistical inference about the correlation coefficient of the regression model

6. For Problem 2, estimate the following:
(1) t Test for the slope of the regression line
(2) Testing the overall model
(3) Statistical inference about the correlation coefficient of the regression model

7. The municipal corporation of a newly formed capital city is planning to launch a new water supply scheme for the city. For this, the Municipal Corporation has considered past data on water consumption in 16 randomly selected weeks of the previous summer and the average temperature in the corresponding week. On the basis of the data, the corporation wants to estimate the water requirement for the current year. Data are given as below:

Weeks	*Temperature (in °F)*	*Water consumption (in million gallons)*
1	37	150
2	38	160
3	39	168
4	35	145
5	34	140
6	33	142
7	37	155
8	40	165
9	41	167
10	42	175
11	44	185
12	42	180
13	40	170
14	38	165
15	42	170
16	44	173

(1) Develop a regression model to predict water consumption from the temperature of the corresponding week.
(2) Calculate the coefficient of determination and interpret it.
(3) Calculate the standard error of the estimate.
(4) Predict the water consumption when temperature is 47 °F.
(5) t Test for the slope of the regression line
(6) Test the overall model
(7) Statistical inference about correlation coefficient of the regression model
(8) Calculate Durbin–Watson statistic and interpret it.

8. A company is a concerned about the high rates of absenteeism among its employees. It organized a training programme to boost the morale of its employees. The following table gives the number of days that sixteen randomly selected employees have received training, and the number of days they have availed leave.

Employee	*Training days*	*Leave*
1	12	20
2	14	18
3	16	16
4	13	22
5	11	18
6	10	19
7	15	14
8	17	12
9	18	10
10	19	9
11	17	11
12	15	16
13	13	19
14	15	17
15	17	15
16	12	21

(1) Develop a regression model to predict leaves based on training days.
(2) Calculate the coefficient of determination and state its interpretation.
(3) Calculate the standard error of the estimate.
(4) Predict the leaves when training days are 25.
(5) t Test for the slope of the regression line
(6) Test the overall model
(7) Statistical inference about the correlation coefficient of the regression model
(8) Calculate Durbin–Watson statistic and interpret it.

FORMULAS

Equation of the simple regression line

$$\hat{y} = b_0 + b_1 x$$

Slope of a regression line

$$b_1 = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = \frac{\sum xy - n(\bar{x}\times\bar{y})}{\sum x^2 - n\bar{x}^2} = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

where

$$SS_{xy} = \sum(x-\bar{x})(y-\bar{y}) = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

and

$$SS_{xx} = \sum(x-\bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$b_1 = \frac{SS_{xy}}{SS_{xx}}$$

y Intercept of the regression line

$$b_0 = \bar{y} - b_1\bar{x} = \frac{\sum y}{n} - b_1\frac{(\sum x)}{n}$$

Coefficient of determination (r^2)

$$r^2 = \frac{\text{Regression sum of squares}}{\text{Total sum of squares}} = \frac{\text{SSR}}{\text{SST}}$$

Residual (e_i)

$$\text{Residual } (e_i) = \text{actual values } (y_i) - \text{regressed values } (\hat{y})$$

Standard error of the estimate

$$S_{yx} = \sqrt{\frac{\text{SSE}}{n-2}} = \sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{n-2}}$$

where y_i is the actual value of y, for observation i and $\hat{y}$ the regressed (predicted) value of y, for observation i.

Durbin–Watson statistic

$$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

where e_i is the residual for the time period i and e_{i-1} the residual for the time period $i-1$.

The test statistic t

$$t = \frac{b_1 - \beta_1}{S_b}$$

where

$$S_b = \frac{S_{yx}}{\sqrt{SS_{xx}}}$$

$$S_{yx} = \sqrt{\frac{SSE}{n-2}}$$

$$SS_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

F statistic for testing slope

$$F = \frac{MSR}{MSE}$$

where $MSR = \frac{SSR}{k}$, $MSE = \frac{SSE}{n-k-1}$, and k = Number of independent (explanatory) variables in the regression model (In case of simple regression $k = 1$)

Estimate of confidence interval for the population slope (β_1)

$$b_1 \pm t_{n-2} S_b$$

t statistic for testing the statistical significant correlation coefficient

$$t = \frac{r - \rho}{\sqrt{\frac{1-r^2}{n-2}}}$$

where

$$r = +\sqrt{r^2}, \text{ if } b_1 \geq 0 \text{ and } r = -\sqrt{r^2}, \text{ if } b_1 < 0.$$

CASE STUDY

Case 14: *Boom in the Indian Cement Industry: ACC's Role*

Introduction

The Indian cement industry was delicensed in 1991. After China, India is the second largest producer of cement. The estimated demand for cement is 265 million metric tonnes by 2114–2115.[1] The Indian cement industry saw a growth of 11.6% in 2006. The financial year 2007 also witnessed a muted growth of 7.1%. In order to meet the increasing demand, several manufacturers have embarked on significant capacity expansion plans.[2]

ACC—A Pioneer in the Indian Cement Industry

Associated Cement Companies Ltd (ACC) came into existence in 1936, after the merger of 10 companies belonging to four important business groups: Tatas, Khataus, Killick Nixon, and F E Dinshaw. The Tata group was associated with ACC since its inception. It sold 14.45% of its share to Gujarat Ambuja Cements Ltd between 1999 and 2000. After this strategic alliance, Gujarat Ambuja Cements Ltd became the largest single stakeholder in ACC. In 2005, ACC entered into a strategic relationship with the Holcim group of Switzerland, a world leader in cement as well as a large supplier of concrete, aggregates, and certain construction related services. These global strategic alliances have strengthened the company.[3]

ACC is India's foremost manufacturer of cement and concrete. The company has a wide range of operations with 14 modern cement factories, more than 30 ready mix concrete plants, 20 sales offices, and several zonal offices. ACC's research and development facility has a unique track record of innovative research, product development, and specialized consultancy services. ACC's brand name is synonymous with cement and it enjoys a high level of equity in the Indian market.[4]

The Impact of Cartelization

Cartelization is one of the major problems in the cement industry. Cartelization takes place when dominant players of the industry join together to control prices and limit competition. In the Indian market, manufacturers have been known to enter into agreements to artificially limit the supply of cement so that the price remains high. When markets are not sufficiently regulated, large companies may be tempted to collude instead of competing with each other. For example, in May 2006, the Competition Council of Romania imposed a combined fine of 27 million euros on France's Lafarge, Switzerland's Holcim, and Germany's Carpatcement for being involved in the cement cartel in the Romanian market. These three companies share 98% of Romanian cement capacity.[4] The government should take appropriate action to check acts of cartelization.

Escalating input and fuel costs have forced manufacturers to tap new sources of supply and increase the quest for alternative fuels and raw materials. The cement industry is faced with the challenge of optimizing the utilization of scare basic raw materials and fossil fuels while simultaneously protecting the environment and maintaining emission levels within acceptable limits. It is vital for the cement industry to achieve high levels of energy utilization efficiencies and to sustain them continuously.[2] Table 14.01 exhibits sales turnover and advertisement expenses of ACC from 1995 to 2007.

TABLE 14.01
Sales turnover and advertisement expenditure of ACC from 1995–2007

Year	*Sales (in million rupees)*	*Advertisement (rs in million rupees)*
1995	20 ,427.0	58.6
1996	23 ,294.6	72.6
1997	24 ,510.5	122.3
1998	23 ,731.1	61.9
1999	25 ,858.3	144.7
2000	26 ,792.2	132.2
2001	29 ,361.2	172.6
2002	32 ,260.0	184.3
2003	33 ,718.8	259.8
2004	39 ,003.7	334.8
2005	45 ,498.0	321.9
2006	37 ,235.1	336.0
2007	64 ,680.6	442.3

1. Develop an appropriate regression model to predict sales from advertisement.
2. Calculate the coefficient of determination and state its interpretation.
3. Calculate the standard error of the estimate.
4. Predict the sales when advertisement is Rs 500 million.
5. Test the significance of the overall model.

NOTES

1. www.indiastat.com, accessed September 2008, reproduced with permission.
2. Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, accessed September 2008, reproduced with permission.
3. www.acclimited.com/newsite/heritage.asp, accessed September 2008.
4. www.acclimited.com/newsite/corprofile.asp, accessed September 2008.
5. ww.businesstoday.org/index.php?option=com_content&task=viewed&id=370&Itemi, accessed September 2008.