Chapter 14

Simple Linear Regression

Learning objectives

After reading this chapter and doing the exercises, you should be able to:

1. Understand how regression analysis can be used to develop an equation that estimates mathematically how two variables are related.
2. Understand the differences between the regression model, the regression equation, and the estimated regression equation.
3. Know how to fit an estimated regression equation to a set of sample data based upon the least-squares method.
4. Determine how good a fit is provided by the estimated regression equation and compute the sample correlation coefficient from the regression analysis output.
5. Understand the assumptions necessary for statistical inference and be able to test for a significant relationship.
6. Know how to develop confidence interval estimates of the mean value of Y and an individual value of Y for a given value of X.
7. Learn how to use a residual plot to make a judgment as to the validity of the regression assumptions, recognise outliers and identify influential observations.
8. Use the Durbin-Watson test to test for autocorrelation.
9. Know the definition of the following terms:
   independent and dependent variable
   simple linear regression
   regression model
   regression equation and estimated regression equation
   scatter diagram
   coefficient of determination
   standard error of the estimate
   confidence interval
   prediction interval
   residual plot
   standardized residual plot
   outlier
   influential observation
   leverage

Managerial decisions are often based on the relationship between two or more variables. For example, after considering the relationship between advertising expenditures and sales, a marketing manager might attempt to predict sales for a given level of advertising expenditure. In another case, a public utility might use the relationship between the daily high temperature and the demand for electricity to predict electricity usage on the basis of next month's anticipated daily high temperatures. Sometimes a manager will rely on intuition to judge how two variables are related. However, if data can be obtained, a statistical procedure called *regression analysis* can be used to develop an equation showing how the variables are related.

In regression terminology, the variable being predicted is called the **dependent variable**. The variable or variables being used to predict the value of the dependent variable are called the **independent variables**. For example, in analyzing the effect of advertising expenditures on sales, a marketing manager's desire to predict sales would suggest making sales the dependent variable. Advertising expenditure would be the independent variable used to help predict sales. In statistical notation, Y denotes the dependent variable and X denotes the independent variable.

Statistics in Practice

Foreign direct investment (FDI) in China

In a recent study by Kingston Business School, regression modelling was used to investigate patterns of FDI in China as well as to assess the particular potential of the autonomous region of Guangxi in SW China as an FDI attractor. A variety of simple models were developed based on positive correlations between GDP and FDI in provinces using data collected from official statistical sources.

Estimated regression equations obtained were as follows:

	$\hat{y} = 1.1m + 21.7x$	1990–1993
	$\hat{y} = 2.1m + 8.9x$	1995–1998
	$\hat{y} = 3.3m + 14.6x$	2000–2003
where	$\hat{y}$ = estimated GDP	
	$x = FDI$	

across all provinces.

In terms of FDI *per capita*, Guangxi has been ranked around 27 of 31 over the last ten years or so. FDI is a key driver of economic growth in modern China. But clearly Guangxi needs to improve its ranking if it is to be able to compete effectively with the more successful eastern coastal provinces and great municipalities.

American coffee shop Starbucks in Shanghai, China. Keren Su/China Span/Alamy.

Source: Foster MJ (2002) 'On evaluation of FDI's: Principles, Actualities and Possibilities' *International Journal of Management and Decision-Making* 3(1) 67–82

In this chapter we consider the simplest type of regression analysis involving one independent variable and one dependent variable in which the relationship between the variables is approximated by a straight line. It is called **simple linear regression**. Regression analysis involving two or more independent variables is called *multiple regression analysis*; multiple regression and cases involving curvilinear relationships are covered in Chapters 15 and 16.

14.1 Simple linear regression model

Armand's Pizza Parlours is a chain of Italian-food restaurants located in northern Italy. Armand's most successful locations are near college campuses. The managers believe that quarterly sales for these restaurants (denoted by Y) are related positively to the size of the student population (denoted by X); that is, restaurants near campuses with a large student population tend to generate more sales than those located near campuses with a small student population. Using regression analysis, we can develop an equation showing how the dependent variable Y is related to the independent variable X.

Regression model and regression equation

In the Armand's Pizza Parlours example, the population consists of all the Armand's restaurants.

For every restaurant in the population, there is a value x of X (student population) and a corresponding value y of Y (quarterly sales). The equation that describes how Y is related to x and an error term is called the **regression model**. The regression model used in simple linear regression follows.

Simple linear regression model

$$Y = \beta_0 + \beta_1 x + \varepsilon \qquad (14.1)$$

β_0 and β_1 are referred to as the parameters of the model, and ε (the Greek letter epsilon) is a random variable referred to as the *error term*. The error term ε accounts for the variability in Y that cannot be explained by the linear relationship between X and Y.

The population of all Armand's restaurants can also be viewed as a collection of subpopulations, one for each distinct value of X. For example, one subpopulation consists of all Armand's restaurants located near college campuses with 8000 students; another subpopulation consists of all Armand's restaurants located near college campuses with 9000 students and so on. Each subpopulation has a corresponding distribution of Y values. Thus, a distribution of Y values is associated with restaurants located near campuses with 8000 students a distribution of Y values is associated with restaurants located near campuses with 9000 students and so on. Each distribution of Y values has its own mean or expected value. The equation that describes how the expected value of Y – denoted by $E(Y)$ or equivalently $E(Y|X = x)$ – is related to x is called the **regression equation**. The regression equation for simple linear regression follows.

Simple linear regression equation

$$E(Y) = \beta_0 + \beta_1 x \qquad (14.2)$$

The graph of the simple linear regression equation is a straight line; β_0 is the y-intercept of the regression line, β_1 is the slope and $E(Y)$ is the mean or expected value of Y for a given value of X.

Examples of possible regression lines are shown in Figure 14.1. The regression line in Panel A shows that the mean value of Y is related positively to X, with larger values of

Figure 14.1 Possible regression lines in simple linear regression

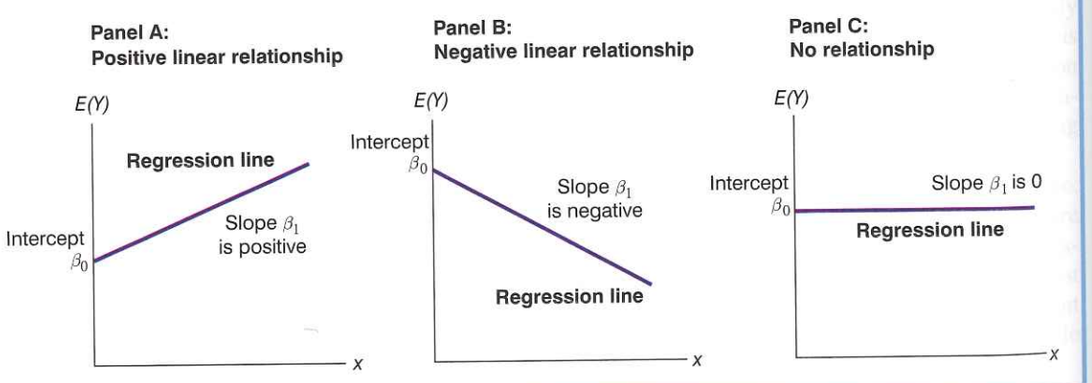
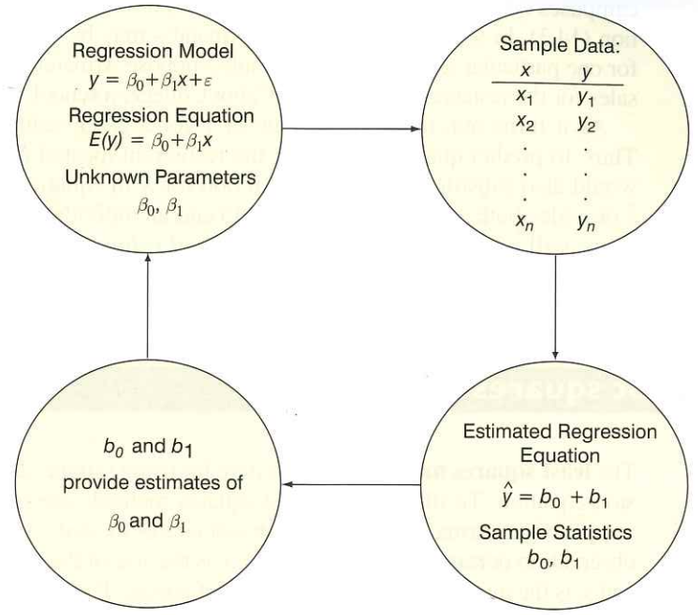


Figure 14.2 The estimation process in simple linear regression



$E(Y)$ associated with larger values of X. The regression line in Panel B shows the mean value of Y is related negatively to X, with smaller values of $E(Y)$ associated with larger values of X. The regression line in Panel C shows the case in which the mean value of Y is not related to X; that is, the mean value of Y is the same for every value of X.

Estimated regression equation

If the values of the population parameters β_0 and β_1 were known, we could use equation (14.2) to compute the mean value of Y for a given value of X. In practice, the parameter values are not known, and must be estimated using sample data. Sample statistics (denoted b_0 and b_1) are computed as estimates of the population parameters β_0 and β_1. Substituting the values of the sample statistics b_0 and b_1 for β_0 and β_1 in the regression equation, we obtain the **estimated regression equation**. The estimated regression equation for simple linear regression follows.

Estimated simple linear regression equation

$$\hat{y} = b_0 + b_1 x \qquad (14.3)$$

The graph of the estimated simple linear regression equation is called the *estimated regression line*; b_0 is the y intercept and b_1 is the slope. In the next section, we show how the least squares method can be used to compute the values of b_0 and b_1 in the estimated regression equation.

In general, $\hat{y}$ is the point estimator of $E(Y)$, the mean value of Y for a given value of X. Thus, to estimate the mean or expected value of quarterly sales for all restaurants located near campuses with 10 000 students, Armand's would substitute the value of 10 000 for X in equation (14.3). In some cases, however, Armand's may be more interested in predicting sales for one particular restaurant. For example, suppose Armand's would like to predict quarterly sales for the restaurant located near Cabot College, a school with 10 000 students.

As it turns out, the best estimate of Y for a given value of X is also provided by $\hat{y}$. Thus, to predict quarterly sales for the restaurant located near Cabot College, Armand's would also substitute the value of 10 000 for X in equation (14.3). Because the value of $\hat{y}$ provides both a point estimate of $E(Y)$ and an individual value of Y for a given value of X, we will refer to $\hat{y}$ simply as the *estimated value of* y.

Figure 14.2 provides a summary of the estimation process for simple linear regression.

14.2 Least squares method

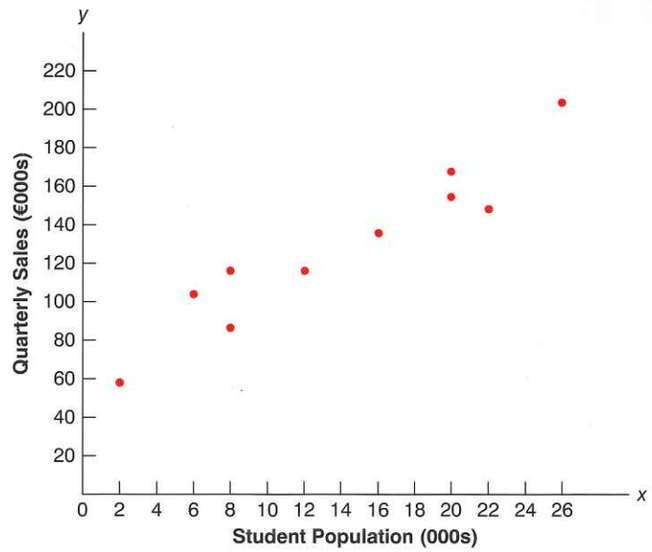
The **least squares method** is a procedure for using sample data to find the estimated regression equation. To illustrate the least squares method, suppose data were collected from a sample of ten Armand's Pizza Parlour restaurants located near college campuses. For the ith observation or restaurant in the sample, x_i is the size of the student population (in thousands) and y_i is the quarterly sales (in thousands of euros). The values of x_i and y_i for the ten restaurants in the sample are summarized in Table 14.1. We see that restaurant 1, with $x_1 = 2$ and $y_1 = 58$, is near a campus with 2000 students and has quarterly sales of €58 000. Restaurant 2, with $x_2 = 6$ and $y_2 = 105$, is near a campus with 6000 students and has quarterly sales of €105 000. The largest sales value is for restaurant 10, which is near a campus with 26 000 students and has quarterly sales of €202 000.

Figure 14.3 is a scatter diagram of the data in Table 14.1. Student population is shown on the horizontal axis and quarterly sales are shown on the vertical axis. **Scatter diagrams** for regression analysis are constructed with the independent variable X on the horizontal axis and the dependent variable Y on the vertical axis. The scatter diagram enables us to observe the data graphically and to draw preliminary conclusions about the possible relationship between the variables.

Table 14.1 Student population and quarterly sales data for ten Armand's Pizza Parlours

Restaurant i	Student population (000s) x_i	Quarterly sales (€000s) y_i
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

Figure 14.3 Scatter diagram of student population and quarterly sales for Armand's Pizza Parlours



What preliminary conclusions can be drawn from Figure 14.3? Quarterly sales appear to be higher at campuses with larger student populations. In addition, for these data the relationship between the size of the student population and quarterly sales appears to be approximated by a straight line; indeed, a positive linear relationship is indicated between X and Y. We therefore choose the simple linear regression model to represent the relationship between quarterly sales and student population. Given that choice, our next task is to use the sample data in Table 14.1 to determine the values of b_0 and b_1 in the estimated simple linear regression equation. For the ith restaurant, the estimated regression equation provides

$$\hat{y}_i = b_0 + b_1 x_i \qquad (14.4)$$

where

$\hat{y}_i$ = estimated value of quarterly sales (€000s) for the ith restaurant
b_0 = the y intercept of the estimated regression line
b_1 = the slope of the estimated regression line
x_i = size of the student population (000s) for the ith restaurant

Every restaurant in the sample will have an observed value of sales y_i and an estimated value of sales $\hat{y}_i$. For the estimated regression line to provide a good fit to the data, we want the differences between the observed sales values and the estimated sales values to be small.

The least squares method uses the sample data to provide the values of b_0 and b_1 that minimize the *sum of the squares of the deviations* between the observed values of the dependent variable y_i and the estimated values of the dependent variable. The criterion for the least squares method is given by expression (14.5).

Least squares criterion

$$\text{Min } \Sigma(y_i - \hat{y}_i)^2 \qquad (14.5)$$

where

y_i = observed value of the dependent variable for the ith observation
$\hat{y}_i$ = estimated value of the dependent variable for the ith observation

Differential calculus can be used to show that the values of b_0 and b_1 that minimize expression (14.5) can be found by using equations (14.6) and (14.7).

Slope and y-intercept for the estimated regression equation*

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - y)}{\Sigma(x_i - \bar{x})^2} \qquad (14.6)$$

$$b_0 = \bar{y} - b_1\bar{x} \qquad (14.7)$$

where

x_i = value of the independent variable for the ith observation
Y_i = value of the dependent variable for the ith observation
$\bar{x}$ = mean value for the independent variable
$\bar{y}$ = mean value for the dependent variable
n = total number of observations

Some of the calculations necessary to develop the least squares estimated regression equation for Armand's Pizza Parlours are shown in Table 14.2. With the sample of ten restaurants, we have $n = 10$ observations. Because equations (14.6) and (14.7) require $\bar{x}$ and $\bar{y}$ we begin the calculations by computing $\bar{x}$ and $\bar{y}$.

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{140}{10} = 14$$

$$\bar{y} = \frac{\Sigma y_i}{n} = \frac{1300}{10} = 130$$

Using equations (14.6) and (14.7) and the information in Table 14.2, we can compute the slope and intercept of the estimated regression equation for Armand's Pizza Parlours. The calculation of the slope (b_1) proceeds as follows.

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{2840}{568} = 5$$

*An alternative formula for b_1 is

$$b_1 = \frac{\Sigma x_i y_i - (\Sigma x_i \Sigma y_i)n}{\Sigma x^2 - (\Sigma x_i)^2 n}$$

This form of equation (14.6) is often recommended when using a calculator to compute b_1.

Table 14.2 Calculations for the least squares estimated regression equation for Armand's Pizza Parlours

Restaurant i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	2	58	−12	−72	864	144
2	6	105	−8	−25	200	64
3	8	88	−6	−42	252	36
4	8	118	−6	−12	72	36
5	12	117	−2	−13	26	4
6	16	137	2	7	14	4
7	20	157	6	27	162	36
8	20	169	6	39	234	36
9	22	149	8	19	152	64
10	26	202	12	72	864	144
Totals	140	1300			2840	568
	Σx_i	Σy_i			$\Sigma(x_i - \bar{x})(y_i - \bar{y})$	$\Sigma(x_i - \bar{x})^2$

The calculation of the y intercept (b_0) follows.

$$b_0 = \bar{y} - b_1\bar{x} = 130 - 5(14) = 60$$

Thus, the estimated regression equation is

$$\hat{y} = 60 + 5x$$

Figure 14.4 shows the graph of this equation on the scatter diagram.

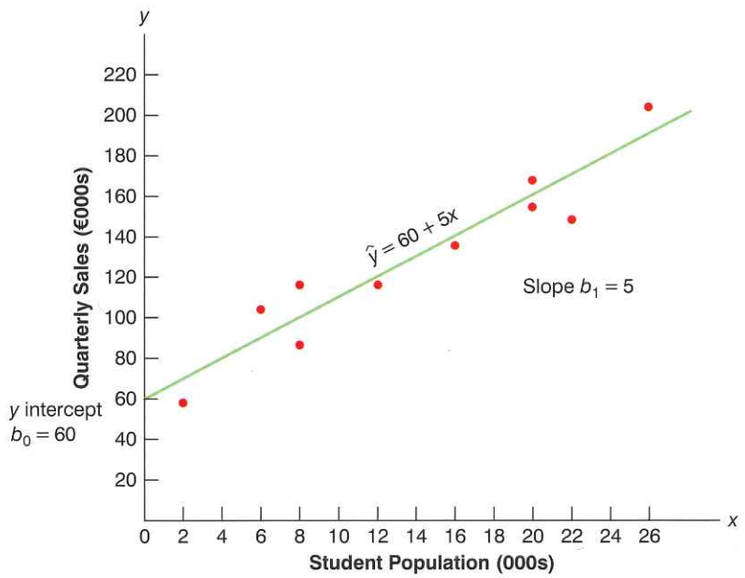
The slope of the estimated regression equation ($b_1 = 5$) is positive, implying that as student population increases, sales increase. In fact, we can conclude (based on sales measured in €000s and student population in 000s) that an increase in the student population of 1000 is associated with an increase of €5000 in expected sales; that is, quarterly sales are expected to increase by €5 per student.

If we believe the least squares estimated regression equation adequately describes the relationship between X and Y, it would seem reasonable to use the estimated regression equation to predict the value of Y for a given value of X. For example, if we wanted to predict quarterly sales for a restaurant to be located near a campus with 16 000 students, we would compute

$$\hat{y} = 60 + 5(16) = 140$$

Therefore, we would predict quarterly sales of €140 000 for this restaurant. In the following sections we will discuss methods for assessing the appropriateness of using the estimated regression equation for estimation and prediction.

Figure 14.4 Graph of the estimated regression equation for Armand's Pizza Parlours $\hat{y} = 60 + 5x$



Exercises

Methods

1 Given are five observations for two variables, X and Y

x_i	1	2	3	4	5
y_i	3	7	5	11	14

a. Develop a scatter diagram for these data.
b. What does the scatter diagram developed in part (a) indicate about the relationship between the two variables?
c. Try to approximate the relationship between X and Y by drawing a straight line through the data.
d. Develop the estimated regression equation by computing the values of b_0 and b_1 using equations (14.6) and (14.7).
e. Use the estimated regression equation to predict the value of Y when $X = 4$.

2 Given are five observations for two variables, X and Y.

x_i	2	3	5	1	8
y_i	25	25	20	30	16

a. Develop a scatter diagram for these data.
b. What does the scatter diagram developed in part (a) indicate about the relationship between the two variables?

c. Try to approximate the relationship between X and Y by drawing a straight line through the data.
d. Develop the estimated regression equation by computing the values of b_0 and b_1 using equations (14.6) and (14.7).
e. Use the estimated regression equation to predict the value of Y when $X = 6$.

3 Given are five observations collected in a regression study on two variables.

x_i	2	4	5	7	8
y_i	2	3	2	6	4

a. Develop a scatter diagram for these data.
b. Develop the estimated regression equation for these data.
c. Use the estimated regression equation to predict the value of Y when $X = 4$.

Applications

4 The following data were collected on the height (cm) and weight (kg) of women swimmers.

Height	173	163	157	165	168
Weight	60	49	46	52	58

a. Develop a scatter diagram for these data with height as the independent variable.
b. What does the scatter diagram developed in part (a) indicate about the relationship between the two variables?
c. Try to approximate the relationship between height and weight by drawing a straight line through the data.
d. Develop the estimated regression equation by computing the values of b_0 and b_1.
e. If a swimmer's height is 160 cm, what would you estimate her weight to be?

5 The Dow Jones Industrial Average (DJIA) and the Standard & Poor's 500 (S&P) indexes are both used as measures of overall movement in the stock market. The DJIA is based on the price movements of 30 large companies; the S&P 500 is an index composed of 500 stocks. Some say the S&P 500 is a better measure of stock market performance because it is broader based. The closing prices for the DJIA and the S&P 500 for ten weeks, beginning with 11 February 2009, follow (*uk.finance.yahoo.com*, 21 April 2009).

DOWS&P

Date	DJIA	S&P
11 Feb 09	7939.53	833.74
18 Feb 09	7555.63	788.42
25 Feb 09	7270.89	764.90
03 Mar 09	6726.02	696.33
10 Mar 09	6926.49	719.60
17 Mar 09	7395.70	778.12
24 Mar 09	7660.21	806.12
31 Mar 09	7608.92	797.87
07 Apr 09	7789.56	815.55
14 Apr 09	7920.18	841.50

a. Develop a scatter diagram for these data with DJIA as the independent variable.
b. Develop the least squares estimated regression equation.
c. Suppose the closing price for the DJIA is 8000. Estimate the closing price for the S&P 500.

6 A sales manager collected the following data on annual sales and years of experience.

Salesperson	Years of experience	Annual sales (€000s)
1	1	80
2	3	97
3	4	92
4	4	102
5	6	103
6	8	111
7	10	119
8	10	123
9	11	117
10	13	136

a. Develop a scatter diagram for these data with years of experience as the independent variable.
b. Develop an estimated regression equation that can be used to predict annual sales given the years of experience.
c. Use the estimated regression equation to predict annual sales for a salesperson with nine years of experience.

14.3 Coefficient of determination

For the Armand's Pizza Parlours example, we developed the estimated regression equation $\hat{y} = 60 + 5x$ to approximate the linear relationship between the size of student population X and quarterly sales Y. A question now is: How well does estimated regression equation fit the data? In this section, we show that **coefficient of determination** provides a measure of the goodness of fit for the estimated regression equation.

For the ith observation, the difference between the observed value of the dependent variable, y_i, and the estimated value of the dependent variable, $\hat{y}_i$, is called the ***i*th residual**. The ith residual represents the error in using y_i to estimate $\hat{y}_i$. Thus, for the ith observation, the residual is $y_i - \hat{y}_i$. The sum of squares of these residuals or errors is the quantity that is minimized by the least squares method. This quantity, also known as the *sum of squares due to error*, is denoted by SSE.

Sum of squares due to error

$$\text{SSE} = \Sigma(y_i - \hat{y}_i)^2 \qquad (14.8)$$

The value of SSE is a measure of the error in using the least squares regression equation to estimate the values of the dependent variable in the sample.

In Table 14.3 we show the calculations required to compute the sum of squares due to error for the Armand's Pizza Parlours example. For instance, for restaurant 1 the values of the independent and dependent variables are $x_1 = 2$ and $y_1 = 58$. Using the estimated regression equation, we find that the estimated value of quarterly sales for restaurant 1 is

Table 14.3 Calculation of SSE for Armand's Pizza Parlours

Restaurant i	x_i = Student population (000s)	y_i = Quarterly sales (€000s)	Predicted sales $\hat{y}_i = 60 + 5x_i$	Error $y_i - \hat{y}_i$	Squared error $(y_i - \hat{y}_i)^2$
1	2	58	70	−12	144
2	6	105	90	15	225
3	8	88	100	−12	144
4	8	118	100	18	324
5	12	117	120	−3	9
6	16	137	140	−3	9
7	20	157	160	−3	9
8	20	169	160	9	81
9	22	149	170	−21	441
10	26	202	190	12	144
					SSE = 1530

$\hat{y}_1 = 60 + 5(2) = 70$. Thus, the error in using $\hat{y}_1$ to estimate y_1 for restaurant 1 is $y_1 - \hat{y}_1 = 58 - 70 = -12$. The squared error, $(-12)^2 = 144$, is shown in the last column of Table 14.3. After computing and squaring the residuals for each restaurant in the sample, we sum them to obtain SSE = 1530. Thus, SSE = 1530 measures the error in using the estimated regression equation $\hat{y}_1 = 60 + 5x$ to predict sales.

Now suppose we are asked to develop an estimate of quarterly sales without knowledge of the size of the student population. Without knowledge of any related variables, we would use the sample mean as an estimate of quarterly sales at any given restaurant. Table 14.2 shows that for the sales data, $\Sigma y_i = 1300$. Hence, the mean value of quarterly sales for the sample of ten Armand's restaurants is $\bar{y} = \Sigma y_i/n = 1300/10 = 130$. In Table 14.4 we show the sum of squared deviations obtained by using the sample mean $\bar{y} = 130$ to estimate the value of quarterly sales for each restaurant in the sample. For the ith restaurant in the sample, the difference $y_i - \bar{y}$ provides a measure of the error involved in using $\bar{y}$ to estimate sales. The corresponding sum of squares, called the *total sum of squares*, is denoted SST.

Total sum of squares

$$\text{SST} = \Sigma(y_i - \bar{y})^2 \qquad (14.9)$$

The sum at the bottom of the last column in Table 14.4 is the total sum of squares for Armand's Pizza Parlours; it is SST = 15 730.

In Figure 14.5 we show the estimated regression line $\hat{y}_i = 60 + 5x$ and the line corresponding to $\bar{y} = 130$. Note that the points cluster more closely around the estimated regression line than they do about the line $\bar{y} = 130$. For example, for the tenth restaurant in the sample we see that the error is much larger when $\bar{y} = 130$ is used as an estimate of y_{10} than when $\hat{y}_i = 60 + 5(26) = 190$ is used. We can think of SST as a measure of how well the observations cluster about the y line and SSE as a measure of how well the observations cluster about the $\hat{y}$ line.

Table 14.4 Computation of the total sum of squares for Armand's Pizza Parlours

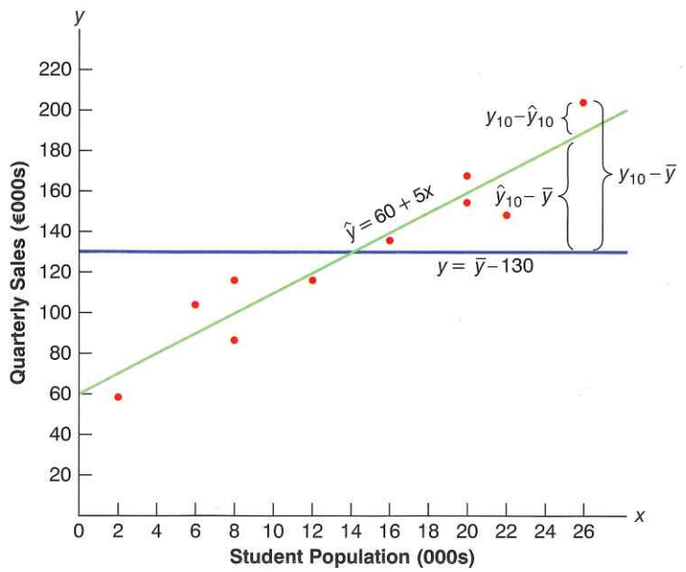
Restaurant i	x_i = Student population (000s)	y_i = Quarterly sales (€000s)	Deviation $y_i - \bar{y}$	Squared deviation $(y_i - \bar{y})^2$
1	2	58	−72	5 184
2	6	105	−25	625
3	8	88	−42	1 764
4	8	118	−12	144
5	12	117	−13	169
6	16	137	7	49
7	20	157	27	729
8	20	169	39	1 521
9	22	149	19	361
10	26	202	72	5 184
				SST = 15 730

To measure how much the $\hat{y}$ values on the estimated regression line deviate from $\bar{y}$, another sum of squares is computed. This sum of squares, called the *sum of squares due to regression*, is denoted SSR.

Sum of squares due to regression

$$SSR = \Sigma(\hat{y}_i - \bar{y})^2 \qquad (14.10)$$

Figure 14.5 Deviations about the estimated regression line and the line $y = \bar{y}$ for Armand's Pizza Parlours



From the preceding discussion, we should expect that SST, SSR and SSE are related. Indeed, the relationship among these three sums of squares provides one of the most important results in statistics.

Relationship among SST, SSR and SSE

$$SST = SSR + SSE \qquad (14.11)$$

where

SST = total sum of squares
SSR = sum of squares due to regression
SSE = sum of squares due to error

Equation (14.11) shows that the total sum of squares can be partitioned into two components, the regression sum of squares and the sum of squares due to error. Hence, if the values of any two of these sum of squares are known, the third sum of squares can be computed easily. For instance, in the Armand's Pizza Parlours example, we already know that SSE = 1530 and SST = 15 730; therefore, solving for SSR in equation (14.11), we find that the sum of squares due to regression is

$$SSR = SST - SSE = 15\,730 - 1530 = 14\,200$$

Now let us see how the three sums of squares, SST, SSR and SSE, can be used to provide a measure of the goodness of fit for the estimated regression equation. The estimated regression equation would provide a perfect fit if every value of the dependent variable y_i happened to lie on the estimated regression line. In this case, $y_i - \hat{y}_i$ would be zero for each observation, resulting in SSE = 0. Because SST = SSR + SSE, we see that for a perfect fit SSR must equal SST and the ratio (SSR/SST) must equal one. Poorer fits will result in larger values for SSE. Solving for SSE in equation (14.11), we see that SSE = SST − SSR. Hence, the largest value for SSE (and hence the poorest fit) occurs when SSR = 0 and SSE = SST. The ratio SSR/SST, which will take values between zero and one, is used to evaluate the goodness of fit for the estimated regression equation. This ratio is called the *coefficient of determination* and is denoted by r^2.

Coefficient of determination

$$r^2 = \frac{SSR}{SST} \qquad (14.12)$$

For the Armand's Pizza Parlours example, the value of the coefficient of determination is

$$r^2 = \frac{SSR}{SST} = \frac{14\,200}{15\,730} = 0.9027$$

When we express the coefficient of determination as a percentage, r^2 can be interpreted as the percentage of the total sum of squares that can be explained by using the estimated regression equation. For Armand's Pizza Parlours, we can conclude that 90.27 per cent of the total sum of squares can be explained by using the estimated regression equation $\hat{y} = 60 + 5x$ to predict quarterly sales. In other words, 90.27 per cent of the variability in sales can be explained by the linear relationship between the size of the student population and sales. We should be pleased to find such a good fit for the estimated regression equation.

Correlation coefficient

In Chapter 3 we introduced the **correlation coefficient** as a descriptive measure of the strength of linear association between two variables, X and Y. Values of the correlation coefficient are always between -1 and $+1$. A value of $+1$ indicates that the two variables X and Y are perfectly related in a positive linear sense. That is, all data points are on a straight line that has a positive slope. A value of -1 indicates that X and Y are perfectly related in a negative linear sense, with all data points on a straight line that has a negative slope. Values of the correlation coefficient close to zero indicate that X and Y are not linearly related.

In Section 3.5 we presented the equation for computing the sample correlation coefficient. If a regression analysis has already been performed and the coefficient of determination r^2 computed, the sample correlation coefficient can be computed as follows.

Sample correlation coefficient

$$r_{XY} = (\text{sign of } b_1)\sqrt{\text{Coefficient of determination}}$$
$$= (\text{sign of } b_1)\sqrt{r^2} \qquad (14.13)$$

where

b_1 = the slope of the estimated regression equation $\hat{y} = b_0 + b_1 x$

The sign for the sample correlation coefficient is positive if the estimated regression equation has a positive slope ($b_1 > 0$) and negative if the estimated regression equation has a negative slope ($b_1 < 0$).

For the Armand's Pizza Parlour example, the value of the co efficient of determination corresponding to the estimated regression equation $\hat{y} = 60 + 5x$ is 0.9027. Because the slope of the estimated regression equation is positive, equation (14.13) shows that the sample correlation coefficient is $= \sqrt{0.9027} = +0.9501$.

With a sample correlation coefficient of $r_{XY} = +0.9501$, we would conclude that a strong positive linear association exists between X and Y.

In the case of a linear relationship between two variables, both the coefficient of determination and the sample correlation coefficient provide measures of the strength of the relationship. The coefficient of determination provides a measure between zero and one whereas the sample correlation coefficient provides a measure between -1 and $+1$. Although the sample correlation coefficient is restricted to a linear relationship between two variables, the coefficient of determination can be used for nonlinear relationships and for relationships that have two or more independent variables. Thus, the coefficient of determination provides a wider range of applicability.

Exercises

Methods

7 The data from exercise 1 follow.

x_i	1	2	3	4	5
y_i	3	7	5	11	14

The estimated regression equation for these data is $\hat{y} = 0.20 + 2.60x$.

a. Compute SSE, SST and SSR using equations (14.8), (14.9) and (14.10).
b. Compute the coefficient of determination r^2. Comment on the goodness of fit.
c. Compute the sample correlation coefficient.

8 The data from exercise 2 follow.

x_i	2	3	5	1	8
y_i	25	25	20	30	16

The estimated regression equation for these data is $\hat{y} = 30.33 - 1.88x$.

a. Compute SSE, SST and SSR.
b. Compute the coefficient of determination r^2. Comment on the goodness of fit.
c. Compute the sample correlation coefficient.

9 The data from exercise 3 follow.

x_i	2	4	5	7	8
y_i	2	3	2	6	4

The estimated regression equation for these data is $\hat{y} = 0.75 + 0.51x$. What percentage of the total sum of squares can be accounted for by the estimated regression equation? What is the value of the sample correlation coefficient?

Applications

10 The estimated regression equation for the data in exercise 5 can be shown to be $\hat{y} = -75.586 + 0.115x$. What percentage of the total sum of squares can be accounted for by the estimated regression equation?

Comment on the goodness of fit. What is the sample correlation coefficient?

11 An important application of regression analysis in accounting is in the estimation of cost. By collecting data on volume and cost and using the least squares method to develop an estimated regression equation relating volume and cost, an accountant can estimate the cost associated with a particular manufacturing volume. Consider the following sample of production volumes and total cost data for a manufacturing operation.

Production volume (units)	Total cost (€)
400	4000
450	5000
550	5400
600	5900
700	6400
750	7000

a. Use these data to develop an estimated regression equation that could be used to predict the total cost for a given production volume.
b. What is the variable cost per unit produced?
c. Compute the coefficient of determination. What percentage of the variation in total cost can be explained by production volume?
d. The company's production schedule shows 500 units must be produced next month. What is the estimated total cost for this operation?

12 *PCWorld* provided details for ten of the most economical laser printers (*PCWorld*, April 2009). The following data show the maximum printing speed in pages per minute (ppm) and the price (in euros including 15 per cent value added tax) for each printer.

PRINTERS-2009

Name	Speed (ppm)	Price (€)
Brother HL 2035	18	61.35
HP Laserjet P1005	15	70.13
Samsung ML-1640	16	77.39
HP Laserjet P1006	17	82.93
Brother HL-2140	22	92.34
Brother DCP7030	22	96.04
HP Laserjet P1009	16	99.52
HP Laserjet P1505	24	119.10
Samsung 4300	18	121.64
Epson EPL-6200 Mono	20	133.53

a. Develop the estimated regression equation with speed as the independent variable.
b. Compute r^2. What percentage of the variation in cost can be explained by the printing speed?
c. What is the sample correlation coefficient between speed and price? Does it reflect a strong or weak relationship between printing speed and cost?

14.4 Model assumptions

We saw in the previous section that the value of the coefficient of determination (r^2) is a measure of the goodness of fit of the estimated regression equation. However, even with a large value of r^2, the estimated regression equation should not be used until further analysis of the appropriateness of the assumed model has been conducted. An important step in determining whether the assumed model is appropriate involves testing for the significance of the relationship. The tests of significance in regression analysis are based on the following assumptions about the error term ε.

Assumptions about the error term ε in the regression model

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

1 The error term ε is a random variable with a mean or expected value of zero; that is, $E(\varepsilon) = 0$.

Implication: β_0 and β_1 are constants, therefore $E(\beta_0) = \beta_0$ and $E(\beta_1) = \beta_1$; thus, for a given value x of X, the expected value of Y is

$$E(Y) = \beta_0 + \beta_1 x \qquad \textbf{(14.14)}$$

As we indicated previously, equation (14.14) is referred to as the regression equation.

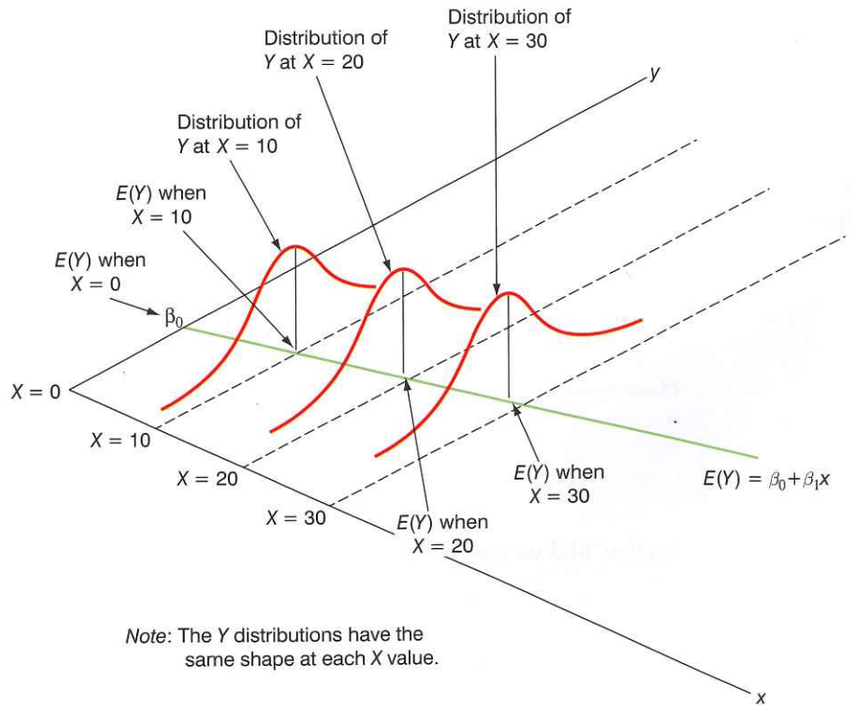
2 The variance of ε, denoted by σ^2, is the same for all values of X
Implication: The variance of Y about the regression line equals σ^2 and is the same for all values of X.

3 The values of ε are independent.
Implication: The value of ε for a particular value of X is not related to the value of ε for any other value of X; thus, the value of Y for a particular value of X is not related to the value of Y for any other value of X.

4 The error term ε is a normally distributed random variable.
Implication: Because Y is a linear function of ε, Y is also a normally distributed random variable.

Figure 14.6 illustrates the model assumptions and their implications; note that in this graphical interpretation, the value of $E(Y)$ changes according to the specific value of X considered. However, regardless of the X value, the probability distribution of ε and hence the probability distributions of Y are normally distributed, each with the same

Figure 14.6 Assumptions for the regression model



variance. The specific value of the error ε at any particular point depends on whether the actual value of Y is greater than or less than $E(Y)$.

At this point, we must keep in mind that we are also making an assumption or hypothesis about the form of the relationship between X and Y. That is, we assume that a straight line represented by $\beta_0 + \beta_1 x$ is the basis for the relationship between the variables. We must not lose sight of the fact that some other model, for instance $Y = \beta_0 + \beta_1 x^2 + \varepsilon$ may turn out to be a better model for the underlying relationship.

14.5 Testing for significance

In a simple linear regression equation, the mean or expected value of Y is a linear function of x: $E(Y) = \beta_0 + \beta_1 x$. If the value of β_1 is zero, $E(Y) = \beta_0 + (0)\, x = \beta_0$. In this case, the mean value of Y does not depend on the value of X and hence we would conclude that X and Y are not linearly related. Alternatively, if the value of β_1 is not equal to zero, we would conclude that the two variables are related. Thus, to test for a significant regression relationship, we must conduct a hypothesis test to determine whether the value of β_1 is zero. Two tests are commonly used. Both require an estimate of σ^2, the variance of ε in the regression model.

Estimate of σ^2

From the regression model and its assumptions we can conclude that σ^2, the variance of ε, also represents the variance of the Y values about the regression line. Recall that the deviations of the Y values about the estimated regression line are called residuals. Thus, SSE, the sum of squared residuals, is a measure of the variability of the actual observations about the estimated regression line. The **mean square error** (MSE) provides the estimate of σ^2; it is SSE divided by its degrees of freedom.

With $\hat{y}_i = b_0 + b_1 x_i$, SSE can be written as

$$\text{SSE} = \Sigma(y_1 - \hat{y}_i)^2 = \Sigma(y_1 - b_0 - b_1 x_i)^2$$

Every sum of squares is associated with a number called its degrees of freedom. Statisticians have shown that SSE has $n - 2$ degrees of freedom because two parameters (β_0 and β_1) must be estimated to compute SSE. Thus, the mean square is computed by dividing SSE by $n - 2$. MSE provides an unbiased estimator of σ^2. Because the value of MSE provides an estimate of σ^2, the notation s^2 is also used.

Mean square error (estimate of σ^2)

$$s^2 = \text{MSE} = \frac{\text{SSE}}{n - 2} \qquad (14.15)$$

In Section 14.3 we showed that for the Armand's Pizza Parlours example, SSE = 1530; hence,

$$s^2 = \text{MSE} = \frac{1530}{8} = 191.25$$

provides an unbiased estimate of σ^2.

To estimate σ we take the square root of s^2. The resulting value, s, is referred to as the **standard error of the estimate**.

Standard error of estimate

$$s = \sqrt{\text{MSE}} = \sqrt{\frac{\text{SSE}}{n - 2}} \qquad (14.16)$$

For the Armand's Pizza Parlours example, $s = \sqrt{\text{MSE}} = \sqrt{191.25} = 13.829$. In the following discussion, we use the standard error of the estimate in the tests for a significant relationship between X and Y.

t test

The simple linear regression model is $Y = \beta_0 + \beta_1 x + \varepsilon$. If X and Y are linearly related, we must have $\beta_1 \neq 0$. The purpose of the t test is to see whether we can conclude that $\beta_1 \neq 0$.

We will use the sample data to test the following hypotheses about the parameter β_1.

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

If H_0 is rejected, we will conclude that $\beta_1 \neq 0$ and that a statistically significant relationship exists between the two variables. However, if H_0 cannot be rejected, we will have insufficient evidence to conclude that a significant relationship exists. The properties of the sampling distribution of b_1, the least squares estimator of β_1, provide the basis for the hypothesis test.

First, let us consider what would happen if we used a different random sample for the same regression study. For example, suppose that Armand's Pizza Parlours used the sales records of a different sample of ten restaurants. A regression analysis of this new sample might result in an estimated regression equation similar to our previous estimated regression equation $\hat{y} = 60 + 5x$. However, it is doubtful that we would obtain exactly the same equation (with an intercept of exactly 60 and a slope of exactly 5). Indeed, b_0 and b_1, the least squares estimators, are sample statistics with their own sampling distributions. The properties of the sampling distribution of b_1 follow.

Sampling distribution of b_1

Expected value $E(b_1) = \beta_1$

Standard deviation

$$\sigma_{b_1} = \frac{\sigma}{\sqrt{\Sigma(x_i - \bar{x})^2}} \qquad (14.17)$$

Distribution form

Normal

Note that the expected value of b_1 is equal to β_1, so b_1 is an unbiased estimator of β_1. As we do not know the value of σ, so we estimate σ_{b_1} by s_{b_1} where s_{b_1} is derived by substituting s for σ in equation (14.17):

Estimated standard deviation of b_1

$$s_{b_1} = \frac{s}{\sqrt{\Sigma(x_i - \bar{x})^2}} \qquad (14.18)$$

For Armand's Pizza Parlours, $s = 13.829$. Hence, using $\Sigma(x_i - \bar{x})^2 = 568$ as shown in Table 14.2, we have

$$s_{b_1} = \frac{13.829}{\sqrt{568}} = 0.5803$$

as the estimated standard deviation of b_1.

The t test for a significant relationship is based on the fact that the test statistic

$$\frac{b_1 - \beta_1}{s_{b_1}}$$

follows a t distribution with $n - 2$ degrees of freedom. If the null hypothesis is true, then $\beta_1 = 0$ and $t = b_1/s_{b_1}$.

Let us conduct this test of significance for Armand's Pizza Parlours at the $\alpha = 0.01$ level of significance. The test statistic is

$$t = \frac{b_1}{s_{b_1}} = \frac{5}{0.5803} = 8.62$$

The t distribution table shows that with $n - 2 = 10 - 2 = 8$ degrees of freedom, $t = 3.355$ provides an area of 0.005 in the upper tail. Thus, the area in the upper tail of the t distribution corresponding to the test statistic $t = 8.62$ must be less than 0.005. Because this test is a two-tailed test, we double this value to conclude that the p-value associated with $t = 8.62$ must be less than $2(0.005) = 0.01$. MINITAB, PASW or EXCEL show the p-value = 0.000. Because the p-value is less than $\alpha = 0.01$, we reject H_0 and conclude that β_1 is not equal to zero. This evidence is sufficient to conclude that a significant relationship exists between student population and quarterly sales. A summary of the t test for significance in simple linear regression follows.

t test for significance in simple linear regression

$H_0: \beta_1 = 0$
$H_1: \beta_1 \neq 0$

Test statistic

$$t = \frac{b_1}{s_{b_1}} \qquad (14.19)$$

Rejection rule

p-value approach: Reject H_0 if p-value $\leq \alpha$
Critical value approach: Reject H_0 if $t \leq -t_{\alpha/2}$ or if $t \geq t_{\alpha/2}$

where $t_{\alpha/2}$ is based on a t distribution with $n - 2$ degrees of freedom.

Confidence interval for β_1

The form of a confidence interval for β_1 is as follows:

$$b_1 \pm t_{\alpha/2} s_{b_1}$$

The point estimator is b_1 and the margin of error is $t_{\alpha/2} s_{b_1}$. The confidence coefficient associated with this interval is $1 - \alpha$, and $t_{\alpha/2}$ is the t value providing an area of $\alpha/2$ in the upper tail of a t distribution with $n - 2$ degrees of freedom. For example, suppose that we wanted to develop a 99 per cent confidence interval estimate of β_1 for Armand's Pizza Parlours. From Table 2 of Appendix B we find that the t value corresponding to $\alpha = 0.01$ and $n - 2 = 10 - 2 = 8$ degrees of freedom is $t_{0.005} = 3.355$. Thus, the 99 per cent confidence interval estimate of β_1 is

$$b_1 \pm t_{\alpha/2} s_{b_1} = 5 \pm 3.355(0.5803) = 5 \pm 1.95$$

or 3.05 to 6.95.

In using the t test for significance, the hypotheses tested were

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

At the $\alpha = 0.01$ level of significance, we can use the 99 per cent confidence interval as an alternative for drawing the hypothesis testing conclusion for the Armand's data. Because 0, the hypothesized value of β_1, is not included in the confidence interval (3.05 to 6.95), we can reject H_0 and conclude that a significant statistical relationship exists between the size of the student population and quarterly sales. In general, a confidence interval can be used to test any two-sided hypothesis about β_1. If the hypothesized value of β_1 is contained in the confidence interval, do not reject H_0. Otherwise, reject H_0.

F test

An F test, based on the F probability distribution, can also be used to test for significance in regression. With only one independent variable, the F test will provide the same conclusion as the t test; that is, if the t test indicates $\beta_1 \neq 0$ and hence a significant relationship, the F test will also indicate a significant relationship*. But with more than one independent variable, only the F test can be used to test for an overall significant relationship.

The logic behind the use of the F test for determining whether the regression relationship is statistically significant is based on the development of two independent estimates of σ^2. We explained how MSE provides an estimate of σ^2. If the null hypothesis $H_0: \beta_1 = 0$ is true, the sum of squares due to regression, SSR, divided by its degrees of freedom provides another independent estimate of σ^2. This estimate is called the *mean square due to regression*, or simply the *mean square regression*, and is denoted MSR. In general,

$$\text{MSR} = \frac{\text{SSR}}{\text{Regression degrees of freedom}}$$

*In fact $F = t^2$ for a simple regression model.

For the models we consider in this text, the regression degrees of freedom is always equal to the number of independent variables in the model:

Mean square regression

$$\text{MSR} = \frac{\text{SSR}}{\text{Number of independent variables}} \qquad (14.20)$$

Because we consider only regression models with one independent variable in this chapter, we have MSR = SSR/1 = SSR. Hence, for Armand's Pizza Parlours, MSR = SSR = 14 200.

If the null hypothesis (H_0: $\beta_1 = 0$) is true, MSR and MSE are two independent estimates of σ^2 and the sampling distribution of MSR/MSE follows an F distribution with numerator degrees of freedom equal to one and denominator degrees of freedom equal to $n - 2$. Therefore, when $\beta_1 = 0$, the value of MSR/MSE should be close to one. However, if the null hypothesis is false ($\beta_1 \neq 0$), MSR will overestimate σ^2 and the value of MSR/MSE will be inflated; thus, large values of MSR/MSE lead to the rejection of H_0 and the conclusion that the relationship between X and Y is statistically significant.

Let us conduct the F test for the Armand's Pizza Parlours example. The test statistic is

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{14\,200}{191.25} = 74.25$$

The F distribution table (Table 4 of Appendix B) shows that with one degree of freedom in the denominator and $n - 2 = 10 - 2 = 8$ degrees of freedom in the denominator, $F = 11.26$ provides an area of 0.01 in the upper tail. Thus, the area in the upper tail of the F distribution corresponding to the test statistic $F = 74.25$ must be less than 0.01. Thus, we conclude that the p-value must be less than 0.01. MINITAB, PASW or EXCEL show the p-value = 0.000. Because the p-value is less than $\alpha = 0.01$, we reject H_0 and conclude that a significant relationship exists between the size of the student population and quarterly sales. A summary of the F test for significance in simple linear regression follows.

F test for significance in simple linear regression

H_0: $\beta_1 = 0$
H_1: $\beta_1 \neq 0$

Test statistic

$$F = \frac{\text{MSR}}{\text{MSE}} \qquad (14.21)$$

Rejection rule

p-value approach: Reject H_0 if p-value $\leq \alpha$
Critical value approach: Reject H_0 if $F \geq F_\alpha$

where F_α is based on an F distribution with 1 degree of freedom in the numerator and $n - 2$ degrees of freedom in the denominator.

Table 14.5 General form of the ANOVA table for simple linear regression

Source of variation	Degrees of freedom	Sum of squares	Mean square	F
Regression	1	SSR	$\text{MSR} = \frac{\text{SSR}}{1}$	$\frac{\text{MSR}}{\text{MSE}}$
Error	$n - 2$	SSE	$\text{MSE} = \frac{\text{SSE}}{n-2}$	
Total	$n - 1$	SST		

In Chapter 13 we covered analysis of variance (ANOVA) and showed how an **ANOVA table** could be used to provide a convenient summary of the computational aspects of analysis of variance. A similar ANOVA table can be used to summarize the results of the F test for significance in regression. Table 14.5 is the general form of the ANOVA table for simple linear regression. Table 14.6 is the ANOVA table with the F test computations performed for Armand's Pizza Parlours. Regression, Error and Total are the labels for the three sources of variation, with SSR, SSE and SST appearing as the corresponding sum of squares in column 3. The degrees of freedom, 1 for SSR, $n - 2$ for SSE and $n - 1$ for SST, are shown in column 2. Column 4 contains the values of MSR and MSE and column 5 contains the value of F = MSR/MSE. Almost all computer printouts of regression analysis include an ANOVA table summary and the F test for significance.

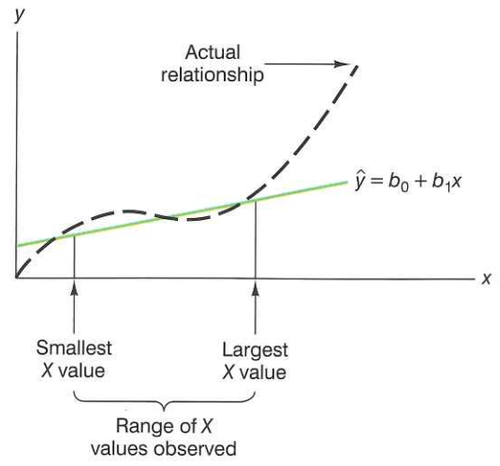
Some cautions about the interpretation of significance tests

Rejecting the null hypothesis H_0: $\beta_1 = 0$ and concluding that the relationship between X and Y is significant does not enable us to conclude that a cause-and-effect relationship is present between X and Y. Concluding a cause-and-effect relationship is warranted only if the analyst can provide some type of theoretical justification that the relationship is in fact causal. In the Armand's Pizza Parlours example, we can conclude that there is a significant relationship between the size of the student population X and quarterly sales Y; moreover, the estimated regression equation $\hat{y} = 60 + 5x$ provides the least squares estimate of the relationship. We cannot, however, conclude that changes in student population X cause changes in quarterly sales Y just because we identified a statistically significant relationship. The appropriateness of such a cause-and-effect conclusion is left to supporting theoretical justification and to good judgment on the part of the analyst.

Table 14.6 ANOVA table for the Armand's Pizza Parlours problem

Source of variation	Degrees of freedom	Sum of squares	Mean square	F
Regression	1	14 200	$\frac{14\,200}{1} = 14\,200$	$\frac{14\,200}{191.25} = 74.25$
Error	8	1 530	$\frac{1530}{8} = 191.25$	
Total	9	15 730		

Figure 14.7 Example of a linear approximation of a nonlinear relationship



Armand's managers felt that increases in the student population were a likely cause of increased quarterly sales. Thus, the result of the significance test enabled them to conclude that a cause-and-effect relationship was present.

In addition, just because we are able to reject H_0: $\beta_1 = 0$ and demonstrate statistical significance does not enable us to conclude that the relationship between X and Y is linear. We can state only that X and Y are related and that a linear relationship explains a significant portion of the variability in Y over the range of values for X observed in the sample. Figure 14.7 illustrates this situation. The test for significance calls for the rejection of the null hypothesis H_0: $\beta_1 = 0$ and leads to the conclusion that X and Y are significantly related, but the figure shows that the actual relationship between X and Y is not linear. Although the linear approximation provided by $\hat{y} = b_0 + b_1x$ is good over the range of X values observed in the sample, it becomes poor for X values outside that range.

Given a significant relationship, we should feel confident in using the estimated regression equation for predictions corresponding to X values within the range of the X values observed in the sample. For Armand's Pizza Parlours, this range corresponds to values of X between 2 and 26. Unless other reasons indicate that the model is valid beyond this range, predictions outside the range of the independent variable should be made with caution. For Armand's Pizza Parlours, because the regression relationship has been found significant at the 0.01 level, we should feel confident using it to predict sales for restaurants where the associated student population is between 2000 and 26 000.

Exercises

Methods

13 The data from exercise 1 follow.

x_i	1	2	3	4	5
y_i	3	7	5	11	14

a. Compute the mean square error using equation (14.15).
b. Compute the standard error of the estimate using equation (14.16).
c. Compute the estimated standard deviation of b_1 using equation (14.18). d. Use the t test to test the following hypotheses ($\alpha = 0.05$):

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

e. Use the F test to test the hypotheses in part (d) at a 0.05 level of significance. Present the results in the analysis of variance table format.

14 The data from exercise 2 follow.

x_i	2	3	5	1	8
y_i	25	25	20	30	16

a. Compute the mean square error using equation (14.15).
b. Compute the standard error of the estimate using equation (14.16).
c. Compute the estimated standard deviation of b_1 using equation (14.18).
d. Use the t test to test the following hypotheses ($\alpha = 0.05$):

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

e. Use the F test to test the hypotheses in part (d) at a 0.05 level of significance. Present the results in the analysis of variance table format.

15 The data from exercise 3 follow.

x_i	2	4	5	7	8
y_i	2	3	2	6	4

a. What is the value of the standard error of the estimate?
b. Test for a significant relationship by using the t test. Use $\alpha = 0.05$.
c. Use the F test to test for a significant relationship. Use $\alpha = 0.05$. What is your conclusion?

Applications

16 Refer to exercise 11, where data on production volume and cost were used to develop an estimated regression equation relating production volume and cost for a particular manufacturing operation. Use $\alpha = 0.05$ to test whether the production volume is significantly related to the total cost. Show the ANOVA table. What is your conclusion?

17 Refer to exercise 12 where the data were used to determine whether the price of a printer is related to the speed for plain text printing (*PC World*, April 2009). Does the evidence indicate a significant relationship between printing speed and price? Conduct the appropriate statistical test and state your conclusion. Use $\alpha = 0.05$.

14.6 Using the estimated regression equation for estimation and prediction

When using the simple linear regression model we are making an assumption about the relationship between X and Y. We then use the least squares method to obtain the estimated simple linear regression equation. If a significant relationship exists between X and

Y, and the coefficient of determination shows that the fit is good, the estimated regression equation should be useful for estimation and prediction.

Point estimation

In the Armand's Pizza Parlours example, the estimated regression equation $\hat{y} = 60 + 5x$ provides an estimate of the relationship between the size of the student population X and quarterly sales Y. We can use the estimated regression equation to develop a point estimate of either the mean value of Y or an individual value of Y corresponding to a given value of X. For instance, suppose Armand's managers want a point estimate of the mean quarterly sales for all restaurants located near college campuses with 10 000 students. Using the estimated regression equation $\hat{y} = 60 + 5x$, we see that for $X = 10$ (or 10 000 students), $\hat{y} = 60 + 5(10) = 110$. Thus, a point estimate of the mean quarterly sales for all restaurants located near campuses with 10 000 students is €110 000.

Now suppose Armand's managers want to predict sales for an individual restaurant located near Cabot College, a school with 10 000 students. Then, as the point estimate for an individual value of Y is the same as the point estimate for the mean value of Y we would predict quarterly sales of $\hat{y} = 60 + 5(10) = 110$ or €110 000 for this one restaurant.

Interval estimation

Point estimates do not provide any information about the precision associated with an estimate. For that we must develop interval estimates much like those in Chapters 8, 10 and 11. The first type of interval estimate, a **confidence interval**, is an interval estimate of the *mean value of Y* for a given value of X. The second type of interval estimate, a **prediction interval**, is used whenever we want an interval estimate of an *individual value* of Y for a given value of X. The point estimate of the mean value of Y is the same as the point estimate of an individual value of Y. But, the interval estimates we obtain for the two cases are different. The margin of error is larger for a prediction interval.

Confidence interval for the mean value of Y

The estimated regression equation provides a point estimate of the mean value of Y for a given value of X. In developing the confidence interval, we will use the following notation.

x_p = the particular or given value of the independent variable X
Y_p = the dependent variable Y corresponding to the given x_p
$E(Y_p)$ = the mean or expected value of the dependent variable Y_p corresponding to the given x_p
$\hat{y}_p = b_0 + b_1 x_p$ = the point estimate of $E(Y_p)$ when $X = x_p$

Using this notation to estimate the mean sales for all Armand's restaurants located near a campus with 10 000 students, we have $x_p = 10$, and $E(Y_p)$ denotes the unknown mean value of sales for all restaurants where $x_p = 10$. The point estimate of $E(Y_p)$ is given by $\hat{y}_p = 60 + 5(10) = 110$.

In general, we cannot expect $\hat{y}_p$ to equal $E(Y_p)$ exactly. If we want to make an inference about how close $\hat{y}_p$ is to the true mean value $E(Y_p)$, we will have to estimate

the variance of $\hat{y}_p$. The formula for estimating the variance of $\hat{y}_p$ given x_p, denoted by $s^2_{\hat{y}_p}$ is

$$s^2_{\hat{y}_p} = s^2\left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}\right]$$

The general expression for a confidence interval follows.

Confidence interval for $E(Y_p)$

$$\hat{y}_p \pm t_{\alpha/2} s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}} \qquad (14.22)$$

where the confidence coefficient is $1 - \alpha$ and $t_{\alpha/2}$ is based on a t distribution with $n - 2$ degrees of freedom.

Using expression (14.22) to develop a 95 per cent confidence interval of the mean quarterly sales for all Armand's restaurants located near campuses with 10 000 students, we need the value of t for $\alpha/2 = 0.025$ and $n - 2 = 10 - 2 = 8$ degrees of freedom. Using Table 2 of Appendix B, we have $t_{0.025} = 2.306$. Thus, with $\hat{y}_p = 110$, the 95 per cent confidence interval estimate is

$$\hat{y}_p \pm t_{\alpha/2} s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}}$$

$$110 \pm 2.306 \times 13.829\sqrt{\frac{1}{10} + \frac{(10 - 14)^2}{568}}$$

$$= 110 \pm 11.415$$

In euros, the 95 per cent confidence interval for the mean quarterly sales of all restaurants near campuses with 10 000 students is €110 000 ± €11 415. Therefore, the 95 per cent confidence interval for the mean quarterly sales when the student population is 10 000 is €98 585 to €121 415.

Note that the estimated standard deviation of $\hat{y}_p$ is smallest when $x_p = \bar{x}$ so that the quantity $x_p - \bar{x} = 0$. In this case, the estimated standard deviation of $\hat{y}_p$ becomes

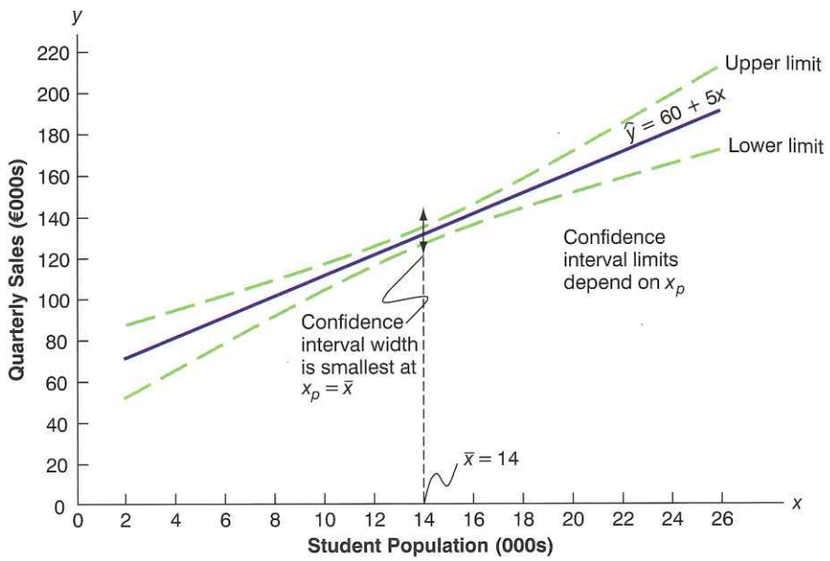
$$s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}} = s\sqrt{\frac{1}{n}}$$

This result implies that the best or most precise estimate of the mean value of Y occurs when $x_p = \bar{x}$. But, the further x_p is from $\bar{x}$ the larger $x_p - \bar{x}$ becomes and thus the wider confidence intervals will be for the mean value of Y. This pattern is shown graphically in Figure 14.8.

Prediction interval for an individual value of Y

Suppose that instead of estimating the mean value of sales for all Armand's restaurants located near campuses with 10 000 students, we want to estimate the sales for an individual restaurant located near Cabot College, a school with 10 000 students. As noted previously,

Figure 14.8 Confidence intervals for the mean sales Y at given values of student population x



the point estimate of y_p, the value of Y corresponding to the given x_p, is provided by the estimated regression equation $\hat{y}_p = b_0 + b_1 x_p$. For the restaurant at Cabot College, we have $x_p = 10$ and a corresponding predicted quarterly sales of $\hat{y}_p = 60 + 5(10) = 110$, or €110 000.

Note that this value is the same as the point estimate of the mean sales for all restaurants located near campuses with 10 000 students.

To develop a prediction interval, we must first determine the variance associated with using $\hat{y}_p$ as an estimate of an individual value of Y when $X = x_p$. This variance is made up of the sum of the following two components.

1. The variance of individual Y values about the mean $E(Y_p)$, an estimate of which is given by s^2.
2. The variance associated with using $\hat{y}_p$ to estimate $E(Y_p)$, an estimate of which is given by

$$s_{\hat{y}_p}^2 = s^2\left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}\right]$$

Thus the formula for estimating the variance of an individual value of Y_p, is

$$s^2 + s_{\hat{y}_p}^2 = s^2 + s^2\left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}\right] = s^2\left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}\right]$$

The general expression for a prediction interval follows.

Prediction interval for y_p

$$\hat{y}_p \pm t_{\alpha/2}\, s\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}} \qquad (14.23)$$

where the confidence coefficient is $1 - \alpha$ and $t_{\alpha/2}$ is based on a t distribution with $n - 2$ degrees of freedom.

Thus the 95 per cent prediction interval of sales for one specific restaurant located near a campus with 10 000 students is

$$\hat{y}_p \pm t_{\alpha/2}\, s\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}}$$

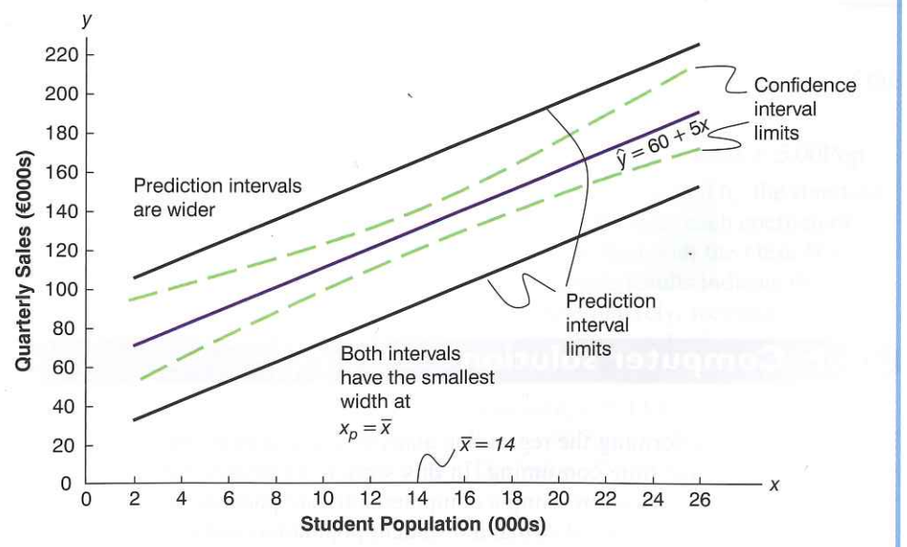
$$= 110 \pm 2.306 \times 13.829\sqrt{1 + \frac{1}{10} + \frac{(10 - 14)^2}{568}}$$

$$= 110 \pm 33.875$$

In euros, this prediction interval is €110 000 ± €33 875 or €76 125 to €143 875. Note that the prediction interval for an individual restaurant located near a campus with 10 000 students is wider than the confidence interval for the mean sales of all restaurants located near campuses with 10 000 students. The difference reflects the fact that we are able to estimate the mean value of Y more precisely than we can an individual value of Y.

Both confidence interval estimates and prediction interval estimates are most precise when the value of the independent variable is $x_p = \bar{x}$. The general shapes of confidence intervals and the wider prediction intervals are shown together in Figure 14.9.

Figure 14.9 Confidence and prediction intervals for sales Y at given values of student population X



Exercises

Methods

18 The data from exercise 1 follow.

x_i	1	2	3	4	5
y_i	3	7	5	11	14

a. Use expression (14.22) to develop a 95 per cent confidence interval for the expected value of Y when $X = 4$.
b. Use expression (14.23) to develop a 95 per cent prediction interval for Y when $X = 4$.

19 The data from exercise 2 follow.

x_i	2	3	5	1	8
y_i	25	25	20	30	16

a. Estimate the standard deviation of $\hat{y}_p$ when $X = 3$.
b. Develop a 95 per cent confidence interval for the expected value of Y when $X = 3$.
c. Estimate the standard deviation of an individual value of Y when $X = 3$.
d. Develop a 95 per cent prediction interval for Y when $X = 3$.

20 The data from exercise 3 follow.

x_i	2	4	5	7	8
y_i	2	3	2	6	4

Develop the 95 per cent confidence and prediction intervals when $X = 3$. Explain why these two intervals are different.

Applications

21 Refer to Exercise 11, where data on the production volume X and total cost Y for a particular manufacturing operation were used to develop the estimated regression equation $\hat{y} = 1246.67 + 7.6x$.

a. The company's production schedule shows that 500 units must be produced next month. What is the point estimate of the total cost for next month?
b. Develop a 99 per cent prediction interval for the total cost for next month.
c. If an accounting cost report at the end of next month shows that the actual production cost during the month was €6000, should managers be concerned about incurring such a high total cost for the month? Discuss.

14.7 Computer solution

Performing the regression analysis computations without the help of a computer can be quite time consuming. In this section we discuss how the computational burden can be minimized by using a computer software package such as MINITAB.

We entered Armand's student population and sales data into a MINITAB worksheet. The independent variable was named Pop and the dependent variable was named Sales to assist with interpretation of the computer output. Using MINITAB, we obtained the

Figure 14.10 MINITAB output for the Armand's Pizza Parlours problem

```
Regression Analysis: Sales versus Pop

The regression equation is
Sales = 60.0 + 5.00 Pop


Predictor      Coef  SE Coef      T      P
Constant     60.000    9.226   6.50  0.000
Pop          5.0000   0.5803   8.62  0.000


S = 13.8293   R-Sq = 90.3%   R-Sq(adj) = 89.1%


Analysis of Variance

Source           DF      SS      MS      F      P
Regression        1   14200   14200  74.25  0.000
Residual Error    8    1530     191
Total             9   15730


Predicted Values for New Observations

New
Obs     Fit  SE Fit        95% CI              95% PI
  1  110.00    4.95  (98.58, 121.42)  (76.13, 143.87)


Values of Predictors for New Observations

New
Obs   Pop
  1  10.0
```

printout for Armand's Pizza Parlours shown in Figure 14.10.* The interpretation of this printout follows.

1 MINITAB prints the estimated regression equation as Sales = 60.0 + 5.00Pop.
2 A table is printed that shows the values of the coefficients b_0 and b_1, the standard deviation of each coefficient, the t value obtained by dividing each coefficient value by its standard deviation, and the p-value associated with the t test. Because the p-value is zero (to three decimal places), the sample results indicate that the null hypothesis (H_0: $\beta_1 = 0$) should be rejected. Alternatively, we could compare 8.62 (located in the t-ratio column) to the appropriate critical value. This procedure for the t test was described in Section 14.5.
3 MINITAB prints the standard error of the estimate, $s = 13.83$, as well as information about the goodness of fit. Note that 'R-sq = 90.3 per cent' is the coefficient of determination expressed as a percentage. The value 'R-Sq (adj) = 89.1 per cent' is discussed in Chapter 15.

*The MINITAB steps necessary to generate the output are given in the software section at the end of the chapter.

4 The ANOVA table is printed below the heading Analysis of Variance. MINITAB uses the label Residual Error for the error source of variation. Note that DF is an abbreviation for degrees of freedom and that MSR is given as 14 200 and MSE as 191.

The ratio of these two values provides the F value of 74.25 and the corresponding p-value of 0.000. Because the p-value is zero (to three decimal places), the relationship between Sales and Pop is judged statistically significant.

5 The 95 per cent confidence interval estimate of the expected sales and the 95 per cent prediction interval estimate of sales for an individual restaurant located near a campus with 10 000 students are printed below the ANOVA table. The confidence interval is (98.58, 121.42) and the prediction interval is (76.12, 143.88) as we showed in Section 14.6.

Exercises

Applications

22 The commercial division of a real estate firm is conducting a regression analysis of the relationship between X, annual gross rents (in thousands of euros), and Y, selling price (in thousands of euros) for apartment buildings. Data were collected on several properties recently sold and the following computer selective output was obtained.

```
The regression equation is
Y = 20.0 + 7.21 X

Predictor       Coef    SE Coef       T
Constant      20.000      3.2213    6.21
X              7.210      1.3626    5.29

Analysis of Variance

SOURCE            DF         SS
Regression         1    41587.3
Residual Error     7
Total              8    51984.1
```

a. How many apartment buildings were in the sample?
b. Write the estimated regression equation.
c. What is the value of s_{b_1}?
d. Use the F statistic to test the significance of the relationship at a 0.05 level of significance.
e. Estimate the selling price of an apartment building with gross annual rents of €50 000.

23 Following is a portion of the computer output for a regression analysis relating Y = maintenance expense (euros per month) to X = usage (hours per week) of a particular brand of computer terminal.

```
The regression equation is
Y = 6.1092 + .8951 X

Predictor       Coef     SE Coef
Constant      6.1092      0.9361
X             0.8951      0.1490

Analysis of Variance

SOURCE            DF         SS         MS
Regression         1    1575.76    1575.76
Residual Error     8     349.14      43.64
Total              9    1924.90
```


a. Write the estimated regression equation.
b. Use a t test to determine whether monthly maintenance expense is related to usage at the 0.05 level of significance.
c. Use the estimated regression equation to predict mean monthly maintenance expense for any terminal that is used 25 hours per week.

24 A regression model relating X, number of salespersons at a branch office, to Y, annual sales at the office (in thousands of euros) provided the following computer output from a regression analysis of the data.

```
The regression equation is
Y = 80.0 + 50.00 X

Predictor       Coef    SE Coef       T
Constant        80.0     11.333    7.06
X               50.0      5.482    9.12

Analysis of Variance

SOURCE            DF         SS         MS
Regression         1     6828.6     6828.6
Residual Error    28     2298.8      82.1
Total             29     9127.4
```

a. Write the estimated regression equation.
b. How many branch offices were involved in the study?
c. Compute the F statistic and test the significance of the relationship at a 0.05 level of significance.
d. Predict the annual sales at the Marseilles branch office. This branch employs 12 salespersons.

14.8 Residual analysis: validating model assumptions

As we noted previously, the *residual* for observation i is the difference between the observed value of the dependent variable (y_i) and the estimated value of the dependent variable ($\hat{y}_i$).

Residual for observation i

$$y_i - \hat{y}_i \qquad (14.24)$$

where
y_i is the observed value of the dependent variable
$\hat{y}_i$ is the estimated value of the dependent variable

In other words, the ith residual is the error resulting from using the estimated regression equation to predict the value of the dependent variable. The residuals for the Armand's Pizza Parlours example are computed in Table 14.7. The observed values of the dependent variable are in the second column and the estimated values of the dependent variable, obtained using the estimated regression equation $\hat{y} = 60 + 5x$, are in the third column. An analysis of the corresponding residuals in the fourth column

Table 14.7 Residuals for Armand's Pizza Parlours

Student population x_i	Sales y_i	Estimated sales $\hat{y}_i = 60 - 5x_i$	Residuals $y_i - \hat{y}_i$
2	58	70	−12
6	105	90	15
8	88	100	−12
8	118	100	18
12	117	120	−3
16	137	140	−3
20	157	160	−3
20	169	160	9
22	149	170	−21
26	202	190	12

will help determine whether the assumptions made about the regression model are appropriate.

Recall that for the Armand's Pizza Parlours example it was assumed the simple linear regression model took the form:

$$Y = \beta_0 + \beta_1 x + \varepsilon \qquad (14.25)$$

In other words we assumed quarterly sales (Y) to be a linear function of the size of the student population (X) plus an error term ε. In Section 14.4 we made the following assumptions about the error term ε.

1. $E(\varepsilon) = 0$.
2. The variance of ε, denoted by σ^2, is the same for all values of X.
3. The values of ε are independent.
4. The error term ε has a normal distribution.

These assumptions provide the theoretical basis for the t test and the F test used to determine whether the relationship between X and Y is significant, and for the confidence and prediction interval estimates presented in Section 14.6. If the assumptions about the error term ε appear questionable, the hypothesis tests about the significance of the regression relationship and the interval estimation results may not be valid.

The residuals provide the best information about ε; hence an analysis of the residuals is an important step in determining whether the assumptions for ε are appropriate. Much of **residual analysis** is based on an examination of graphical plots. In this section, we discuss the following residual plots.

1. A plot of the residuals against values of the independent variable X.
2. A plot of residuals against the predicted values $\hat{y}$ of the dependent variable.
3. A standardized residual plot.
4. A normal probability plot.

Residual plot against X

A **residual plot** against the independent variable X is a graph in which the values of the independent variable are represented by the horizontal axis and the corresponding residual values are represented by the vertical axis. A point is plotted for each residual. The first coordinate for each point is given by the value of x_i and the second coordinate is given by the corresponding value of the residual $y_i - \hat{y}_i$. For a residual plot against X with the Armand's Pizza Parlours data from Table 14.7, the coordinates of the first point are (2,− 12), corresponding to $x_1 = 2$ and $y_1 - \hat{y}_1 = -12$; the coordinates of the second point are (6, 15), corresponding to $x_2 = 6$ and $y_2 - \hat{y}_2 = 15$ and so on. Figure 14.11 shows the resulting residual plot.

Before interpreting the results for this residual plot, let us consider some general patterns that might be observed in any residual plot. Three examples appear in Figure 14.12.

If the assumption that the variance of ε is the same for all values of X and the assumed regression model is an adequate representation of the relationship between the variables, the residual plot should give an overall impression of a horizontal band of points such as the one in Panel A of Figure 14.12. However, if the variance of ε is not the same for all values of X − for example, if variability about the regression line is greater for larger values of X – a pattern such as the one in Panel B of Figure 14.12 could be observed. In this case, the assumption of a constant variance of ε is violated. Another possible residual plot is shown in Panel C. In this case, we would conclude that the assumed regression model is not an adequate representation of the relationship between the variables. A curvilinear regression model or multiple regression model should be considered.

Now let us return to the residual plot for Armand's Pizza Parlours shown in Figure 14.11. The residuals appear to approximate the horizontal pattern in Panel A of Figure 14.12. Hence, we conclude that the residual plot does not provide evidence that the assumptions made for Armand's regression model should be challenged. At

Figure 14.11 Plot of the residuals against the independent variable for Armand's Pizza Parlours

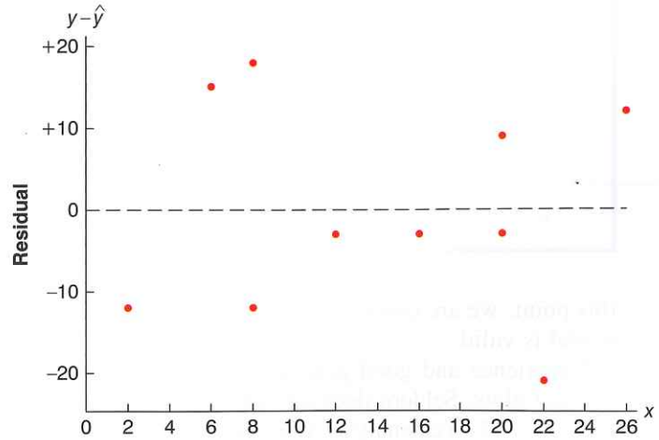
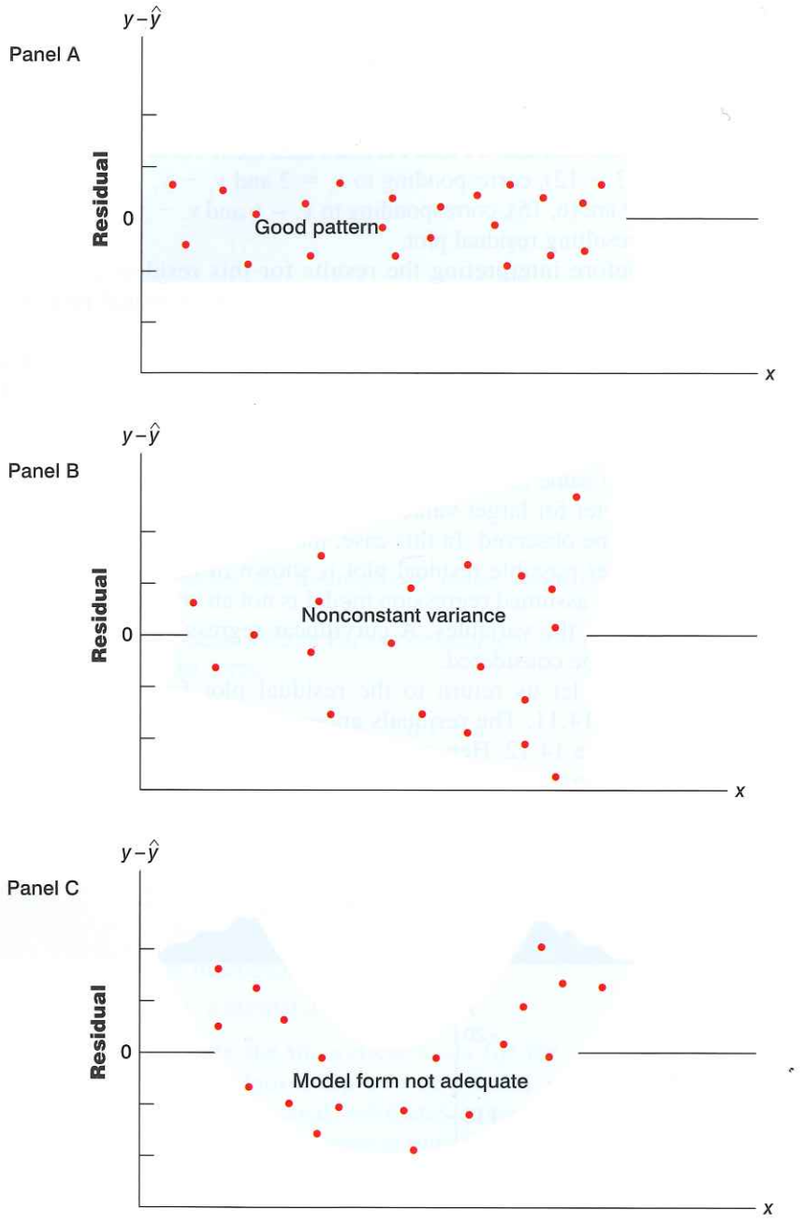


Figure 14.12 Residual plots from three regression studies



this point, we are confident in the conclusion that Armand's simple linear regression model is valid.

Experience and good judgment are always factors in the effective interpretation of residual plots. Seldom does a residual plot conform precisely to one of the patterns in Figure 14.12. Yet analysts who frequently conduct regression studies and frequently review residual plots become adept at understanding the differences between patterns that are reasonable and patterns that indicate the assumptions of the model should be

questioned. A residual plot provides one technique to assess the validity of the assumptions for a regression model.

Residual plot against $\hat{y}$

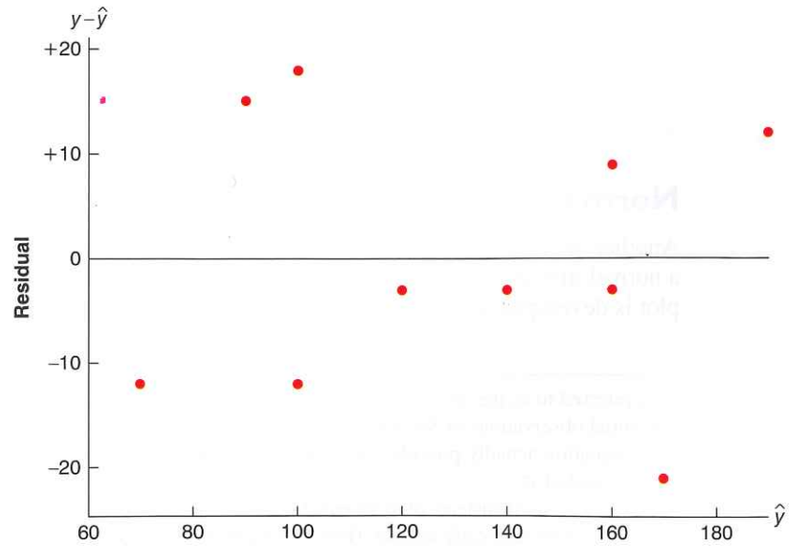
Another residual plot represents the predicted value of the dependent variable $\hat{y}$ on the horizontal axis and the residual values on the vertical axis. A point is plotted for each residual. The first coordinate for each point is given by $\hat{y}_i$ and the second coordinate is given by the corresponding value of the ith residual $y_i - \hat{y}_i$. With the Armand's data from Table 14.7, the coordinates of the first point are (70, − 12), corresponding to $\hat{y}_1 = 70$ and $y_1 - \hat{y}_1 = -12$; the coordinates of the second point are (90, 15), and so on. Figure 14.13 provides the residual plot. Note that the pattern of this residual plot is the same as the pattern of the residual plot against the independent variable X. It is not a pattern that would lead us to question the model assumptions. For simple linear regression, both the residual plot against X and the residual plot against $\hat{y}$ provide the same pattern. For multiple regression analysis, the residual plot against $\hat{y}$ is more widely used because of the presence of more than one independent variable.

Standardized residuals

Many of the residual plots provided by computer software packages use a standardized version of the residuals. As demonstrated in preceding chapters, a random variable is standardized by subtracting its mean and dividing the result by its standard deviation. With the least squares method, the mean of the residuals is zero. Thus, simply dividing each residual by its standard deviation provides the **standardized residual**.

It can be shown that the standard deviation of residual i depends on the standard error of the estimate s and the corresponding value of the independent variable x_i.

Figure 14.13 Plot of the residuals against the predicted values $\hat{y}$ for Armand's Pizza Parlours



Note that equation (14.26) shows that the standard deviation of the ith residual depends on x_i because of the presence of h_i in the formula.† Once the standard deviation of each residual is calculated, we can compute the standardized residual by dividing each residual by its corresponding standard deviation.

Standard deviation of the ith residual*

$$s_{y_i - \hat{y}_i} = s\sqrt{1 - h_i} \qquad (14.26)$$

where

$s_{y_i - \hat{y}_i}$ = the standard deviation of residual i
s = the standard error of the estimate

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\Sigma(x_i - \bar{x})^2} \qquad (14.27)$$

Standardized residual for observation i

$$\frac{y_i - \hat{y}_i}{s_{y_i - \hat{y}_i}} \qquad (14.28)$$

Table 14.8 shows the calculation of the standardized residuals for Armand's Pizza Parlours. Recall that previous calculations showed $s = 13.829$. Figure 14.14 is the plot of the standardized residuals against the independent variable X.

The standardized residual plot can provide insight about the assumption that the error term ε has a normal distribution. If this assumption is satisfied, the distribution of the standardized residuals should appear to come from a standard normal probability distribution.‡

Thus, when looking at a standardized residual plot, we should expect to see approximately 95 per cent of the standardized residuals between -2 and $+2$. We see in Figure 14.14 that for the Armand's example all standardized residuals are between -2 and $+2$. Therefore, on the basis of the standardized residuals, this plot gives us no reason to question the assumption that ε has a normal distribution.

Because of the effort required to compute the estimated values $\hat{y}$, the residuals, and the standardized residuals, most statistical packages provide these values as optional regression output. Hence, residual plots can be easily obtained. For large problems computer packages are the only practical means for developing the residual plots discussed in this section.

Normal probability plot

Another approach for determining the validity of the assumption that the error term has a normal distribution is the **normal probability plot**. To show how a normal probability plot is developed, we introduce the concept of *normal scores.*

†h_i is referred to as the *leverage* of observation i. Leverage will be discussed further when we consider influential observations in Section 14.9.

*This equation actually provides an estimate of the standard deviation of the ith residual, because s is used instead of σ.

‡Because s is used instead of σ in equation (14.26), the probability distribution of the standardized residuals is not technically normal. However, in most regression studies, the sample size is large enough that a normal approximation is very good.

Table 14.8 Computation of standardized residuals for Armand's Pizza Parlours

Restaurant i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$\frac{(x_i - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}$	h_i	$s_{y_i - \hat{y}_i}$	$y_i - \hat{y}_i$	Standardized residual
1	2	−12	144	0.2535	0.3535	11.1193	−12	−1.0792
2	6	−8	64	0.1127	0.2127	12.2709	15	1.2224
3	8	−6	36	0.0634	0.1634	12.6493	−12	−0.9487
4	8	−6	36	0.0634	0.1634	12.6493	18	1.4230
5	12	−2	4	0.0070	0.1070	13.0682	−3	−0.2296
6	16	2	4	0.0070	0.1070	13.0682	−3	−0.2296
7	20	6	36	0.0634	0.1634	12.6493	−3	−0.2372
8	20	6	36	0.0634	0.1634	12.6493	9	0.7115
9	22	8	64	0.1127	0.2127	12.2709	−21	−1.7114
10	26	12	144	0.2535	0.3535	11.1193	12	1.0792
		Total	568					

Note: The values of the residuals were computed in Table 14.7.

Suppose ten values are selected randomly from a normal probability distribution with a mean of zero and a standard deviation of one, and that the sampling process is repeated over and over with the values in each sample of ten ordered from smallest to largest. For now, let us consider only the smallest value in each sample. The random variable representing the smallest value obtained in repeated sampling is called the first order statistic.

Figure 14.14 Plot of the standardized residuals against the independent variable X for Armand's Pizza Parlours

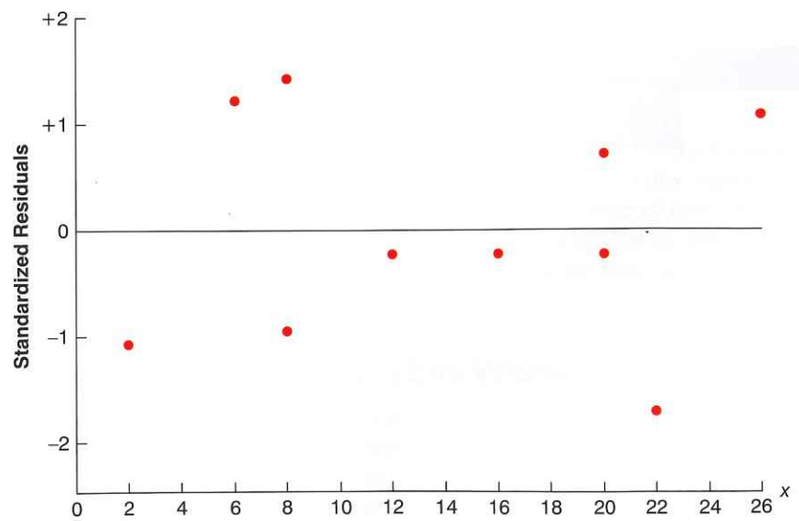


Table 14.9 Normal scores for $n = 10$

Order Statistic	Normal score
1	−1.55
2	−1.00
3	−0.65
4	−0.37
5	−0.12
6	0.12
7	0.37
8	0.65
9	1.00
10	1.55

Statisticians show that for samples of size ten from a standard normal probability distribution, the expected value of the first-order statistic is – 1.55. This expected value is called a normal score. For the case with a sample of size $n = 10$, there are ten order statistics and ten normal scores (see Table 14.9). In general, a data set consisting of n observations will have n order statistics and hence n normal scores.

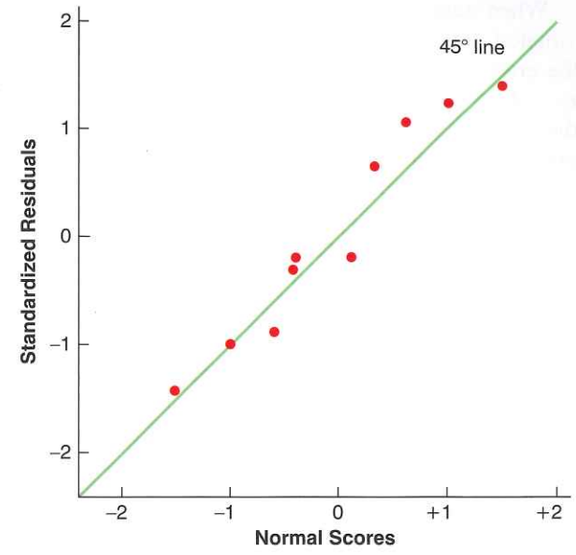
Let us now show how the ten normal scores can be used to determine whether the standardized residuals for Armand's Pizza Parlours appear to come from a standard normal probability distribution. We begin by ordering the ten standardized residuals from Table 14.8. The ten normal scores and the ordered standardized residuals are shown together in Table 14.10. If the normality assumption is satisfied, the smallest standardized residual should be close to the smallest normal score, the next smallest standardized residual should be close to the next smallest normal score and so on. If we were to develop a plot with the normal scores on the horizontal axis and the corresponding standardized residuals on the vertical axis, the plotted points should cluster closely around a 45-degree line passing through the origin if the standardized residuals are approximately normally distributed. Such a plot is referred to as a *normal probability plot.*

Figure 14.15 is the normal probability plot for the Armand's Pizza Parlours example. Judgment is used to determine whether the pattern observed deviates from

Table 14.10 Normal scores and ordered standardized residuals for Armand's Pizza Parlours

Ordered normal scores	Standardized residuals
−1.55	−1.7114
−1.00	−1.0792
−0.65	−0.9487
−0.37	−0.2372
−0.12	−0.2296
0.12	−0.2296
0.37	0.7115
0.65	1.0792
1.00	1.2224
1.55	1.4230

Figure 14.15 Normal probability plot for Armand's Pizza Parlours



the line enough to conclude that the standardized residuals are not from a standard normal probability distribution. In Figure 14.15, we see that the points are grouped closely about the line. We therefore conclude that the assumption of the error term having a normal probability distribution is reasonable. In general, the more closely the points are clustered about the 45-degree line, the stronger the evidence supporting the normality assumption. Any substantial curvature in the normal probability plot is evidence that the residuals have not come from a normal distribution. Normal scores and the associated normal probability plot can be obtained easily from statistical packages such as MINITAB.

14.9 Residual analysis: autocorrelation

In the last section we showed how residual plots can be used to detect violations of assumptions about the error term ε in the regression model. In many regression studies, particularly involving data collected over time, a special type of correlation among the error terms can cause problems; it is called **serial correlation** or **autocorrelation**. In this section we show how the **Durbin-Watson test** can be used to detect significant autocorrelation.

Autocorrelation and the Durbin-Watson test

Often, the data used for regression studies in business and economics are collected over time. It is not uncommon for the value of Y at time t, denoted by y_t, to be related to the value of Y at previous time periods. In such cases, we say autocorrelation (also called serial correlation) is present in the data. If the value of Y in time period t is related to its value in time period $t - 1$, first-order autocorrelation is present. If the value of Y in time

period t is related to the value of Y in time period $t - 2$, second-order autocorrelation is present and so on.

When autocorrelation is present, one of the assumptions of the regression model is violated: the error terms are not independent. In the case of first-order autocorrelation, the error at time t, denoted ε_t, will be related to the error at time period $t - 1$, denoted ε_{t-1}. Two cases of first-order autocorrelation are illustrated in Figure 14.16. Panel A is the case of positive autocorrelation; panel B is the case of negative autocorrelation. With positive autocorrelation we expect a positive residual in one period to be followed by a positive residual in the next period, a negative residual in one period to be followed by a negative residual in the next period and so on. With negative autocorrelation, we expect a positive residual in one period to be followed by a negative residual in the next period, then a positive residual and so on. When autocorrelation is present, serious errors can be made in performing tests of statistical significance based upon the assumed regression model. It is therefore important to be able to detect autocorrelation and take corrective action. We will show how the Durbin-Watson statistic can be used to detect first-order autocorrelation.

Suppose the values of ε are not independent but are related in the following manner:

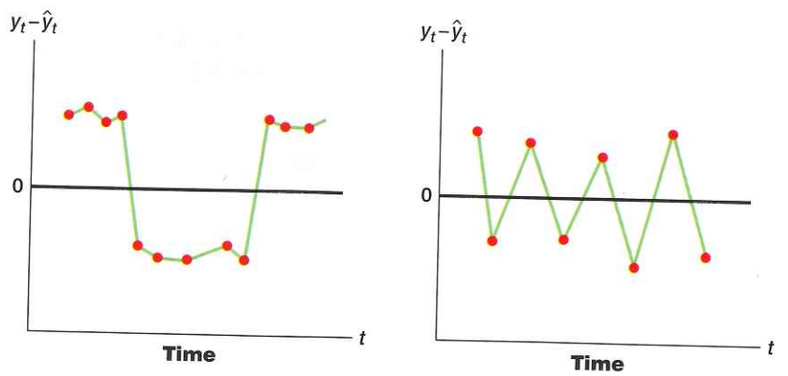
First order autocorrelation

$$\varepsilon_t = \rho\varepsilon_{t-1} + z_t \qquad (14.29)$$

where ρ is a parameter with an absolute value less than one and z_t is a normally and independently distributed random variable with a mean of zero and a variance of σ^2. From equation (16.16) we see that if $\rho = 0$, the error terms are not related, and each has a mean of zero and a variance of σ^2. In this case, there is no autocorrelation and the regression assumptions are satisfied. If $\rho > 0$, we have positive autocorrelation; if $\rho < 0$, we have negative autocorrelation. In either of these cases, the regression assumptions about the error term are violated.

The Durbin-Watson test for autocorrelation uses the residuals to determine whether $\rho = 0$. To simplify the notation for the Durbin-Watson statistic, we denote the ith residual by $e_t = y_t - \hat{y}_t$. The Durbin-Watson test statistic is computed as follows.

Figure 14.16 Two data sets with first-order autocorrelation



Panel A. Positive Autocorrelation **Panel B. Negative Autocorrelation**

Durbin-Watson test statistic

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \qquad (14.30)$$

If successive values of the residuals are close together (positive autocorrelation), the value of the Durbin-Watson test statistic will be small. If successive values of the residuals are far apart (negative autocorrelation), the value of the Durbin-Watson statistic will be large.

The Durbin-Watson test statistic ranges in value from zero to four, with a value of two indicating no autocorrelation is present. Durbin and Watson developed tables that can be used to determine when their test statistic indicates the presence of autocorrelation. Table 8 in Appendix B shows lower and upper bounds (d_L and d_U) for hypothesis tests using $\alpha = 0.05$, $\alpha = 0.025$, and $\alpha = 0.01$; n denotes the number of observations.

The null hypothesis to be tested is always that there is no autocorrelation.

$$H_0: \rho = 0$$

The alternative hypothesis to test for positive autocorrelation is

$$H_1: \rho > 0$$

The alternative hypothesis to test for negative autocorrelation is

$$H_1: \rho < 0$$

A two-sided test is also possible. In this case the alternative hypothesis is

$$H_1: \rho \neq 0$$

Figure 14.17 shows how the values of d_L and d_U in Table 8 are used to test for autocorrelation.

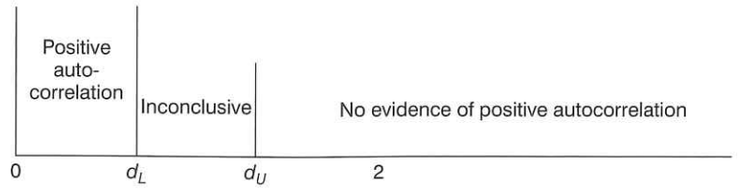
Panel A illustrates the test for positive autocorrelation. If $d < d_L$, we conclude that positive autocorrelation is present. If $d_L \leq d \leq d_U$, we say the test is inconclusive. If $d > d_U$, we conclude that there is no evidence of positive autocorrelation.

Panel B illustrates the test for negative autocorrelation. If $d > 4 - d_L$, we conclude that negative autocorrelation is present. If $4 - d_U \leq d \leq 4 - d_L$, we say the test is inconclusive. If $d < 4 - d_U$, we conclude that there is no evidence of negative autocorrelation.

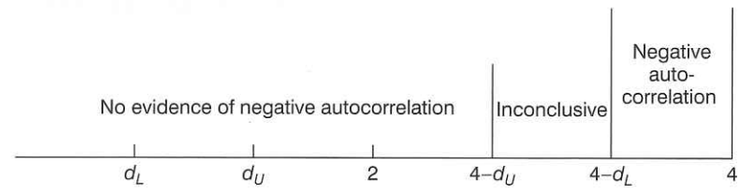
Note: Entries in Table 8 are the critical values for a one-tailed Durbin-Watson test for autocorrelation. For a two-tailed test, the level of significance is doubled.

Panel C illustrates the two-sided test. If $d < d_L$ or $d > 4 - d_L$, we reject H_0 and conclude that autocorrelation is present. If $d_L \leq d \leq d_U$ or $4 - d_U \leq d \leq 4 - d_L$, we say the test is inconclusive. If $d_U \leq d \leq 4 - d_U$, we conclude that there is no evidence of autocorrelation.

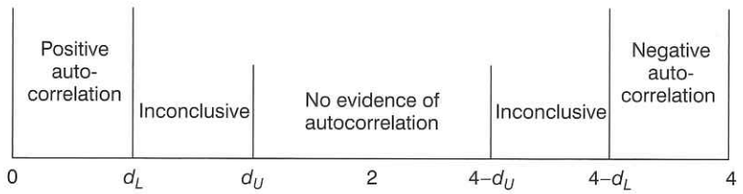
Figure 14.17 Hypothesis test for autocorrelation using the Durbin-Watson test



Panel A. Test for Positive Autocorrelation



Panel B. Test for Negative Autocorrelation



Panel C. Two-Sided Test for Autocorrelation

If significant autocorrelation is identified, we should investigate whether we omitted one or more key independent variables that have time-ordered effects on the dependent variable. If no such variables can be identified, including an independent variable that measures the time of the observation (for instance, the value of this variable could be one for the first observation, two for the second observation and so on) will sometimes eliminate or reduce the autocorrelation. When these attempts to reduce or remove autocorrelation do not work, transformations on the dependent or independent variables can prove helpful; a discussion of such transformations can be found in more advanced texts on regression analysis.

Note that the Durbin-Watson tables list the smallest sample size as 15. The reason is that the test is generally inconclusive for smaller sample sizes; in fact, many statisticians believe the sample size should be at least 50 for the test to produce worthwhile results.

Exercises

Methods

25 Given are data for two variables, X and Y.

x_i	6	11	15	18	20
y_i	6	8	12	20	30

a. Develop an estimated regression equation for these data.
b. Compute the residuals.
c. Develop a plot of the residuals against the independent variable X. Do the assumptions about the error terms seem to be satisfied?
d. Compute the standardized residuals.
e. Develop a plot of the standardized residuals against $\hat{y}$. What conclusions can you draw from this plot?

26 The following data were used in a regression study.

Observation	x_i	y_i	Observation	x_i	y_i
1	2	4	6	7	6
2	3	5	7	7	9
3	4	4	8	8	5
4	5	6	9	9	11
5	7	4			

a. Develop an estimated regression equation for these data.
b. Construct a plot of the residuals. Do the assumptions about the error term seem to be satisfied?

Applications

27 Data on advertising expenditures and revenue (in thousands of euros) for the Four Seasons Restaurant follow.

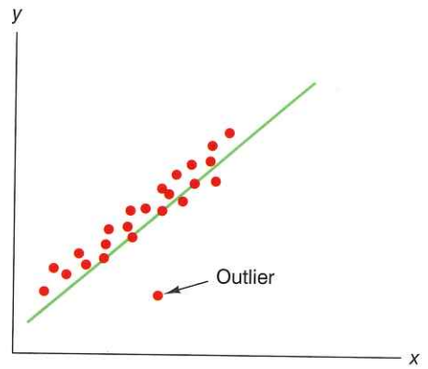
Advertising expenditures	Revenue
1	19
2	32
4	44
6	40
10	52
14	53
20	54

a. Let X equal advertising expenditures and Y equal revenue. Use the method of least squares to develop a straight line approximation of the relationship between the two variables.
b. Test whether revenue and advertising expenditures are related at a 0.05 level of significance.
c. Prepare a residual plot of $y - \hat{y}$ versus $\hat{y}$. Use the result from part (a) to obtain the values of $\hat{y}$.
d. What conclusions can you draw from residual analysis? Should this model be used, or should we look for a better one?

28 Refer to exercise 6, where an estimated regression equation relating years of experience and annual sales was developed.

a. Compute the residuals and construct a residual plot for this problem.
b. Do the assumptions about the error terms seem reasonable in light of the residual plot?

Figure 14.18 A data set with an outlier



14.10 Residual analysis: outliers and influential observations

In Section 14.8 we showed how residual analysis could be used to determine when violations of assumptions about the regression model occur. In this section, we discuss how residual analysis can be used to identify observations that can be classified as outliers or as being especially influential in determining the estimated regression equation. Some steps that should be taken when such observations occur are discussed.

Detecting outliers

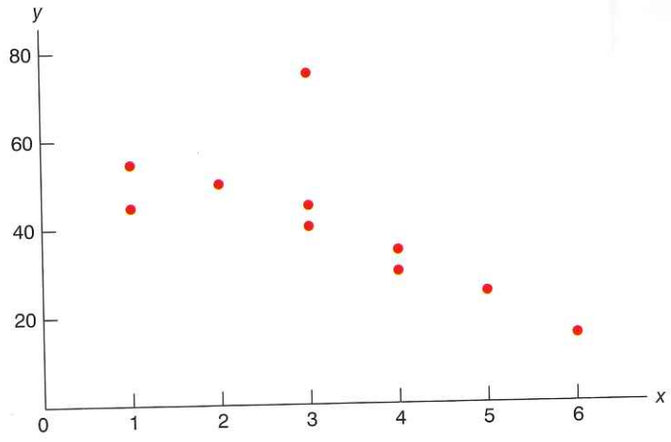
Figure 14.18 is a scatter diagram for a data set that contains an **outlier**, a data point (observation) that does not fit the trend shown by the remaining data. Outliers represent observations that are suspect and warrant careful examination. They may represent erroneous data; if so, the data should be corrected. They may signal a violation of model assumptions; if so, another model should be considered. Finally, they may simply be unusual values that occurred by chance. In this case, they should be retained.

To illustrate the process of detecting outliers, consider the data set in Table 14.11; Figure 14.19 is a scatter diagram. Except for observation 4 ($x_4 = 3$, $y_4 = 75$), a pattern

Table 14.11 Data set illustrating the effect of an outlier

x_i	y_i
1	45
1	55
2	50
3	75
3	40
3	45
4	30
4	35
5	25
6	15

Figure 14.19 Scatter diagram for outlier data set



suggesting a negative linear relationship is apparent. Indeed, given the pattern of the rest of the data, we would expect y_4 to be much smaller and hence would identify the corresponding observation as an outlier. For the case of simple linear regression, one can often detect outliers by simply examining the scatter diagram.

The standardized residuals can also be used to identify outliers. If an observation deviates greatly from the pattern of the rest of the data (e.g. the outlier in Figure 14.18), the corresponding standardized residual will be large in absolute value. Many computer packages automatically identify observations with standardized residuals that are large in absolute value. In Figure 14.20 we show the MINITAB output from a regression analysis of the data in Table 14.11. The next to last line of the output shows that the standardized residual for observation 4 is 2.67. MINITAB identifies any observation with a standardized residual of less than -2 or greater than $+2$ as an unusual observation; in such cases, the observation is printed on a separate line with an R next to the standardized residual, as shown in Figure 14.20. With normally distributed errors, standardized residuals should be outside these limits approximately 5 per cent of the time.

In deciding how to handle an outlier, we should first check to see whether it is a valid observation. Perhaps an error was made in initially recording the data or in entering the data into the computer file. For example, suppose that in checking the data for the outlier in Table 14.11, we find an error; the correct value for observation 4 is $x_4 = 3$, $y_4 = 30$. Figure 14.21 is the MINITAB output obtained after correction of the value of y_4. We see that using the incorrect data value substantially affected the goodness of fit. With the correct data, the value of R-sq increased from 49.7 per cent to 83.8 per cent and the value of b_0 decreased from 64.958 to 59.237. The slope of the line changed from -7.331 to -6.949. The identification of the outlier enabled us to correct the data error and improve the regression results.

Detecting influential observations

Sometimes one or more observations exert a strong influence on the results obtained. Figure 14.22 shows an example of an **influential observation** in simple linear regression. The estimated regression line has a negative slope. However, if the influential observation were dropped from the data set, the slope of the estimated regression line

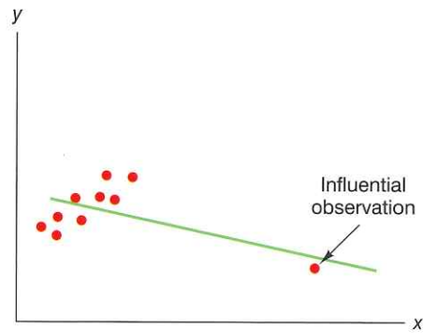
Figure 14.20 MINITAB output for regression analysis of the outlier data set

```
Regression Analysis: y versus x

The regression equation is
y = 65.0 - 7.33 x


Predictor     Coef  SE Coef      T      P
Constant   64.958    9.258   7.02  0.000
x          -7.331    2.608  -2.81  0.023


S = 12.6704   R-Sq = 49.7%   R-Sq(adj) = 43.4%


Analysis of Variance

Source           DF      SS      MS     F      P
Regression        1  1268.2  1268.2  7.90  0.023
Residual Error    8  1284.3   160.5
Total             9  2552.5


Unusual Observations

Obs     x      y     Fit  SE Fit  Residual  St Resid
  4  3.00  75.00  42.97    4.04     32.03      2.67R

R denotes an observation with a large standardized residual.
```

Figure 14.21 MINITAB output for the revised outlier data set

```
Regression Analysis: y versus x

The regression equation is
y = 59.2 - 6.95 x


Predictor     Coef  SE Coef      T      P
Constant   59.237    3.835  15.45  0.000
x          -6.949    1.080  -6.43  0.000


S = 5.24808   R-Sq = 83.8%   R-Sq(adj) = 81.8%


Analysis of Variance

Source           DF      SS      MS      F      P
Regression        1  1139.7  1139.7  41.38  0.000
Residual Error    8   220.3    27.5
Total             9  1360.0
```


Figure 14.22 A data set with an influential observation



would change from negative to positive and the y-intercept would be smaller. Clearly, this one observation is much more influential in determining the estimated regression line than any of the others; dropping one of the other observations from the data set would have little effect on the estimated regression equation.

Influential observations can be identified from a scatter diagram when only one independent variable is present. An influential observation may be an outlier (an observation with a Y value that deviates substantially from the trend), it may correspond to an X value far away from its mean (e.g. see Figure 14.22), or it may be caused by a combination of the two (a somewhat off-trend Y value and a somewhat extreme X value).

Because influential observations may have such a dramatic effect on the estimated regression equation, they must be examined carefully. We should first check to make sure that no error was made in collecting or recording the data. If an error occurred, it can be corrected and a new estimated regression equation can be developed. If the observation is valid, we might consider ourselves fortunate to have it. Such a point, if valid, can contribute to a better understanding of the appropriate model and can lead to a better estimated regression equation. The presence of the influential observation in Figure 14.22, if valid, would suggest trying to obtain data on intermediate values of X to understand better the relationship between X and Y.

Observations with extreme values for the independent variables are called **high leverage points**. The influential observation in Figure 14.22 is a point with high leverage. The leverage of an observation is determined by how far the values of the independent variables are from their mean values. For the single-independent-variable case, the leverage of the ith observation, denoted h_i, can be computed by using equation (14.31).

Leverage of observation *i*

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\Sigma(x_i - \bar{x})^2} \qquad (14.31)$$

From the formula, it is clear that the farther x_i is from its mean $\bar{x}$, the higher the leverage of observation i.

Many statistical packages automatically identify observations with high leverage as part of the standard regression output. As an illustration of how the MINITAB statistical package identifies points with high leverage, let us consider the data set in Table 14.12.

Table 14.12 Data set with a high leverage observation

x_i	y_i
10	125
10	130
15	120
20	115
20	120
25	110
70	100

From Figure 14.23, a scatter diagram for the data set in Table 14.12, it is clear that observation 7 ($X = 70$, $Y = 100$) is an observation with an extreme value of X. Hence, we would expect it to be identified as a point with high leverage. For this observation, the leverage is computed by using equation (14.31) as follows.

$$h_7 = \frac{1}{n} + \frac{(x_7 - \bar{x})^2}{\Sigma (x_i - \bar{x})^2} = \frac{1}{7} + \frac{(70 - 24.286)^2}{2621.43} = 0.94$$

For the case of simple linear regression, MINITAB identifies observations as having high leverage if $h_i > 6/n$; for the data set in Table 14.12, $6/n = 6/7 = 0.86$. Because $h_7 = 0.94 > 0.86$, MINITAB will identify observation 7 as an observation whose X value gives it large influence. Figure 14.24 shows the MINITAB output for a regression analysis of this data set. Observation 7 ($X = 70$, $Y = 100$) is identified as having large influence; it is printed on a separate line at the bottom, with an X in the right margin.

Influential observations that are caused by an interaction of large residuals and high leverage can be difficult to detect. Diagnostic procedures are available that take both into account in determining when an observation is influential. One such measure, called Cook's D statistic, will be discussed in Chapter 15.

Figure 14.23 Scatter diagram for the data set with a high leverage observation

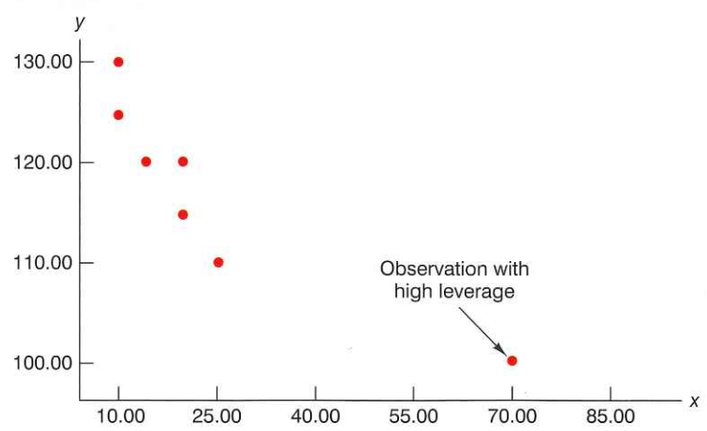


Figure 14.24 MINITAB output for the data set with a high leverage observation

```
Regression Analysis: y versus x

The regression equation is
y = 127 - 0.425 x

Predictor       Coef  SE Coef      T      P
Constant     127.466    2.961  43.04  0.000
x           -0.42507  0.09537  -4.46  0.007

S = 4.88282   R-Sq = 79.9%   R-Sq(adj) = 75.9%

Analysis of Variance

Source           DF      SS      MS      F      P
Regression        1  473.65  473.65  19.87  0.007
Residual Error    5  119.21   23.84
Total             6  592.86

Unusual Observations

Obs     x       y     Fit  SE Fit  Residual  St Resid
  7  70.0  100.00   97.71    4.73      2.29      1.91 X

X denotes an observation whose X value gives it large leverage.
```

Exercises

Methods

29 Consider the following data for two variables, X and Y.

x_i	135	110	130	145	175	160	120
y_i	145	100	120	120	130	130	110

a. Compute the standardized residuals for these data. Do there appear to be any outliers in the data? Explain.
b. Plot the standardized residuals against $\hat{y}$. Does this plot reveal any outliers?
c. Develop a scatter diagram for these data. Does the scatter diagram indicate any outliers in the data? In general, what implications does this finding have for simple linear regression?

30 Consider the following data for two variables, X and Y.

x_i	4	5	7	8	10	12	12	22
y_i	12	14	16	15	18	20	24	19

a. Compute the standardized residuals for these data. Do there appear to be any outliers in the data? Explain.

b. Compute the leverage values for these data. Do there appear to be any influential observations in these data? Explain.
c. Develop a scatter diagram for these data. Does the scatter diagram indicate any influential observations? Explain.

For additional online summary questions and answers go to the companion website at www.cengage.co.uk/aswsbe2

Summary

In this chapter we showed how regression analysis can be used to determine how a dependent variable Y is related to an independent variable X. In simple linear regression, the regression model is $Y = \beta_0 + \beta_1 x + \varepsilon$. The simple linear regression equation $E(\hat{Y}) = \beta_0 + \beta_1 x$ describes how the mean or expected value of Y is related to X. We used sample data and the least squares method to develop the estimated regression equation $\hat{y} = b_0 + b_1 x$ for a given value x of X. In effect, b_0 and b_1 are the sample statistics used to estimate the unknown model parameters β_0 and β_1.

The coefficient of determination was presented as a measure of the goodness of fit for the estimated regression equation; it can be interpreted as the proportion of the variation in the dependent variable Y that can be explained by the estimated regression equation. We reviewed correlation as a descriptive measure of the strength of a linear relationship between two variables.

The assumptions about the regression model and its associated error term ε were discussed, and t and F tests, based on those assumptions, were presented as a means for determining whether the relationship between two variables is statistically significant. We showed how to use the estimated regression equation to develop confidence interval estimates of the mean value of Y and prediction interval estimates of individual values of Y.

The chapter concluded with a section on the computer solution of regression problems and two sections on the use of residual analysis to validate the model assumptions and to identify outliers and influential observations.

Key terms

ANOVA table
Autocorrelation
Coefficient of determination
Confidence interval
Correlation coefficient
Dependent variable
Durbin-Watson test
Estimated regression equation
High leverage points
Independent variable
Influential observation
*i*th residual
Least squares method
Mean square error
Normal probability plot
Outlier
Prediction interval
Regression equation
Regression model
Residual analysis
Residual plot
Scatter diagram
Serial correlation
Simple linear regression
Standard error of the estimate
Standardized residual

Key formulae

Simple linear regression model

$$Y = \beta_0 + \beta_1 x + \varepsilon \quad (14.1)$$

Simple linear regression equation

$$E(Y) = \beta_0 + \beta_1 x \quad (14.2)$$

Estimated simple linear regression equation

$$\hat{y} = b_0 + b_1 x \quad (14.3)$$

Least squares criterion

$$\text{Min } \Sigma (y_i - \hat{y}_i)^2 \quad (14.5)$$

Slope and y-intercept for the estimated regression equation

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} \quad (14.6)$$

$$b_0 = \bar{y} - b_1 \bar{x} \quad (14.7)$$

Sum of squares due to error

$$\text{SSE} = \Sigma(y_i - \hat{y}_i)^2 \quad (14.8)$$

Total sum of squares

$$\text{SST} = \Sigma(y_i - \bar{y})^2 \quad (14.9)$$

Sum of squares due to regression

$$\text{SSR} = \Sigma(\hat{y}_i - \bar{y})^2 \quad (14.10)$$

Relationship among SST, SSR, and SSE

$$\text{SST} = \text{SSR} + \text{SSE} \quad (14.11)$$

Coefficient of determination

$$r^2 = \frac{\text{SSR}}{\text{SST}} \quad (14.12)$$

Sample correlation coefficient

$$r_{xy} = (\text{sign of } b_1)\sqrt{\text{Coefficient of determination}} = (\text{sign of } b_1)\sqrt{r^2} \quad (14.13)$$

Mean square error (estimate of s^2)

$$s^2 = \text{MSE} = \frac{\text{SSE}}{n-2} \quad (14.15)$$

Standard error of the estimate

$$s = \sqrt{\frac{SSE}{(n-2)}} \quad (14.16)$$

Standard deviation of b_1

$$\sigma_{b_1} = \frac{\sigma}{\sqrt{\Sigma(x_i - \bar{x})^2}} \quad (14.17)$$

Estimated standard deviation of b_1

$$s_{b_1} = \frac{s}{\sqrt{\Sigma(x_i - \bar{x})^2}} \quad (14.18)$$

t test statistic

$$t = \frac{b_1}{s_{b_1}} \quad (14.19)$$

Mean square regression

$$\text{MSR} = \frac{\text{SSR}}{\text{Number of independent variables}} \quad (14.20)$$

***F* test statistic**

$$F = \frac{\text{MSR}}{\text{MSE}} \quad (14.21)$$

Confidence interval for $E(Y_p)$

$$\hat{y}_p \pm t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}} \quad (14.22)$$

Prediction interval for Y_p

$$\hat{y}_p \pm t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}} \quad (14.23)$$

Residual for observation *i*

$$y_i - \hat{y}_i \quad (14.24)$$

Standard deviation of the *i*th residual

$$s_{y_i - \hat{y}_i} = s\sqrt{1 - h_i} \quad (14.26)$$

Standardized residual for observation *i*

$$\frac{y_i - \hat{y}_i}{s_{y_i - \hat{y}_i}} \quad (14.28)$$

First order autocorrelation

$$\varepsilon_t = \rho \varepsilon_{t-1} + z_t \quad (14.29)$$

Durbin-Watson test statistic

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \qquad (14.30)$$

Leverage of observation *i*

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\Sigma(x_i - \bar{x})^2} \qquad (14.31)$$

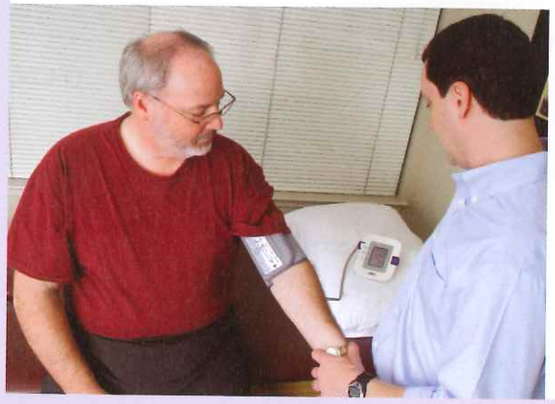
Case problem 1 Investigating the relationship between weight loss and triglyceride level reduction†

Epidemiological studies have shown that there is a relationship between raised blood levels of triglyceride and coronary heart disease but it is not certain how important a risk factor triglycerides are. It is believed that exercise and lower consumption of fatty acids can help to reduce triglyceride levels.*

In 1998 Knoll Pharmaceuticals received authorization to market sibutramine for the treatment of obesity in the US. One of their suite of studies involved 35 obese patients who followed a treatment regime comprising a combination of diet, exercise and drug treatment.

Each patient's weight and triglyceride level were recorded at the start (known as *baseline*) and at week eight. The information recorded for each patient was:

- Patient ID.
- Weight at baseline (kg).
- Weight at week 8 (kg).
- Triglyceride level at baseline (mg/dl).
- Triglyceride level at week 8 (mg/dl).



Doctor checking an overweight patient's blood pressure. Digital readout indicates high blood pressure and pulse rate. © Eliza Snow.

The results are shown below.

Patient ID	Weight at baseline	Weight at week 8	Triglyceride level at baseline	Triglyceride level at week 8
201	84.0	82.4	90	131
202	88.8	87.0	137	82
203	87.0	81.8	182	152
204	84.5	80.4	72	72
205	69.4	69.0	143	126
206	104.7	102.0	96	157
207	90.0	87.6	115	88
208	89.4	86.8	124	123
209	95.2	92.8	188	255
210	108.1	100.9	167	87
211	93.9	90.2	143	213
212	83.4	75.0	143	102
213	104.4	102.9	276	313
214	103.7	95.7	84	84
215	99.2	99.2	142	135
216	95.6	88.5	64	114
217	126.0	123.2	226	152
218	103.7	95.5	199	120
219	133.1	130.8	212	156
220	85.0	80.0	268	250
221	83.8	77.9	111	107
222	104.5	98.3	132	117
223	76.8	73.2	165	96
224	90.5	88.9	57	63
225	106.9	103.7	163	131
226	81.5	78.9	111	54
227	96.5	94.9	300	241
228	103.0	97.2	192	124
229	127.5	124.7	176	215
230	103.2	102.0	146	138

(continued)

Patient ID	Weight at baseline	Weight at week 8	Triglyceride level at baseline	Triglyceride level at week 8
231	113.5	115.0	446	795
232	107.0	99.2	232	63
233	106.0	103.5	255	204
234	114.9	105.3	187	144
235	103.4	96.0	154	96

Managerial report

1 Are weight loss and triglyceride level reduction (linearly) correlated?

2 Is there a linear relationship between weight loss and triglyceride level reduction?

3 How can a more detailed regression analysis be undertaken?

†Data in this case study reproduced with permission from STARS (www.stars.ac.uk).
*Triglycerides are lipids (fats) which are formed from glycerol and fatty acids. They can be absorbed into the body from food intake, particularly from fatty food, or produced in the body itself when the uptake of energy (food) exceeds the expenditure (exercise). Triglycerides provide the principal energy store for the body. Compared with carbohydrates or proteins, triglycerides produce a substantially higher number of calories per gram.

Case Problem 2 US Department of Transportation

As part of a study on transportation safety, the US Department of Transportation collected data on the number of fatal accidents per 1000 licences and the percentage of licensed drivers under the age of 21 in a sample of 42 cities. Data collected over a one-year period follow. These data are available on the CD accompanying the text in the file named Safety.

Percentage under 21	Fatal accidents per 1000 licences	Percentage under 21	Fatal accidents per 1000 licences
13	2.962	17	4.100
12	0.708	8	2.190
8	0.885	16	3.623
12	1.652	15	2.623
11	2.091	9	0.835
17	2.627	8	0.820
18	3.830	14	2.890
8	0.368	8	1.267
13	1.142	15	3.224
8	0.645	10	1.014
9	1.028	10	0.493
16	2.801	14	1.443
12	1.405	18	3.614
9	1.433	10	1.926
10	0.039	14	1.643
9	0.338	16	2.943
11	1.849	12	1.913
12	2.246	15	2.814
14	2.855	13	2.634
14	2.352	9	0.926
11	1.294	17	3.256

A fatal car accident. © Celso Pupo.

Managerial report

1. Develop numerical and graphical summaries of the data.
2. Use regression analysis to investigate the relationship between the number of fatal accidents and the percentage of drivers under the age of 21. Discuss your findings.
3. What conclusion and recommendations can you derive from your analysis?

Case Problem 3 Can we detect dyslexia?*

Data were collected on 34 pre-school children and then in follow-up tests (on the same children) three years later when they were seven years old.

Scores were obtained from a variety of tests on all the children at age four when they were at nursery school. The tests were:

- Knowledge of vocabulary, measured by the British Picture Vocabulary Test (BPVT) in three versions – as raw scores, standardized scores and percentile norms.
- Another vocabulary test – non-word repetition.
- Motor skills, where the children were scored on the time in seconds to complete five different peg board tests.
- Knowledge of prepositions, scored as the number correct out of ten.
- Three tests on the use of rhyming, scored as the number correct out of ten.

Three years later the same children were given a reading test, from which a reading deficiency was calculated as Reading Age – Chronological Age (in months), this being known as Reading Age Deficiency (RAD). The children were then classified into 'poor' or 'normal' readers, depending on their RAD scores. Poor reading ability is taken as an indication of potential dyslexia.

One purpose of this study is to identify which of the tests at age four might be used as predictors of poor reading ability, which in turn is a possible indication of dyslexia.

*Data in this case study reproduced with permission from STARS (www.stars.ac.uk)

Data

The data set *Dyslexia* contains 18 variables:

- Child Code — an identification number for each child (1–34)
- Sex — m for male, f for female

The BPVT scores:

- BPVT raw — the raw score
- BPVT std — the standardized score
- BPVT % norm — cumulative percentage scores
- Non-wd repn — score for non-word repetition

Scores in motor skills:

- Pegboard set1 to Pegboard set5 — the time taken to complete each test
- Mean — child's average over the pegboard tests
- Preps Score — knowledge of prepositions (6–10)

Scores in rhyming tests (2–10):

- Rhyme set1
- Rhyme set2
- Rhyme set3
- RAD
- Poor/Normal — RAD scores, categorized as 1 = normal, 2 = poor

Details for ten records from the dataset are shown below.

Child code	Sex	BPVT raw	BPVT std	BPVT % norm	Non-wd repn	Pegboard set1	Pegboard set2	Pegboard set3	Pegboard set4	Pegboard set5
1	m	29	88	22	15	20.21	28.78	28.04	20.00	24.37
2	m	21	77	6	11	26.34	26.20	20.35	28.25	20.87
3	m	50	107	68	17	21.13	19.88	17.63	16.25	19.76
4	m	23	80	9	5	16.46	16.47	16.63	14.16	17.25
5	f	35	91	28	13	17.88	15.13	17.81	18.41	15.99
6	m	36	97	42	16	20.41	18.64	17.03	16.69	14.47
7	f	47	109	72	25	21.31	18.06	28.00	21.88	18.03
8	m	32	92	30	12	14.57	14.22	13.47	12.29	18.38
9	f	38	101	52	14	22.07	22.69	21.19	22.72	20.62
10	f	44	105	63	15	16.40	14.48	13.83	17.59	34.68

Child code	Mean	Preps score	Rhyme set1	Rhyme set2	Rhyme set3	RAD	Poor/ normal
1	24.3	6	5	5	5	−6.50	P
2	24.4	9	3	3	4	−7.33	P
3	18.9	10	9	8	*	49.33	N
4	16.2	7	4	6	4	−11.00	P
5	17.0	10	10	6	6	−2.67	N
6	17.5	10	6	5	5	−8.33	P
7	21.5	8	9	10	10	26.33	N
8	14.6	10	8	6	3	9.00	N
9	21.9	9	10	10	7	2.67	N
10	19.4	10	7	8	4	9.67	N

A young boy with dyslexia reads a book. © karen squires.

Managerial report

1. Is there a (linear) relationship between scores in tests at ages four and seven?
2. Can we predict RAD from scores at age four?

Software Section for Chapter 14

Regression analysis using MINITAB

In Section 14.7 we discussed the computer solution of regression problems by showing MINITAB's output for the Armand's Pizza Parlours problem. In this section, we describe the steps required to generate the MINITAB computer solution. First, the data must be entered in a MINITAB worksheet. Student population data are entered in column C1 and quarterly sales data are entered in column C2. The variable names Pop and Sales are entered as the column headings on the worksheet. In subsequent steps, we refer to the data by using the variable names Pop and Sales or the column indicators C1 and C2. The steps involved in using MINITAB to produce the regression results shown in Figure 14.10 follow.

Step 1 **Stat > Regression > Regression** [Main menu bar]

Step 2 Enter Sales in the **Response** box [**Regression** panel]
Enter Pop in the **Predictors** box
Click the **Options** button
Enter 10 in the **Prediction intervals for new observations** box
Click **OK**
(The MINITAB regression panel provides additional capabilities that can be obtained by selecting the desired options. For instance, to obtain a residual plot that shows the predicted value of the dependent variable on the horizontal axis and the standardized residual values on the vertical axis, click the **Graphs** button. Select **Standardized** under Residuals for Plots. Select **Residuals versus fits** under Residual Plots). Click **OK.** When the Regression panel reappears: Click **OK**.

Regression analysis using EXCEL

In this section we will illustrate how EXCEL's Regression tool can be used to perform the regression analysis computations for the Armand's Pizza Parlours problem. Refer to Figure 14.25 as we describe the steps involved. The labels Restaurant, Population and Sales are entered into cells A1:C1 of the worksheet. To identify each of the ten observations, we entered the numbers 1 through 10 into cells A2:A11. The sample data are entered into cells B2:C11. The steps involved in using the Regression tool for regression analysis follow.

Figure 14.25 EXCEL solution to the Armand's Pizza Parlours problem

	A	B	C	D	E	F	G	H	I
1	**Restaurant**	**Population**	**Sales**						
2	1	2	58						
3	2	6	105						
4	3	8	88						
5	4	8	118						
6	5	12	117						
7	6	16	137						
8	7	20	157						
9	8	20	169						
10	9	22	149						
11	10	26	202						
12									
13	SUMMARY OUTPUT								
14									
15	*Regression Statistics*								
16	Multiple R	0.9501							
17	R Square	0.9027							
18	Adjusted R Sq	0.8906							
19	Standard Error	13.8293							
20	Observations	10							
21									
22	ANOVA								
23		*df*	*SS*	*MS*	*F*	*Significance F*			
24	Regression	1	14200	14200	74.2484	0.0000			
25	Residual	8	1530	191.25					
26	Total	9	15730						
27									
28		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
29	Intercept	60	9.2260	6.5033	0.0002	38.7247	81.2753	29.0431	90.9569
30	Population	5	0.5803	8.6167	0.0000	3.6619	6.3381	3.0530	6.9470
31									

Step 1 **Data > Data Analysis > Regression** [Main menu bar]

Step 2 Enter C1:C11 in the **Input Y Range** box [**Regression** panel]
Enter B1:B11 in the **Input X Range** box
Select **Labels**
Select Confidence **Level.** Enter 99 in the **Confidence Level** box
Select **Output Range**
Enter A13 in the **Output Range** box (to identify the upper left corner of the section of the worksheet where the output will appear)
Click **OK**

The first section of the output, titled *Regression Statistics*, contains summary statistics such as the coefficient of determination (R Square). The second section of the output, titled ANOVA, contains the analysis of variance table. The last section of the output, which is not titled, contains the estimated regression coefficients and related information. We will begin our discussion of the interpretation of the regression output with the information contained in cells A28:I30.

Interpretation of estimated regression equation output

The y intercept of the estimated regression line, $b_0 = 60$, is shown in cell B29, and the slope of the estimated regression line, $b_1 = 5$, is shown in cell B30. The label Intercept in cell A29 and the label Population in cell A30 are used to identify these two values. In Section 14.5 we showed that the estimated standard deviation of b_1 is $s_{b_1} = 0.5803$.

Note that the value in cell C30 is the standard error, or standard deviation, s_{b_1} of b_1. Recall that the t test for a significant relationship required the computation of the t statistic, $t = b_1/s_{b_1}$. For the Armand's data, the value of t that we computed was $t = 5/0.5803 = 8.62$. The label in cell D28, *t Stat*, reminds us that cell D30 contains the value of the t test statistic.

The value in cell E30 is the p-value associated with the t test for significance. EXCEL has displayed the p-value in cell E30 using scientific notation. To obtain the decimal value, we move the decimal point five places to the left, obtaining a value of 0.0000255. Because the p-value $= 0.0000255 < \alpha = 0.01$, we can reject H_0 and conclude that we have a significant relationship between student population and quarterly sales.

Cells F28:I30 refer to confidence interval estimates of the y intercept and slope of the estimated regression equation. EXCEL always provides the lower and upper limits for a 95 per cent confidence interval. Recall that in step 4 we selected Confidence Level and entered 99 in the Confidence Level box. As a result, EXCEL's Regression tool also provides the lower and upper limits for a 99 per cent confidence interval. The value in cell H30 is the lower limit for the 99 per cent confidence interval estimate of β_1 and the value in cell I30 is the upper limit. Thus, after rounding, the 99 per cent confidence interval estimate of β_1 is 3.05 to 6.95. The values in cells F30 and G30 provide the lower and upper limits for the 95 per cent confidence interval. Thus, the 95 per cent confidence interval is 3.66 to 6.34.

Interpretation of ANOVA output

The information in cells A22:F26 is a summary of the analysis of variance computations. The three sources of variation are labelled Regression, Residual and Total. The label *df* in cell B23 stands for degrees of freedom, the label *SS* in cell C23 stands for sum of squares, and the label *MS* in cell D23 stands for mean square.

In Section 14.5 we stated that the mean square error, obtained by dividing the error or residual sum of squares by its degrees of freedom, provides an estimate of σ^2. The value s^2 in cell D25, 191.25, is the mean square error for the Armand's regression output. In Section 14.5 we showed that an F test could also be used to test for significance in regression.

The value in cell F24, 0.0000, is the p-value associated with the F test for significance. Because the p-value $= 0.0000 < \alpha = 0.01$, we can reject H_0 and conclude that we have a significant relationship between student population and quarterly sales. The label EXCEL uses to identify the p-value for the F test for significance, shown in cell F23, is *Significance F*.

Interpretation of regression statistics output

The coefficient of determination, 0.9027, appears in cell B17; the corresponding label, R Square, is shown in cell A17. The square root of the coefficient of determination provides the sample correlation coefficient (though EXCEL always shows the positive square root of R^2) of 0.9501 shown in cell B16. Note that EXCEL uses the label Multiple R (cell A16) to identify this value. In cell A19, the label Standard Error is used to identify the value of the standard error of the estimate shown in cell B19. Thus, the standard error of the estimate is 13.8293. We caution the reader to keep in mind that in the EXCEL output, the label Standard Error appears in two different places. In the Regression Statistics section of the output, the label Standard Error refers to the estimate of σ. In the Estimated Regression Equation section of the output, the label *Standard Error* refers to s_{b_1}, the standard deviation of the sampling distribution of b_1.

Regression analysis using PASW

First, the data must be entered in a PASW worksheet. In 'Data View' mode, restaurants are entered in rows 1 to 10 of the leftmost column. This is automatically labelled by the system V1. Similarly population and sales details are entered in the two immediately adjacent columns to the right and are labelled V2 and V3 respectively. The latter variable names can then be changed to Restaurants, Pop and Sales in 'Variable View' mode. The following command sequence describes how PASW generates the regression results shown in Figure 14.26.

Step 1 **Analyze > Regression > Linear** [Main menu bar]

Step 2 Enter Sales in the **Dependent** box [**Linear** panel]
Enter Pop in the **Independent(s)** box
(In an analogous way to MINITAB, by clicking on the **Plots** button, a variety of residual plots can also be obtained.)
Click **OK**

Figure 14.26 PASW solution to the Armand's Pizza Parlours problem

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.950[a]	.903	.891	13.82932

a. Predictors: (Constant), Population

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	14200.000	1	14200.000	74.248	.000[a]
	Residual	1530.000	8	191.250		
	Total	15730.000	9			

a. Predictors: (Constant), Population

b. Dependent Variable: Sales

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	60.000	9.226		6.503	.000
	Population	5.000	.580	.950	8.617	.000

a. Dependent Variable: Sales