Chapter 15

Multiple Regression

Learning objectives

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

1 Understand how multiple regression analysis can be used to develop relationships involving one dependent variable and several independent variables.

2 Interpret the coefficients in a multiple regression analysis.

3 Appreciate the background assumptions necessary to conduct statistical tests involving the hypothesized regression model.

4 Understand the role of computer packages in performing multiple regression analysis.

5 Interpret and use computer output to develop the estimated regression equation.

6 Determine how good a fit is provided by the estimated regression equation.

7 Test the significance of the regression equation.

8 Understand how multicollinearity affects multiple regression analysis.

9 Understand how residual analysis can be used to make a judgment as to the appropriateness of the model, identify outliers and determine which observations are influential.

10 Understand how logistic regression is used for regression analyses involving a binary dependent variable.

In Chapter 14 we presented simple linear regression and demonstrated its use in developing an estimated regression equation that describes the relationship between two variables. Recall that the variable being predicted or explained is called the dependent variable and the variable being used to predict or explain the dependent variable is called the independent variable. In this chapter we continue our study of regression analysis by considering situations involving two or more independent variables. This subject area, called **multiple regression analysis**, enables us to consider more than one potential predictor and thus obtain better estimates than are possible with simple linear regression.

15.1 Multiple regression model

Multiple regression analysis is the study of how a dependent variable Y is related to two or more independent variables. In the general case, we will use p to denote the number of independent variables.

Regression model and regression equation

The concepts of a regression model and a regression equation introduced in the preceding chapter are applicable in the multiple regression case. The equation that describes how the dependent variable Y is related to the independent variables $X_1, X_2, \ldots X_p$ and an error term is called the **multiple regression model**. We begin with the assumption that the multiple regression model takes the following form.

Stastics in Practice

Jura

Jura is a large island (380 sq km) off the South West of Scotland famous for its malt whisky and the large deer population that wander the quartz mountains ('the Paps') that dominate the landscape. With a population of a mere 461 it has one of the lowest population densities of any place in the UK. Currently Jura is only accessible via the adjoining island, Islay, which has three ferry services a day – crossings taking about two hours. However, because Jura is only four miles from the mainland it has been suggested that a direct car ferry taking less than half an hour would be preferable and more economical than existing provisions.

In exploring the case for an alternative service, Riddington (1994) arrives at a number of alternative mathematical formulations that essentially reduce to multiple regression analysis. In particular using historical data that also encompasses other inner Hebridean islands of Arran, Bute, Mull and Skye he obtains the estimated binary logistic regression model:

$$\text{Log}_e \frac{Q_{1it}}{Q_{2it}} = 6.48 - 0.89\frac{P_{1it}}{P_{2it}} + 0.129\frac{F_{1it}}{F_{2it}} - 6.18\frac{J_{1it}}{J_{2it}}$$

where

Q_{1it}/Q_{2it} it is the number of cars travelling by route 1 relative to the number travelling by route 2 to island *i* in year *t*

P_{1it}/P_{2it} is the relative price between route 1 and route 2 to *i* in year *t*

F_{1it}/F_{2it} is the relative frequency between route 1 and route 2 to *i* in year *t*

J_{1it}/J_{2it} is the relative journey time between route 1and route 2 to *i* in year *t*

Based on appropriate economic assumptions he estimates from this that some 132 000 passengers and 38 000 cars would use the new service each year rising over time. Initially this would yield a revenue of £426 000. Allowing for annual running costs of £322 000, the resultant gross profit would therefore be of the order of £100 000.

Isle of Jura off the west coast of Scotland. The mountains in the distance are the distinctive Paps of Jura. © Martin McCarthy.

Source: Riddington, Geoff (1996) How many for the ferry boat? *OR Insight* Vol 9 Issue 2 pp 26–32.

Multiple regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \qquad \textbf{(15.1)}$$

where $X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p$.

In the multiple regression model, $\beta_0, \beta_1, \ldots, \beta_p$ are the parameters and ε (the Greek letter epsilon) is a random variable. A close examination of this model reveals that Y is a linear function of $x_1, x_2, \ldots, x_p$ (the $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots\cdots + \beta_p x_p$ part) plus an error term ε. The error term accounts for the variability in Y that cannot be explained by the linear effect of the p independent variables.

In Section 15.4 we will discuss the assumptions for the multiple regression model and ε. One of the assumptions is that the mean or expected value of ε is zero. A consequence of this assumption is that the mean or expected value of Y, denoted $E(Y)$, is equal to $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots\cdots + \beta_p x_p$. The equation that describes how the mean value of Y is related to $x_1, x_2, \ldots x_p$ is called the **multiple regression equation**

Multiple regression equation

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad \textbf{(15.2)}$$

Estimated multiple regression equation

If the values of $\beta_0, \beta_1, \ldots, \beta_p$ were known, equation (15.2) could be used to compute the mean value of Y at given values of $x_1, x_2, \ldots x_p$. Unfortunately, these parameter values will not, in general, be known and must be estimated from sample data. A simple random sample is used to compute sample statistics $b_0, b_1, \ldots, b_p$ that are used as the point estimators of the parameters $\beta_0, \beta_1, \ldots, \beta_p$. These sample statistics provide the following **estimated multiple regression equation**.

Estimated multiple regression equation

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p \qquad \textbf{(15.3)}$$

where

$b_0, b_1, \ldots, b_p$ are the estimates of $\beta_0, \beta_1, \ldots, \beta_p$
$\hat{y}$ = estimated value of the dependent variable

The estimation process for multiple regression is shown in Figure 15.1.

15.2 Least squares method

In Chapter 14, we used the **least squares method** to develop the estimated regression equation that best approximated the straight-line relationship between the dependent and independent variables. This same approach is used to develop the estimated multiple regression equation. The least squares criterion is restated as follows.

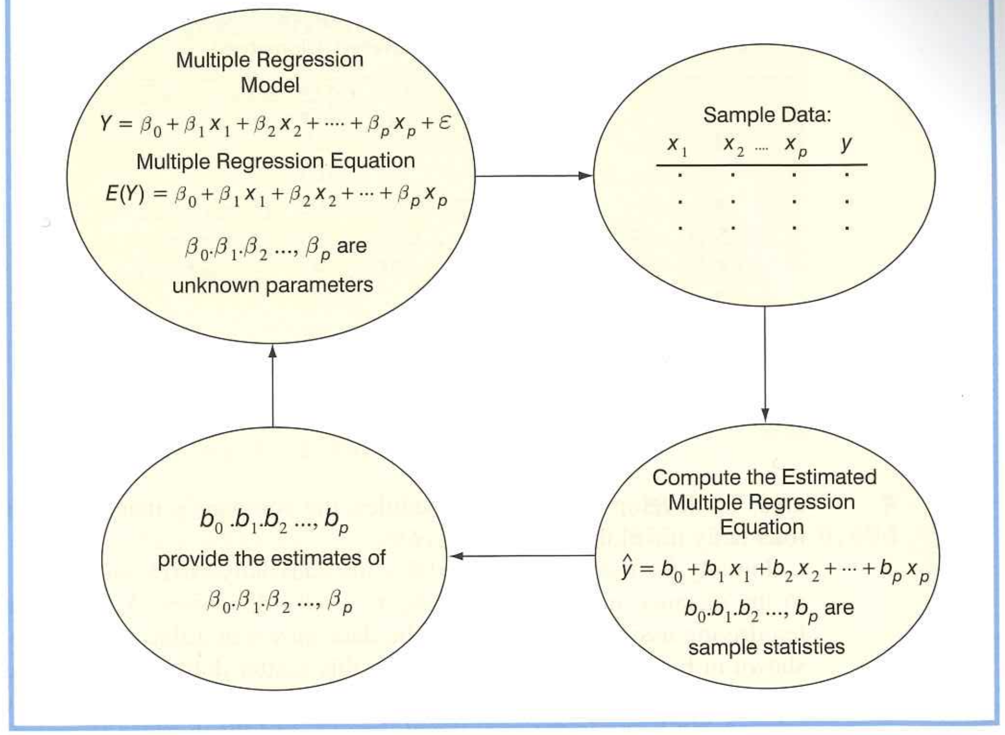
Least squares criterion

$$\min \Sigma (y_i - \hat{y}_i)^2 \qquad \textbf{(15.4)}$$

where

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

Figure 15.1 The estimation process for multiple regression


Multiple Regression Model
$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$
Multiple Regression Equation
$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$
$\beta_0, \beta_1, \beta_2 \ldots, \beta_p$ are unknown parameters
Sample Data:
x_1 x_2 ... x_p y
Compute the Estimated Multiple Regression Equation
$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$
$b_0, b_1, b_2 \ldots, b_p$ are sample statistics
$b_0, b_1, b_2 \ldots, b_p$
provide the estimates of
$\beta_0, \beta_1, \beta_2 \ldots, \beta_p$

The estimated values of the dependent variable are computed by using the estimated multiple regression equation,

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

As expression (15.4) shows, the least squares method uses sample data to provide the values of $b_0, b_1, \ldots, b_p$ that make the sum of squared residuals {the deviations between the observed values of the dependent variable (y_i) and the estimated values of the dependent variable $\hat{y}_i$} a minimum.

In Chapter 14 we presented formulae for computing the lea st squares estimators b_0 and b_1 for the estimated simple linear regression equation $\hat{y} = b_0 + b_1 x$. With relatively small data sets, we were able to use those formulae to compute b_0 and b_1 by manual calculations. In multiple regression, however, the presentation of the formulae for the regression coefficients $b_0, b_1, \ldots, b_p$ involves the use of matrix algebra and is beyond the scope of this text. Therefore, in presenting multiple regression, we focus on how computer software packages can be used to obtain the estimated regression equation and other information. The emphasis will be on how to interpret the computer output rather than on how to make the multiple regression computations.

An example: Eurodistributor Company

As an illustration of multiple regression analysis, we will consider a problem faced by the Eurodistributor Company, an independent distribution company in the Netherlands. A major portion of Eurodistributor's business involves deliveries throughout its local

Table 15.1 Preliminary data for Eurodistributor

Driving assignment	X_1 = Distance travelled (kilometres)	Y = Travel time (hours)
1	100	9.3
2	50	4.8
3	100	8.9
4	100	6.5
5	50	4.2
6	80	6.2
7	75	7.4
8	65	6.0
9	90	7.6
10	90	6.1

area. To develop better work schedules, the company's managers want to estimate the total daily travel time for their drivers.

Initially the managers believed that the total daily travel time would be closely related to the distance travelled in making the daily deliveries. A simple random sample of ten driving assignments provided the data shown in Table 15.1 and the scatter diagram shown in Figure 15.2. After reviewing this scatter diagram, the managers hypothesized that the simple linear regression model $Y = \beta_0 + \beta_1 x_1 + \varepsilon$ could be used to describe the relationship between the total travel time (Y) and the distance travelled (X_1). To estimate the parameters β_0 and β_1, the least squares method was used to develop the estimated regression equation.

$$\hat{y} = b_0 + b_1 x_1 \qquad (15.5)$$

In Figure 15.3, we show the MINITAB computer output from applying simple linear regression to the data in Table 15.1. The estimated regression equation is

$$\hat{y} = 1.27 + 0.0678x_1$$

EURODIS-TRIBUTOR

Figure 15.2 Scatter diagram of preliminary data for Eurodistributor

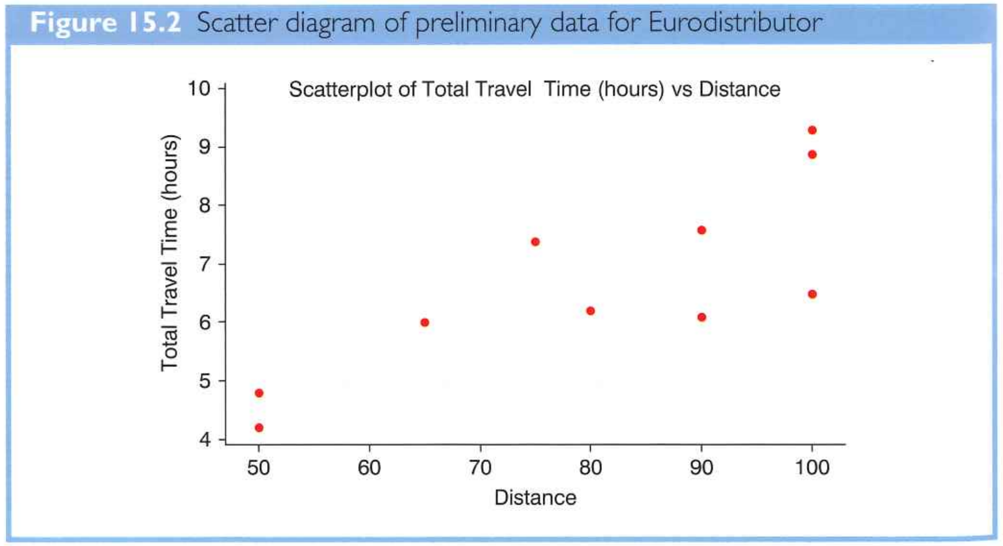
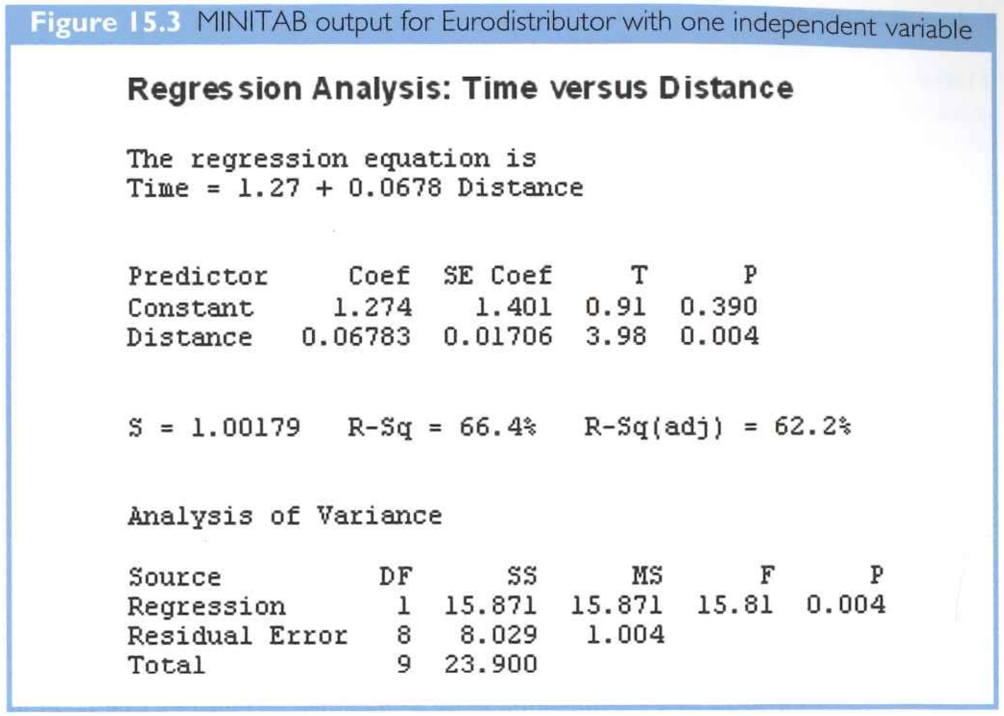


Figure 15.3 MINITAB output for Eurodistributor with one independent variable



Regression Analysis: Time versus Distance

```
The regression equation is
Time = 1.27 + 0.0678 Distance


Predictor      Coef   SE Coef     T      P
Constant      1.274     1.401  0.91  0.390
Distance    0.06783   0.01706  3.98  0.004


S = 1.00179   R-Sq = 66.4%   R-Sq(adj) = 62.2%


Analysis of Variance

Source          DF      SS      MS      F      P
Regression       1  15.871  15.871  15.81  0.004
Residual Error   8   8.029   1.004
Total            9  23.900
```

At the 0.05 level of significance, the F value of 15.81 and its corresponding p-value of 0.004 indicate that the relationship is significant; that is, we can reject H_0: $\beta_1 = 0$ because the p-value is less than $\alpha = 0.05$. Thus, we can conclude that the relationship between the total travel time and the distance travelled is significant; longer travel times are associated with more distance. With a coefficient of determination (expressed as a percentage) of R-sq = 66.4 per cent, we see that 66.4 per cent of the variability in travel time can be explained by the linear effect of the distance travelled. This finding is fairly good, but the managers might want to consider adding a second independent variable to explain some of the remaining variability in the dependent variable.

In attempting to identify another independent variable, the managers felt that the number of deliveries could also contribute to the total travel time. The Eurodistributor data, with the number of deliveries added, are shown in Table 15.2. The MINITAB computer solution with both distance (X_1) and number of deliveries (X_2) as independent variables is shown in Figure 15.4. The estimated regression equation is

$$\hat{y} = -0.869 + 0.0611x_1 + 0.923x_2 \qquad (15.6)$$

In the next section we will discuss the use of the coefficient of multiple determination in measuring how good a fit is provided by this estimated regression equation. Before doing so, let us examine more carefully the values of $b_1 = 0.0611$ and $b_2 = 0.923$ in equation (15.6).

Note on interpretation of coefficients

One observation can be made at this point about the relationship between the estimated regression equation with only the distance as an independent variable and the equation that includes the number of deliveries as a second independent variable. The value of b_1

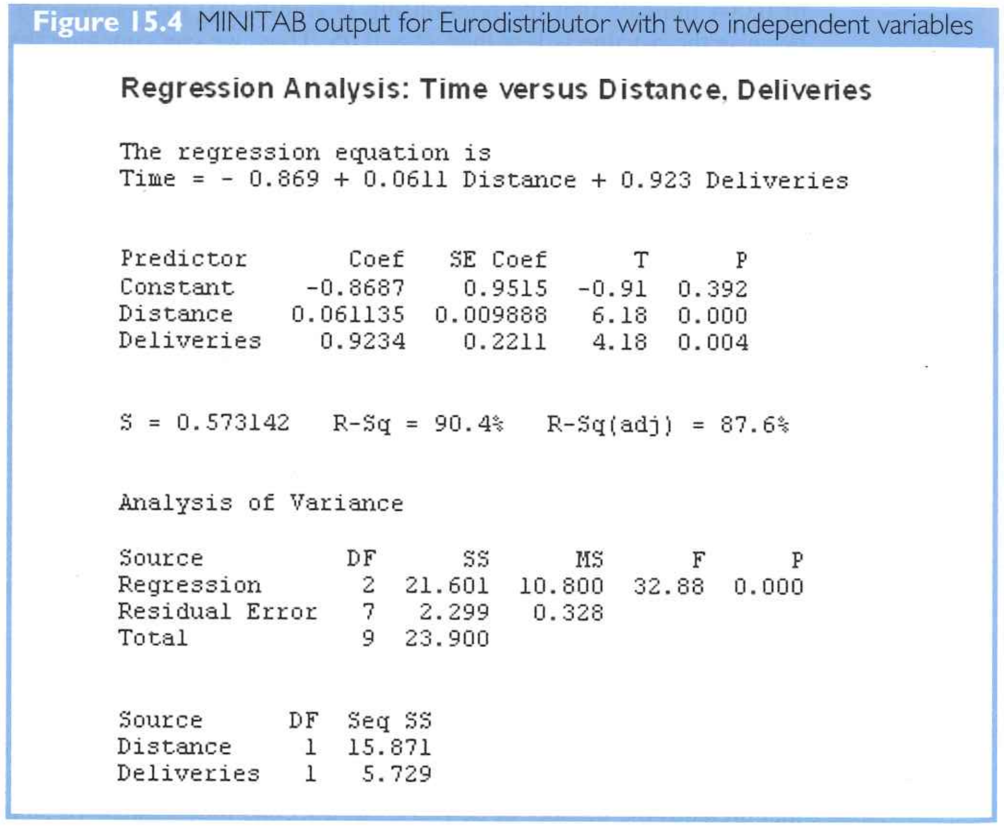
Table 15.2 Data for Eurodistributor with distance (X_1) and number of deliveries (X_2) as the independent variables

Driving assignment	X_1 = Distance travelled (kilometres)	X_2 = Number of deliveries	Y = Travel time (hours)
1	100	4	9.3
2	50	3	4.8
3	100	4	8.9
4	100	2	6.5
5	50	2	4.2
6	80	2	6.2
7	75	3	7.4
8	65	4	6.0
9	90	3	7.6
10	90	2	6.1

is not the same in both cases. In simple linear regression, we interpret b_1 as an estimate of the change in Y for a one-unit change in the independent variable. In multiple regression analysis, this interpretation must be modified somewhat. That is, in multiple regression analysis, we interpret each regression coefficient as follows: b_i represents an estimate of the change in Y corresponding to a one-unit change in X_i when all other independent variables are held constant.

In the Eurodistributor example involving two independent variables, $b_1 = 0.0611$. Thus, 0.0611 hours is an estimate of the expected increase in travel time corresponding to an increase of one kilometre in the distance travelled when the number of deliveries is held constant. Similarly, because $b_2 = 0.923$, an estimate of the expected increase in travel time corresponding to an increase of one delivery when the distance travelled is held constant is 0.923 hours.

Figure 15.4 MINITAB output for Eurodistributor with two independent variables



```
Regression Analysis: Time versus Distance, Deliveries

The regression equation is
Time = - 0.869 + 0.0611 Distance + 0.923 Deliveries


Predictor        Coef    SE Coef      T      P
Constant      -0.8687     0.9515  -0.91  0.392
Distance     0.061135   0.009888   6.18  0.000
Deliveries     0.9234     0.2211   4.18  0.004


S = 0.573142   R-Sq = 90.4%   R-Sq(adj) = 87.6%


Analysis of Variance

Source          DF      SS      MS      F      P
Regression       2  21.601  10.800  32.88  0.000
Residual Error   7   2.299   0.328
Total            9  23.900


Source      DF  Seq SS
Distance     1  15.871
Deliveries   1   5.729
```

Exercises

Note to student: The exercises involving data in this and subsequent sections were designed to be solved using a computer software package.

Methods

1 The estimated regression equation for a model involving two independent variables and ten observations follows.

$$\hat{y} = 29.1270 + 0.5906x_1 + 0.4980x_2$$

a. Interpret b_1 and b_2 in this estimated regression equation.
b. Estimate Y when X_1 = 180 and X_2 = 310.

2 Consider the following data for a dependent variable Y and two independent variables, X_1 and X_2.

EXER2

x_1	x_2	y
30	12	94
47	10	108
25	17	112
51	16	178
40	5	94
51	19	175
74	7	170
36	12	117
59	13	142
76	16	211

a. Develop an estimated regression equation relating Y to X_1. Estimate Y if X_1 = 45.
b. Develop an estimated regression equation relating Y to X_2. Estimate Y if X_2 = 15.
c. Develop an estimated regression equation relating Y to X_1 and X_2. Estimate Y if X_1 = 45 and X_2 = 15.

3 In a regression analysis involving 30 observations, the following estimated regression equation was obtained.

$$\hat{y} = 17.6 + 0\ 3.8x_1 - 2.3x_2 + 7.6x_3 + 2.7x_4$$

a. Interpret b_1, b_2, b_3, and b_4 in this estimated regression equation.
b. Estimate Y when X_1 = 10, X_2 = 5, X_3 = 1, and X_4 =2.

Applications

4 A shoe store developed the following estimated regression equation relating sales to inventory investment and advertising expenditures.

$$\hat{y} = 25 + 10x_1 + 8x_2$$

where

$$X_1 = \text{inventory investment (€000s)}$$
$$X_2 = \text{advertising expenditures (€000s)}$$
$$Y = \text{sales (€000s)}$$

a. Estimate sales resulting from a €15 000 investment in inventory and an advertising budget of €10 000.

b. Interpret b_1 and b_2 in this estimated regression equation.

TOULON

5 The owner of Toulon Theatres would like to estimate weekly gross revenue as a function of advertising expenditures. Historical data for a sample of eight weeks follow.

Weekly gross revenue (€000s)	Television advertising (€000s)	Newspaper advertising (€000s)
96	5.0	1.5
90	2.0	2.0
95	4.0	1.5
92	2.5	2.5
95	3.0	3.3
94	3.5	2.3
94	2.5	4.2
94	3.0	2.5

a. Develop an estimated regression equation with the amount of television advertising as the independent variable.

b. Develop an estimated regression equation with both television advertizing and newspaper advertising as the independent variables.

c. Is the estimated regression equation coefficient for television advertising expenditures the same in part (a) and in part (b)? Interpret the coefficient in each case.

d. What is the estimate of the weekly gross revenue for a week when €3500 is spent on television advertising and €1800 is spent on newspaper advertising?

6 The following table gives the annual return, the safety rating (0 = riskiest, 10 = safest), and the annual expense ratio for 20 foreign funds.

FORFUNDS

	Annual safety rating	Expense ratio (%)	Annual return (%)
Accessor Int'l Equity 'Adv'	7.1	1.59	49
Aetna 'I' International	7.2	1.35	52
Amer Century Int'l Discovery 'Inv'	6.8	1.68	89
Columbia International Stock	7.1	1.56	58
Concert Inv 'A' Int'l Equity	6.2	2.16	131
Dreyfus Founders Int'l Equity 'F'	7.4	1.80	59
Driehaus International Growth	6.5	1.88	99
Excelsior 'Inst' Int'l Equity	7.0	0.90	53
Julius Baer International Equity	6.9	1.79	77
Marshall International Stock 'Y'	7.2	1.49	54
MassMutual Int'l Equity 'S'	7.1	1.05	57
Morgan Grenfell Int'l Sm Cap 'Inst'	7.7	1.25	61
New England 'A' Int'l Equity	7.0	1.83	88
Pilgrim Int'l Small Cap 'A'	7.0	1.94	122
Republic International Equity	7.2	1.09	71
Sit International Growth	6.9	1.50	51
Smith Barney 'A' Int'l Equity	7.0	1.28	60
State St Research 'S' Int'l Equity	7.1	1.65	50
Strong International Stock	6.5	1.61	93
Vontobel International Equity	7.0	1.50	47

a. Develop an estimated regression equation relating the annual return to the safety rating and the annual expense ratio.

b. Estimate the annual return for a firm that has a safety rating of 7.5 and annual expense ratio of 2.

15.3 Multiple coefficient of determination

In simple linear regression we showed that the total sum of squares can be partitioned into two components: the sum of squares due to regression and the sum of squares due to error. The same procedure applies to the sum of squares in multiple regression.

Relationship among SST, SSR and SSE

$$\text{SST} = \text{SSR} + \text{SSE} \qquad (15.7)$$

where

$$\text{SST} = \text{total sum of squares} = \Sigma(y_i - \bar{y})^2$$
$$\text{SSR} = \text{sum of squares due to regression} = \Sigma(\hat{y}_i - \bar{y})^2$$
$$\text{SSE} = \text{sum of squares due to error} = \Sigma(y_i - \hat{y}_i)^2$$

Because of the computational difficulty in computing the three sums of squares, we rely on computer packages to determine those values. The analysis of variance part of the MINITAB output in Figure 15.4 shows the three values for the Eurodistributor problem with two independent variables: SST = 23.900, = SSR 21.601 and SSE = 2.299. With only one independent variable (distance travelled), the MINITAB output in Figure 15.3 shows that SST = 23.900, SSR = 15.871 and SSE = 8.029. The value of SST is the same in both cases because it does not depend on $\hat{y}$ but SSR increases and SSE decreases when a second independent variable (number of deliveries) is added. The implication is that the estimated multiple regression equation provides a better fit for the observed data.

In Chapter 14, we used the coefficient of determination, $R^2 = \text{SSR/SST}$, to measure the goodness of fit for the estimated regression equation. The same concept applies to multiple regression. The term **multiple coefficient of determination** indicates that we are measuring the goodness of fit for the estimated multiple regression equation. The multiple coefficient of determination, denoted R^2, is computed as follows.

Multiple coefficient of determination

$$R^2 = \frac{\text{SSR}}{\text{SST}} \tag{15.8}$$

The multiple coefficient of determination can be interpreted as the proportion of the variability in the dependent variable that can be explained by the estimated multiple regression equation. Hence, when multiplied by 100, it can be interpreted as the percentage of the variability in Y that can be explained by the estimated regression equation.

In the two-independent-variable Eurodistributor example, with SSR = 21.601 and SST = 23.900, we have

$$R^2 = \frac{21.601}{23.900} = 0.904$$

Therefore, 90.4 per cent of the variability in travel time Y is explained by the estimated multiple regression equation with distance and number of deliveries as the independent variables. In Figure 15.4, we see that the multiple coefficient of determination is also provided by the MINITAB output; it is denoted by R-sq = 90.4 per cent.

Figure 15.3 shows that the R-sq value for the estimated regression equation with only one independent variable, distance travelled (X_1), is 66.4 per cent. Thus, the percentage of the variability in travel times that is explained by the estimated regression equation increases from 66.4 per cent to 90.4 per cent when number of deliveries is added as a second independent variable. In general, R^2 increases as independent variables are added to the model.

Many analysts prefer adjusting R^2 for the number of independent variables to avoid overestimating the impact of adding an independent variable on the amount of variability explained by the estimated regression equation. With n denoting the number of observations and p denoting the number of independent variables, the **adjusted multiple coefficient of determination** is computed as follows.

Adjusted multiple coefficient of determination

$$\text{adj } R^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1} \tag{15.9}$$

For the Eurodistributor example with $n = 10$ and $p = 2$, we have

$$\text{adj } R^2 = 1 - (1 - 0.904)\frac{10-1}{10-2-1} = 0.88$$

Therefore, after adjusting for the two independent variables, we have an adjusted multiple coefficient of determination of 0.88. This value, allowing for rounding, corresponds with the value in the MINITAB output in Figure 15.4 of R-sq(adj) = 87.6 per cent.

Exercises

Methods

7 In exercise 1, the following estimated regression equation based on ten observations was presented.

$$\hat{y} = 29.1270 + 0.5906x_1 + 0.4980x_2$$

The values of SST and SSR are 6724.125 and 6216.375, respectively.

a. Find SSE.
b. Compute R^2.
c. Compute *Adj* R^2.
d. Comment on the goodness of fit.

EXER2

8 In exercise 2, ten observations were provided for a dependent variable Y and two independent variables X_1 and X_2; for these data SST = 15 182.9, and SSR = 14 052.2.

a. Compute R^2.
b. Compute *Adj* R^2.
c. Does the estimated regression equation explain a large amount of the variability in the data? Explain.

9 In exercise 3, the following estimated regression equation based on 30 observations was presented.

$$\hat{y} = 17.6 + 3.8x_1 - 2.3x_2 + 7.6x_3 + 2.7x_4$$

The values of SST and SSR are 1805 and 1760, respectively.

a. Compute R^2.
b. Compute *Adj* R^2.
c. Comment on the goodness of fit.

Applications

10 In exercise 4, the following estimated regression equation relating sales to inventory investment and advertising expenditures was given.

$$\hat{y} = 25 + 10x_1 + 8x_2$$

The data used to develop the model came from a survey of ten stores; for those data, SST = 16 000 and SSR = 12 000.

a. For the estimated regression equation given, compute R^2.
b. Compute *Adj* R^2.
c. Does the model appear to explain a large amount of variability in the data? Explain.

TOULON

11 In exercise 5, the owner of Toulon Theatres used multiple regression analysis to predict gross revenue (Y) as a function of television advertising (X_1) and newspaper advertising (X_2). The estimated regression equation was

$$\hat{y} = 83.2 + 2.29x_1 + 1.30x_2$$

The computer solution provided SST = 25.5 and SSR = 23.435.

a. Compute and interpret R^2 and *Adj* R^2.
b. When television advertising was the only independent variable, R^2 = 0.653 and *Adj* R^2 = 0.595. Do you prefer the multiple regression results? Explain.

15.4 Model assumptions

In Section 15.1 we introduced the following multiple regression model.

Multiple regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \qquad \textbf{(15.10)}$$

The assumptions about the error term ε in the multiple regression model parallel those for the simple linear regression model.

Assumptions about the error term in the multiple regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta x_2 + \cdots + \beta_p x_p + \varepsilon$$

1 The error ε is a random variable with mean or expected value of zero; that is, $E(\varepsilon) = 0$.
Implication: For given values of $X_1, X_2, \ldots X_p$, the expected, or average, value of Y is given by

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad \textbf{(15.11)}$$

Equation (15.11) is the multiple regression equation we introduced in Section 15.1. In this equation, $E(Y)$ represents the average of all possible values of Y that might occur for the given values of $X_1, X_2, \ldots, X_p$.

2 The variance of ε is denoted by σ^2 and is the same for all values of the independent variables $X_1, X_2, \ldots, X_p$.
Implication: The variance of Y about the regression line equals σ^2 and is the same for all values of $X_1, X_2, \ldots, X_p$.

3 The values of ε are independent.
Implication: The size of the error for a particular set of values for the independent variables is not related to the size of the error for any other set of values.

4 The error ε is a normally distributed random variable reflecting the deviation between the Y value and the expected value of Y given by $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$.
Implication: Because $\beta_0, \beta_1, \ldots, \beta_p$ are constants for the given values of $x_1, x_2, \ldots x_p$, the dependent variable Y is also a normally distributed random variable.

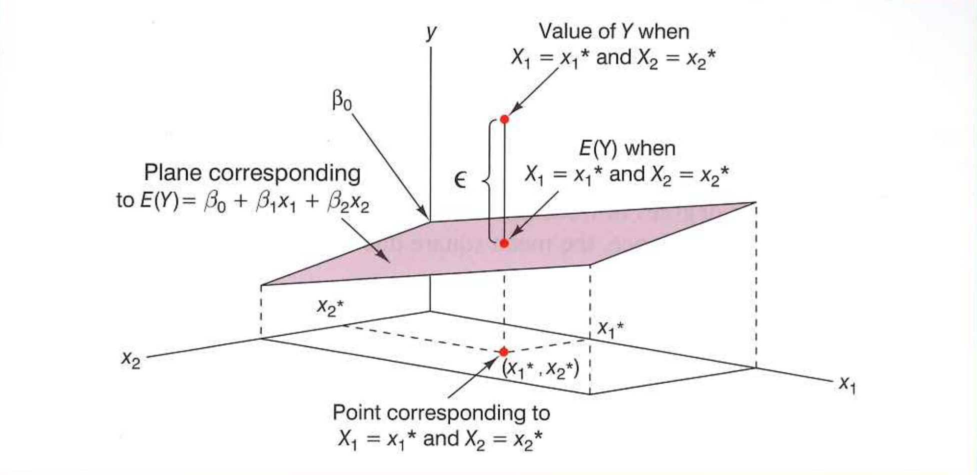
To obtain more insight about the form of the relationship given by equation (15.11), consider the following two-independent-variable multiple regression equation.

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

The graph of this equation is a plane in three-dimensional space. Figure 15.5 provides an example of such a graph. Note that the value of ε shown is the difference between the actual Y value and the expected value of y, $E(Y)$, when $X_1 = x_1^*$ and $X_2 = x_2^*$.

In regression analysis, the term *response variable* is often used in place of the term *dependent variable*. Furthermore, since the multiple regression equation generates a plane or surface, its graph is called a *response surface*.

Figure 15.5 Graph of the regression equation for multiple regression analysis with two independent variables



15.5 Testing for significance

In this section we show how to conduct significance tests for a multiple regression relationship.

The significance tests we used in simple linear regression were a t test and an F test. In simple linear regression, both tests provide the same conclusion; that is, if the null hypothesis is rejected, we conclude that the slope parameter $\beta_1 \neq 0$. In multiple regression, the t test and the F test have different purposes.

1 The F test is used to determine whether a significant relationship exists between the dependent variable and the set of all the independent variables; we will refer to the F test as the test for *overall significance*.

2 If the F test shows an overall significance, the t test is used to determine whether each of the individual independent variables is significant. A separate t test is conducted for each of the independent variables in the model; we refer to each of these t tests as a test for *individual significance*.

In the material that follows, we will explain the F test and the t test and apply each to the Eurodistributor Company example.

F test

Given the multiple regression model defined in (15.1)

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots\cdots + \beta_p x_p + \varepsilon$$

the hypotheses for the F test can be written as follows:

$$H_0\colon \beta_1 = \beta_2 = \ldots\ldots = \beta_p = 0$$
$$H_1\colon \text{One or more of the parameters is not equal to zero}$$

If H_0 is rejected, the test gives us sufficient statistical evidence to conclude that one or more of the parameters is not equal to zero and that the overall relationship between Y and the set of independent variables $X_1, X_2, \ldots X_p$ is significant. However, if H_0 cannot be rejected, we deduce there is not sufficient evidence to conclude that a significant relationship is present.

Before confirming the steps involved in performing the F test, it might be helpful if we first review the concept of *mean square*. A mean square is a sum of squares divided by its corresponding degrees of freedom. In the multiple regression case, the total sum of squares has $n - 1$ degrees of freedom, the sum of squares due to regression (SSR) has p degrees of freedom, and the sum of squares due to error has $n - p - 1$ degrees of freedom. Hence, the mean square due to regression (MSR) is

Mean square regression

$$\text{MSR} = \frac{\text{SSR}}{p} \qquad (15.12)$$

and

Mean square error

$$\text{MSE} = s^2 = \frac{\text{SSE}}{n - p - 1} \qquad (15.13)$$

As has already been acknowledged in Chapter 14, MSE provides an unbiased estimate of σ^2, the variance of the error term ε. If H_0: $\beta_1 = \beta_2 = \ldots\ldots = \beta_p = 0$ is true, MSR also provides an unbiased estimate of σ^2, and the value of MSR/MSE should be close to 1. However, if H_0 is false, MSR overestimates σ^2 and the value of MSR/MSE becomes larger. To determine how large the value of MSR/MSE must be to reject H_0, we make use of the fact that if H_0 is true and the assumptions about the multiple regression model are valid, the sampling distribution of MSR/MSE is an F distribution with p degrees of freedom in the numerator and $n - p - 1$ in the denominator. A summary of the F test for significance in multiple regression follows.

***F* test for overall significance**

$$H_0\text{: } \beta_1 = \beta_2 = \ldots = \beta_p = 0$$
H_1: One or more of the parameters is not equal to zero

Test statistic

$$F = \frac{\text{MSR}}{\text{MSE}} \qquad (15.14)$$

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 p degrees of freedom in the numerator and $n - p - 1$ degrees of freedom in the denominator.

Applying the F test to the Eurodistributor Company multiple regression problem with two independent variables, the hypotheses can be written as follows.

$$H_0\text{: } \beta_1 = \beta_2 = 0$$
$$H_1\text{: } \beta_1 \text{ and/or } \beta_2 \text{ is not equal to zero}$$

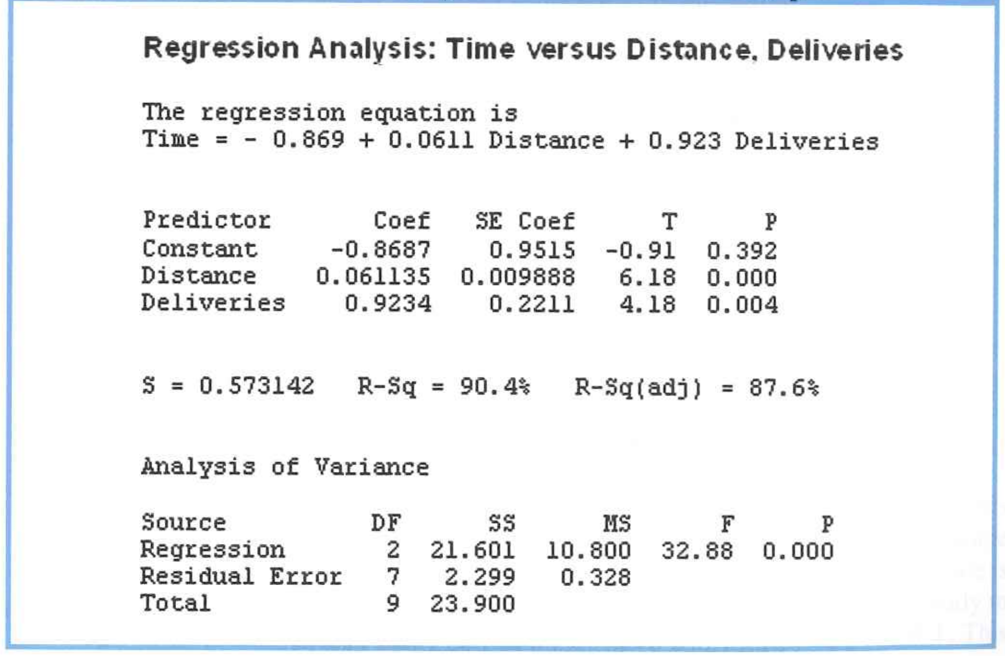
Figure 15.6 shows the MINITAB output for the multiple regression model with distance (X_1) and number of deliveries (X_2) as the two independent variables. In the analysis of variance part of the output, we see that MSR = 10.8 and MSE = 0.328. Using equation (15.14), we obtain the test statistic.

$$F = \frac{10.8}{0.328} = 32.9$$

Note that the F value on the MINITAB output is $F = 32.88$; the value we calculated differs because we used rounded values for MSR and MSE in the calculation. Using $\alpha = 0.01$, the p-value = 0.000 in the last column of the analysis of variance table (Figure 15.6) indicates that we can reject H_0: $\beta_1 = \beta_2 = 0$ because the p-value is less than $\alpha = 0.01$. Alternatively, Table 4 of Appendix B shows that with two degrees of freedom in the numerator and seven degrees of freedom in the denominator, $F_{0.01} = 9.55$. With $32.9 > 9.55$, we reject H_0: $\beta_1 = \beta_2 = 0$ and conclude that a significant relationship is present between travel time Y and the two independent variables, distance and number of deliveries.

As noted previously, the mean square error provides an unbiased estimate of σ^2, the variance of the error term ε. Referring to Figure 15.6, we see that the estimate of σ^2 is MSE = 0.328. The square root of MSE is the estimate of the standard deviation of the error term. As defined in Section 14.5, this standard deviation is called the standard error

Figure 15.6 MINITAB output for Eurodistributor with two independent variables, distance (X_1) and number of deliveries (X_2)



```
Regression Analysis: Time versus Distance, Deliveries

The regression equation is
Time = - 0.869 + 0.0611 Distance + 0.923 Deliveries


Predictor        Coef    SE Coef       T      P
Constant      -0.8687     0.9515   -0.91  0.392
Distance     0.061135   0.009888    6.18  0.000
Deliveries     0.9234     0.2211    4.18  0.004


S = 0.573142   R-Sq = 90.4%   R-Sq(adj) = 87.6%


Analysis of Variance

Source          DF      SS      MS      F      P
Regression       2  21.601  10.800  32.88  0.000
Residual Error   7   2.299   0.328
Total            9  23.900
```

Table 15.3 ANOVA table for a multiple regression model with p independent variables

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

of the estimate and is denoted s. Hence, we have $s = \sqrt{\text{MSE}} = \sqrt{0.328} = 0.573$. Note that the value of the standard error of the estimate appears in the MINITAB output in Figure 15.6.

Table 15.3 is the general analysis of variance (ANOVA) table that provides the F test results for a multiple regression model. The value of the F test statistic appears in the last column and can be compared to F_α with p degrees of freedom in the numerator and $n - p - 1$ degrees of freedom in the denominator to make the hypothesis test conclusion.

By reviewing the MINITAB output for Eurodistributor Company in Figure 15.6, we see that MINITAB's analysis of variance table contains this information. In addition, MINITAB provides the p-value corresponding to the F test statistic.

t test

If the F test shows that the multiple regression relationship is significant, a t test can be conducted to determine the significance of each of the individual parameters. The t test for individual significance follows.

t test for individual significance

For any parameter β_i

$$H_0\colon \beta_i = 0$$
$$H_1\colon \beta_i \neq 0$$

Test statistic

$$t = \frac{b_i}{s_{b_i}}$$

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 $n - p - 1$ degrees of freedom.

In the test statistic, s_{b_i} is the estimate of the standard deviation of b_i. The value of s_{b_i} will be provided by the computer software package.

Let us conduct the t test for the Eurodistributor regression problem. Refer to the section of Figure 15.6 that shows the MINITAB output for the t-ratio calculations. Values of b_1, b_2, s_{b_1} and s_{b_2} are as follows.

$$b_1 = 0.061135 \qquad s_{b_1} = 0.009888$$
$$b_2 = 0.9234 \qquad s_{b_2} = 0.2211$$

Using equation (15.15), we obtain the test statistic for the hypotheses involving parameters β_1 and β_2.

$$t = 0.061135/0.009888 = 6.18$$
$$t = 0.9234/0.2211 = 4.18$$

Note that both of these t-ratio values and the corresponding p-values are provided by the MINITAB output in Figure 15.6. Using $\alpha = 0.01$, the p-values of 0.000 and 0.004 from the MINITAB output indicate that we can reject $H_0\colon \beta_1 = 0$ and $H_0\colon \beta_2 = 0$. Hence, both parameters are statistically significant. Alternatively, Table 2 of Appendix B shows that with $n - p - 1 = 10 - 2 - 1 = 7$ degrees of freedom, $t_{0.005} = 3.499$. With $6.18 > 3.499$, we reject $H_0\colon \beta_1 = 0$. Similarly, with $4.18 > 3.499$, we reject $H_0\colon \beta_2 = 0$.

Multicollinearity

In multiple regression analysis, **multicollinearity** refers to the correlation among the independent variables. We used the term independent variable in regression analysis to refer to any variable being used to predict or explain the value of the dependent variable. The term does not mean, however, that the independent variables themselves are independent in any statistical sense. On the contrary, most independent variables in a multiple regression problem are correlated to some degree with one another. For example, in the Eurodistributor example involving the two independent variables X_1 (distance) and X_2 (number of deliveries), we could treat the distance as the dependent variable and the number of deliveries as the independent variable to determine whether those two variables are themselves related. We could then compute the sample correlation coefficient to determine the extent to which the variables are related. Doing so yields

$$\text{Pearson correlation of Distance and Deliveries} = 0.162$$

which suggests only a small degree of linear association exists between the two variables. The implication from this would be that multicollinearity is not a problem for the data. If however the association had been more pronounced the resultant multicollinearity might seriously have jeopardized the estimation of the model.

To provide a better perspective of the potential problems of multicollinearity, let us consider a modification of the Eurodistributor example. Instead of X_2 being the number of deliveries, let X_2 denote the number of litres of petrol consumed. Clearly, X_1 (the distance) and X_2 are related; that is, we know that the number of litres of petrol used depends on the distance travelled. Hence, we would conclude logically that X_1 and X_2 are highly correlated independent variables.

Assume that we obtain the equation $\hat{y} = b_0 + b_1x_1 + b_2x_2$ and find that the F test shows the relationship to be significant. Then suppose we conduct a t test on β_1 to determine whether $\beta_1 = 0$, and we cannot reject $H_0\colon \beta_1 = 0$. Does this result mean that travel time is not related to distance? Not necessarily. What it probably means is that with X_2 already in the model, X_1 does not make a significant contribution to determining the value of Y. This interpretation makes sense in our example; if we know the amount of petrol consumed, we do not gain much additional information useful in predicting Y by knowing the distance. Similarly, a t test might lead us to conclude $\beta_2 = 0$ on the grounds that, with X_1 in the model, knowledge of the amount of petrol consumed does not add much.

One useful way of detecting multicollinearity is to calculate the **variance inflation factor** (VIF) for each independent variable (X_j) in the model. The VIF is defined as

Variance inflation factor

$$\text{VIF}(X_j) \quad \frac{1}{1 - R_j^2} \qquad \textbf{(15.16)}$$

where R_j^2 is the coefficient of determination obtained when X_j ($j = 1, 2, \ldots, p$) is regressed on all remaining independent variables in the model. If X_j is not correlated with other predictors $R_j^2 \approx 0$ and VIF ≈ 1. Correspondingly if R_j^2 is close to 1 the VIF will be very large. Typically VIF values of ten or more are regarded as problematic.

For the Eurodistributor data, the VIF for X_1 (and also X_2 by symmetry) would be

$$\text{VIF}(X_j) = \frac{1}{1 - 0.162^2} = 1.027$$

signifying as before there is no problem with multicollinearity.

To summarize, for t tests associated with testing for the significance of individual parameters, the difficulty caused by multicollinearity is that it is possible to conclude that none of the individual parameters are significantly different from zero when an F test on the overall multiple regression equation indicates there is a significant relationship. This problem is avoided however when little correlation among the independent variables exists.

If possible, every attempt should be made to avoid including independent variables that are highly correlated. In practice, however, strict adherence to this policy is not always possible. When decision-makers have reason to believe substantial multicollinearity is present, they must realize that separating the effects of the individual independent variables on the dependent variable is difficult.

Exercises

Methods

12 In exercise 1, the following estimated regression equation based on ten observations was presented.

$$\hat{y} = 29.1270 + 0.5906x_1 + 0.4980x_2$$

Here SST = 6724.125, SSR = 6216.375, $s_{b_1} = 0.0813$ and $s_{b_2} = 0.0567$.

a. Compute MSR and MSE.
b. Compute F and perform the appropriate F test. Use $\alpha = 0.05$.
c. Perform a t test for the significance of β_1. Use $\alpha = 0.05$.
d. Perform a t test for the significance of β_2. Use $\alpha = 0.05$.

EXER2

13 Refer to the data presented in exercise 2. The estimated regression equation for these data is

$$\hat{y} = -18.4 + 2.01x_1 + 4.74x_2$$

Here SST = 15 182.9, SSR = 14 052.2, $s_{b_1} = 0.2471$ and $s_{b_2} = 0.9484$.

a. Test for a significant relationship among X_1, X_2 and Y. Use $\alpha = 0.05$.
b. Is β_1 significant? Use $\alpha = 0.05$.
c. Is β_2 significant? Use $\alpha = 0.05$.

14 The following estimated regression equation was developed for a model involving two independent variables.

$$\hat{y} = 40.7 + 8.63x_1 + 2.71x_2$$

After X_2 was dropped from the model, the least squares method was used to obtain an estimated regression equation involving only X_1 as an independent variable.

$$\hat{y} = 42.0 + 9.01x_1$$

a. Give an interpretation of the coefficient of X_1 in both models.
b. Could multicollinearity explain why the coefficient of X_1 differs in the two models? If so, how?

Applications

15 In exercise 4 the following estimated regression equation relating sales to inventory investment and advertizing expenditures was given.

$$\hat{y} = 25 + 10x_1 + 8x_2$$

The data used to develop the model came from a survey of ten stores; for these data SST = 16 000 and SSR = 12 000.

a. Compute SSE, MSE and MSR.
b. Use an F test and a 0.05 level of significance to determine whether there is a relationship among the variables.

16 Refer to exercise 5.

a. Use $\alpha = 0.01$ to test the hypotheses

$$H_0\colon \beta_1 = \beta_2 = 0$$
$$H_1\colon \beta_1 \text{ and/or } \beta_2 \text{ is not equal to zero}$$

for the model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, where

X_1 = television advertising (€1000s)
X_2 = newspaper advertising (€1000s)

b. Use $\alpha = 0.05$ to test the significance of β_1. Should X_1 be dropped from the model?
c. Use $\alpha = 0.05$ to test the significance of β_2. Should X_2 be dropped from the model?

15.6 Using the estimated regression equation for estimation and prediction

The procedures for estimating the mean value of Y and predicting an individual value of Y in multiple regression are similar to those in regression analysis involving one independent variable. First, recall that in Chapter 14 we showed that the point estimate of the expected value of Y for a given value of X was the same as the point estimate of an individual value of Y. In both cases, we used $\hat{y} = b_0 + b_1x$ as the point estimate.

Table 15.4 The 95% confidence and prediction intervals for Eurodistributor

		Confidence Interval		Prediction Interval	
Value of X_1	Value of X_2	Lower Limit	Upper Limit	Lower Limit	Upper Limit
50	2	3.146	4.924	2.414	5.656
50	3	4.127	5.789	3.368	6.548
50	4	4.815	6.948	4.157	7.607
100	2	6.258	7.926	5.500	8.683
100	3	7.385	8.645	6.520	9.510
100	4	8.135	9.742	7.362	10.515

In multiple regression we use the same procedure. That is, we substitute the given values of $X_1, X_2, \ldots X_p$ into the estimated regression equation and use the corresponding value of $\hat{y}$ as the point estimate. Suppose that for the Eurodistributor example we want to use the estimated regression equation involving X_1 (distance) and X_2 (number of deliveries) to develop two interval estimates:

1 A *confidence interval* of the mean travel time for all trucks that travel 100 kilometres and make two deliveries.

2 A *prediction interval* of the travel time for *one specific* truck that travels 100 kilometres and makes two deliveries.

Using the estimated regression equation $\hat{y} = -0.869 + 0.0611x_1 + 0.923x_2$ with $X_1 = 100$ and $X_2 = -2$, we obtain the following value of $\hat{y}$.

$$\hat{y} = -0.869 + 0.0611(100) + 0.923(2) = 7.09$$

Hence, the point estimate of travel time in both cases is approximately seven hours.

To develop interval estimates for the mean value of Y and for an individual value of Y, we use a procedure similar to that for regression analysis involving one independent variable.

The formulae required are beyond the scope of the text, but computer packages for multiple regression analysis will often provide confidence intervals once the values of $X_1, X_2, \ldots X_p$ are specified by the user. In Table 15.4 we show the 95 per cent confidence and prediction intervals for the Eurodistributor example for selected values of X_1 and X_2; these values were obtained using MINITAB. Note that the interval estimate for an individual value of Y is wider than the interval estimate for the expected value of Y. This difference simply reflects the fact that for given values of X_1 and X_2 we can estimate the mean travel time for all trucks with more precision than we can predict the travel time for one specific truck.

Exercises

Methods

17 In exercise 1, the following estimated regression equation based on ten observations was presented.

$$\hat{y} = 29.1270 + 0.5906x_1 + 0.4980x_2$$

a. Develop a point estimate of the mean value of Y when $X_1 = 180$ and $X_2 = 310$.

b. Develop a point estimate for an individual value of Y when $X_1 = 180$ and $X_2 = 310$.

EXER2

18 Refer to the data in exercise 2. The estimated regression equation for those data is

$$\hat{y} = -18.4 + 2.01x_1 + 4.74x_2$$

a. Develop a 95 per cent confidence interval for the mean value of Y when $X_1 = 45$ and $X_2 = 15$.

b. Develop a 95 per cent prediction interval for Y when $X_1 = 45$ and $X_2 = 15$.

Applications

TOULON

19 In exercise 5, the owner of Toulon Theatres used multiple regression analysis to predict gross revenue (Y) as a function of television advertising (X_1) and newspaper advertising (X_2). The estimated regression equation was

$$\hat{y} = 83.2 + 2.29x_1 + 1.30x_2$$

a. What is the gross revenue expected for a week when €3500 is spent on television advertising ($X_1 = 3.5$) and €1800 is spent on newspaper advertising ($X_2 = 1.8$)?

b. Provide a 95 per cent confidence interval for the mean revenue of all weeks with the expenditures listed in part (a).

c. Provide a 95 per cent prediction interval for next week's revenue, assuming that the advertising expenditures will be allocated as in part (a).

15.7 Qualitative independent variables

Thus far, the examples we considered involved quantitative independent variables such as distance travelled and number of deliveries. In many situations, however, we must work with **qualitative independent variables** such as gender (male, female), method of payment (cash, credit card, cheque) and so on. The purpose of this section is to show how qualitative variables are handled in regression analysis. To illustrate the use and interpretation of a qualitative independent variable, we will consider a problem facing the managers of Johansson Filtration.

An example: Johansson Filtration

Johansson Filtration provides maintenance service for water-filtration systems throughout southern Denmark. Customers contact Johansson with requests for maintenance service on their water-filtration systems. To estimate the service time and the service cost, Johansson's managers wish to predict the repair time necessary for each maintenance request. Hence, repair time in hours is the dependent variable. Repair time is believed to be related to two factors, the number of months since the last maintenance service and the type of repair problem (mechanical or electrical). Data for a sample of ten service calls are reported in Table 15.5.

Let Y denote the repair time in hours and X_1 denote the number of months since the last maintenance service. The regression model that uses only X_1 to predict Y is

$$Y = \beta_0 + \beta_1 x_1 + \varepsilon$$

JOHANNSON

Table 15.5 Data for the Johansson Filtration example

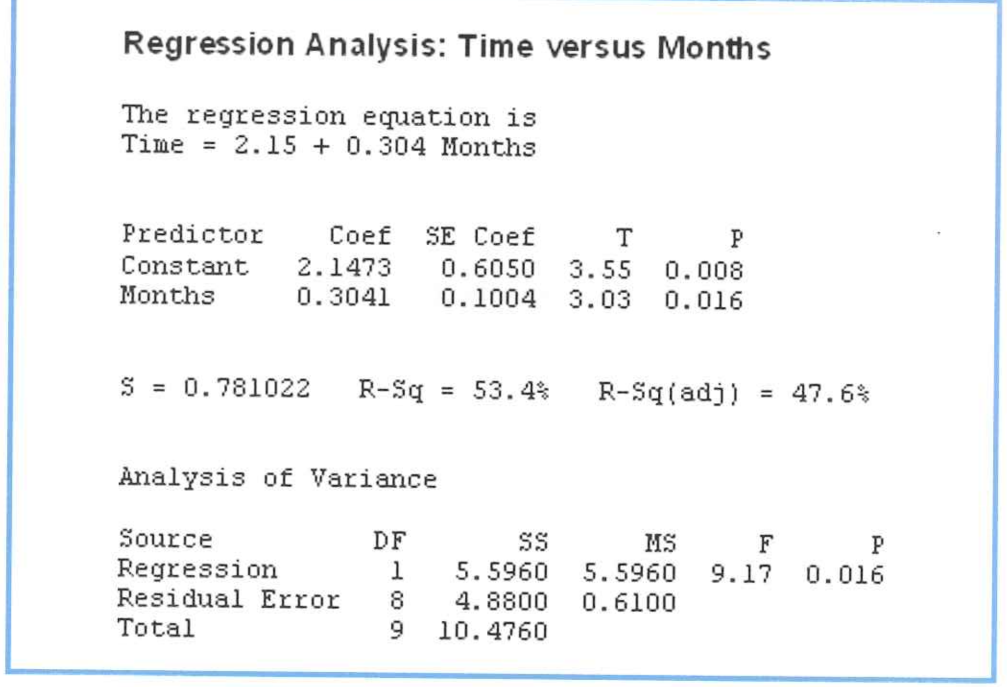
Service call	Months since last service	Type of repair	Repair time in hours
1	2	electrical	2.9
2	6	mechanical	3.0
3	8	electrical	4.8
4	3	mechanical	1.8
5	2	electrical	2.9
6	7	electrical	4.9
7	9	mechanical	4.2
8	8	mechanical	4.8
9	4	electrical	4.4
10	6	electrical	4.5

Using MINITAB to develop the estimated regression equation, we obtained the output shown in Figure 15.7. The estimated regression equation is

$$\hat{y} = 2.15 + 0.304x_1 \quad \textbf{(15.17)}$$

At the 0.05 level of significance, the p-value of 0.016 for the t (or F) test indicates that the number of months since the last service is significantly related to repair time. R-sq = 53.4 per cent indicates that X_1 alone explains 53.4 per cent of the variability in repair time.

Figure 15.7 MINITAB output for Johansson Filtration with months since last service (X_1) as the independent variable



```
Regression Analysis: Time versus Months

The regression equation is
Time = 2.15 + 0.304 Months

Predictor     Coef   SE Coef     T      P
Constant    2.1473    0.6050  3.55  0.008
Months      0.3041    0.1004  3.03  0.016

S = 0.781022   R-Sq = 53.4%   R-Sq(adj) = 47.6%

Analysis of Variance

Source          DF       SS      MS     F      P
Regression       1   5.5960  5.5960  9.17  0.016
Residual Error   8   4.8800  0.6100
Total            9  10.4760
```

To incorporate the type of failure into the regression model, we define the following variable.

$$X_2 = 0 \text{ if the type of repair is mechanical}$$
$$X_2 = 1 \text{ if the type of repair is electrical}$$

In regression analysis X_2 is called a **dummy** or *indicator* **variable**. Using this dummy variable, we can write the multiple regression model as

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Table 15.6 is the revised data set that includes the values of the **dummy variable**. Using MINITAB and the data in Table 15.6, we can develop estimates of the model parameters. The MINITAB output in Figure 15.8 shows that the estimated multiple regression equation is

$$\hat{y} = 0.93 + 0.388x_1 + 1.26x_2 \quad \textbf{(15.18)}$$

At the 0.05 level of significance, the p-value of 0.001 associated with the F test (F = 21.36) indicates that the regression relationship is significant. The t test part of the printout in Figure 15.8 shows that both months since last service (p-value = 0.000) and type of repair (p-value = 0.005) are statistically significant. In addition, R-sq = 85.9 per cent and R-sq(adj) = 81.9 per cent indicate that the estimated regression equation does a good job of explaining the variability in repair times. Thus, equation (15.18) should prove helpful in estimating the repair time necessary for the various service calls.

Interpreting the parameters

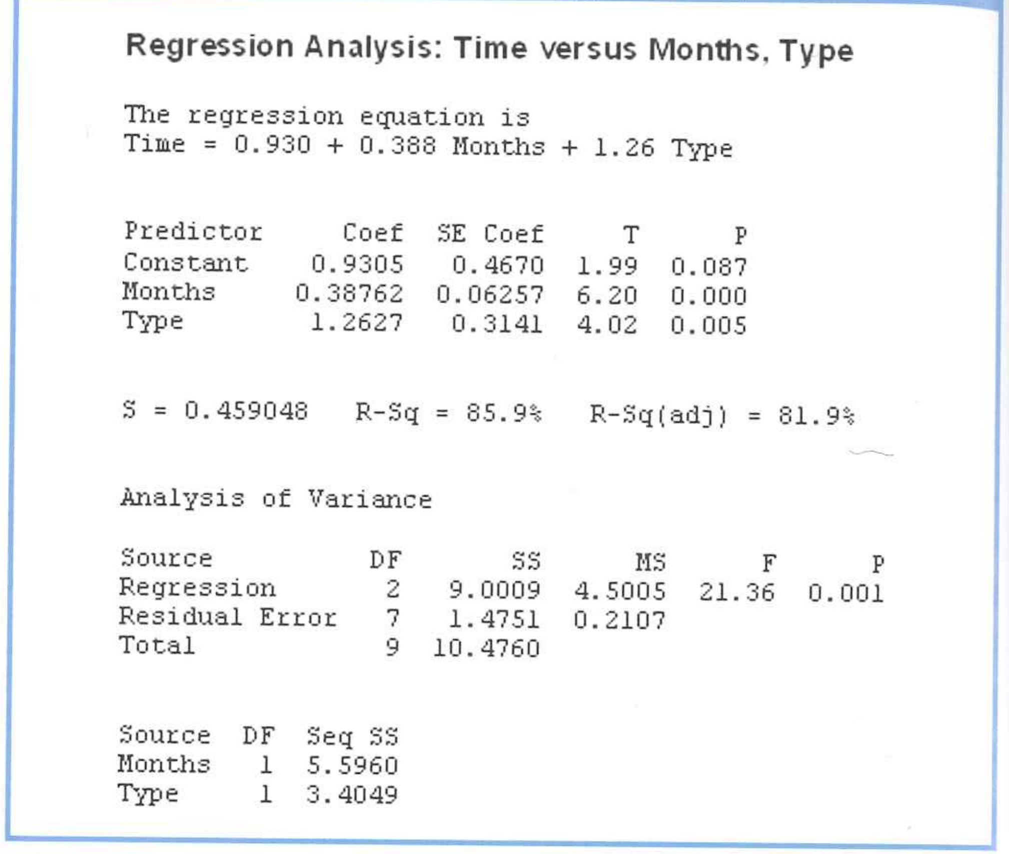
The multiple regression equation for the Johansson Filtration example is

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \quad \textbf{(15.19)}$$

Table 15.6 Data for the Johansson Filtration example with type of repair indicated by a dummy variable (X_2 = 0 for mechanical; X_2 = 1 for electrical)

Customer	Months since last service (X_1)	Type of repair (X_2)	Repair time in hours (Y)
1	2	1	2.9
2	6	0	3.0
3	8	1	4.8
4	3	0	1.8
5	2	1	2.9
6	7	1	4.9
7	9	0	4.2
8	8	0	4.8
9	4	1	4.4
10	6	1	4.5

Figure 15.8 MINITAB output for Johansson Filtration with months since last service (X_1) and type of repair (X_2) as the independent variables



Regression Analysis: Time versus Months, Type

```
The regression equation is
Time = 0.930 + 0.388 Months + 1.26 Type


Predictor      Coef   SE Coef      T      P
Constant     0.9305    0.4670   1.99  0.087
Months      0.38762   0.06257   6.20  0.000
Type         1.2627    0.3141   4.02  0.005


S = 0.459048    R-Sq = 85.9%    R-Sq(adj) = 81.9%


Analysis of Variance

Source          DF       SS      MS      F      P
Regression       2   9.0009  4.5005  21.36  0.001
Residual Error   7   1.4751  0.2107
Total            9  10.4760


Source  DF  Seq SS
Months   1  5.5960
Type     1  3.4049
```

To understand how to interpret the parameters β_0, β_1, and β_2 when a qualitative variable is present, consider the case when $X_2 = 0$ (mechanical repair). Using $E(Y|\text{mechanical})$ to denote the mean or expected value of repair time *given* a mechanical repair, we have

$$E(Y|\text{mechanical}) = \beta_0 + \beta_1 x_1 + \beta_2(0) = \beta_0 + \beta_1 x_1 \qquad (15.20)$$

Similarly, for an electrical repair ($X_2 = 1$), we have

$$\begin{aligned} E(Y|\text{electrical}) &= \beta_0 + \beta_1 x_1 + \beta_2(1) = \beta_0 + \beta_1 x_1 + \beta_2 \\ &= (\beta_0 + \beta_2) + \beta_1 x_1 \end{aligned} \qquad (15.21)$$

Comparing equations (15.20) and (15.21), we see that the mean repair time is a linear function of X_1 for both mechanical and electrical repairs. The slope of both equations is β_1, but the y-intercept differs. The y-intercept is β_0 in equation (15.20) for mechanical repairs and $(\beta_0 + \beta_2)$ in equation (15.21) for electrical repairs. The interpretation of β_2 is that it indicates the difference between the mean repair time for an electrical repair and the mean repair time for a mechanical repair.

If β_2 is positive, the mean repair time for an electrical repair will be greater than that for a mechanical repair; if β_2 is negative, the mean repair time for an electrical repair will be less than that for a mechanical repair. Finally, if $\beta_2 = 0$, there is no difference in the mean repair time between electrical and mechanical repairs and the type of repair is not related to the repair time.

Using the estimated multiple regression equation $\hat{y} = 0.93 + 0.388x_1 + 1.26x_2$, we see that 0.93 is the estimate of β_0 and 1.26 is the estimate of β_2. Thus, when $X_2 = 0$ (mechanical repair)

$$\hat{y} = 0.93 + 0.388x_1 \qquad (15.22)$$

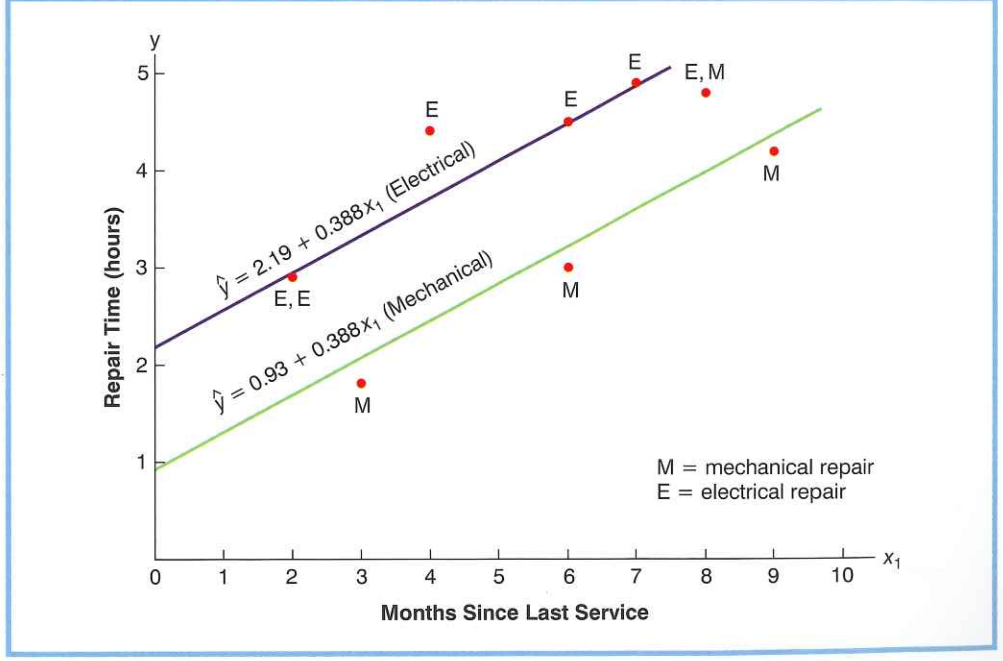
and when $X_2 = 1$ (electrical repair)

$$\begin{aligned} \hat{y} &= 0.93 + 0.388x_1 + 1.26(1) \\ &= 2.19 + 0.388x_1 \end{aligned} \qquad (15.23)$$

In effect, the use of a dummy variable for type of repair provides two equations that can be used to predict the repair time, one corresponding to mechanical repairs and one corresponding to electrical repairs. In addition, with $b_2 = 1.26$, we learn that, on average, electrical repairs require 1.26 hours longer than mechanical repairs.

Figure 15.9 is the plot of the Johansson data from Table 15.6. Repair time in hours (Y) is represented by the vertical axis and months since last service (X_1) is represented by the horizontal axis. A data point for a mechanical repair is indicated by an M and a data point for an electrical repair is indicated by an E. Equations (15.22) and (15.23) are plotted on the graph to show graphically the two equations that can be used to predict the repair time, one corresponding to mechanical repairs and one corresponding to electrical repairs.

Figure 15.9 Scatter diagram for the Johansson Filtration repair data from Table 15.6



More complex qualitative variables

Because the qualitative variable for the Johansson Filtration example had two levels (mechanical and electrical), defining a dummy variable with zero indicating a mechanical repair and one indicating an electrical repair was easy. However, when a qualitative variable has more than two levels, care must be taken in both defining and interpreting the dummy variables. As we will show, if a qualitative variable has k levels, $k - 1$ dummy variables are required, with each dummy variable being coded as 0 or 1.

For example, suppose a manufacturer of copy machines organized the sales territories for a particular area into three regions: A, B and C. The managers want to use regression analysis to help predict the number of copiers sold per week. With the number of units sold as the dependent variable, they are considering several independent variables (the number of sales personnel, advertising expenditures and so on). Suppose the managers believe sales region is also an important factor in predicting the number of copiers sold. Because sales region is a qualitative variable with three levels, A, B and C, we will need $3 - 1 = 2$ dummy variables to represent the sales region. Each variable can be coded 0 or 1 as follows.

$$X_1 = \begin{cases} 1 \text{ if sales region B} \\ 0 \text{ otherwise} \end{cases}$$

$$X_2 = \begin{cases} 1 \text{ if sales region C} \\ 0 \text{ otherwise} \end{cases}$$

With this definition, we have the following values of X_1 and X_2.

Region	X_1	X_2
A	0	0
B	1	0
C	0	1

Observations corresponding to region A would be coded $X_1 = 0$, $X_2 = 0$; observations corresponding to region B would be coded $X_1 = 1$, $X_2 = 0$; and observations corresponding to region C would be coded $X_1 = 0$, $X_2 = 1$.

The regression equation relating the expected value of the number of units sold, $E(Y)$, to the dummy variables would be written as

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

To help us interpret the parameters β_0, β_1 and β_2, consider the following three variations of the regression equation.

$$\begin{aligned} E(Y \mid \text{region A}) &= \beta_0 + \beta_1(0) + \beta_2(0) = \beta_0 \\ E(Y \mid \text{region B}) &= \beta_0 + \beta_1(1) + \beta_2(0) = \beta_0 + \beta_1 \\ E(Y \mid \text{region C}) &= \beta_0 + \beta_1(0) + \beta_2(1) = \beta_0 + \beta_2 \end{aligned}$$

Therefore, β_0 is the mean or expected value of sales for region A; β_1 is the difference between the mean number of units sold in region B and the mean number of units sold in region A; and β_2 is the difference between the mean number of units sold in region C and the mean number of units sold in region A.

Two dummy variables were required because sales region is a qualitative variable with three levels. But the assignment of $X_1 = 0$, $X_2 = 0$ to indicate region A, $X_1 = 1$, $X_2 = 0$ to indicate region B, and $X_1 = 0$, $X_2 = 1$ to indicate region C was arbitrary. For example, we could have chosen $X_1 = 1$, $X_2 = 0$ to indicate region A, $X_1 = 0$, $X_2 = 0$ to indicate region B, and $X_1 = 0$, $X_2 = 1$ to indicate region C. In that case, β_1 would have been interpreted as the mean difference between regions A and B and β_2 as the mean difference between regions C and B.

Exercises

Methods

20 Consider a regression study involving a dependent variable Y, a quantitative independent variable X_1 and a qualitative variable with two levels (level 1 and level 2).

a. Write a multiple regression equation relating X_1 and the qualitative variable to Y.
b. What is the expected value of Y corresponding to level 1 of the qualitative variable?
c. What is the expected value of Y corresponding to level 2 of the qualitative variable?
d. Interpret the parameters in your regression equation.

21 Consider a regression study involving a dependent variable Y, a quantitative independent variable X_1, and a qualitative independent variable with three possible levels (level 1, level 2 and level 3).

a. How many dummy variables are required to represent the qualitative variable?
b. Write a multiple regression equation relating X_1 and the qualitative variable to Y.
c. Interpret the parameters in your regression equation.

Applications

22 Management proposed the following regression model to predict sales at a fast-food outlet.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

where

X_1 = number of competitors within one kilometre
X_2 = population within one kilometre (000s)
X_3 = 1 if drive-up window present
0 otherwise
Y = sales (€000s)

The following estimated regression equation was developed after 20 outlets were surveyed.

$$\hat{y} = 10.1 - 4.2x_1 + 6.8x_2 + 15.3x_3$$

a. What is the expected amount of sales attributable to the drive-up window?
b. Predict sales for a store with two competitors, a population of 8000 within one kilometre and no drive-up window.
c. Predict sales for a store with one competitor, a population of 3000 within one kilometre and a drive-up window.

23 Refer to the Johansson Filtration problem introduced in this section. Suppose that in addition to information on the number of months since the machine was serviced and whether a mechanical or an electrical failure had occurred, the managers obtained a list showing which engineer performed the service. The revised data follow.

REPAIR

Repair time in hours	Months since last service	Type of repair	Engineer
2.9	2	Electrical	Heinz Kolb
3.0	6	Mechanical	Heinz Kolb
4.8	8	Electrical	Wolfgang Linz
1.8	3	Mechanical	Heinz Kolb
2.9	2	Electrical	Heinz Kolb
4.9	7	Electrical	Wolfgang Linz
4.2	9	Mechanical	Wolfgang Linz
4.8	8	Mechanical	Wolfgang Linz
4.4	4	Electrical	Wolfgang Linz
4.5	6	Electrical	Heinz Kolb

a. Ignore for now the months since the last maintenance service (X_1) and the engineer who performed the service. Develop the estimated simple linear regression equation to predict the repair time (Y) given the type of repair (X_2). Recall that $X_2 = 0$ if the type of repair is mechanical and 1 if the type of repair is electrical.

b. Does the equation that you developed in part (a) provide a good fit for the observed data? Explain.

c. Ignore for now the months since the last maintenance service and the type of repair associated with the machine. Develop the estimated simple linear regression equation to predict the repair time given the engineer who performed the service. Let $X_3 = 0$ if Heinz Kolb performed the service and $X_3 = 1$ if Wolfgang Linz performed the service.

d. Does the equation that you developed in part (c) provide a good fit for the observed data? Explain.

24 In a multiple regression analysis by McIntyre (1994), Tar, Nicotine and Weight are considered as possible predictors of Carbon Monoxide (CO) content for 25 different brands of cigarette. Details of variables and data follow.

Brand	The cigarette brand
Tar	The tar content (in mg)
Nicotine	The nicotine content (in mg)
Weight	The weight (in g)
CO	The carbon monoxide (CO) content (in mg)

GIGARETTES

Brand	Tar	Nicotine	Weight	CO
Alpine	14.1	0.86	.9853	13.6
Benson&Hedges	16.0	1.06	1.0938	16.6
BullDurham	29.8	2.03	1.1650	23.5
CamelLights	8.0	0.67	0.9280	10.2
Carlton	4.1	0.40	0.9462	5.4
Chesterfield	15.0	1.04	0.8885	15.0
.GoldenLights	8.8	0.76	1.0267	9.0
Kent	12.4	0.95	0.9225	12.3
Kool	16.6	1.12	0.9372	16.3
L&M	14.9	1.02	0.8858	15.4
LarkLights	13.7	1.01	0.9643	13.0
Marlboro	15.1	0.90	0.9316	14.4
Merit	7.8	0.57	0.9705	10.0
MultiFilter	11.4	0.78	1.1240	10.2
NewportLights	9.0	0.74	0.8517	9.5
Now	1.0	0.13	0.7851	1.5
OldGold	17.0	1.26	0.9186	18.5
PallMallLight	12.8	1.08	1.0395	12.6
Raleigh	15.8	0.96	0.9573	17.5
SalemUltra	4.5	0.42	0.9106	4.9
Tareyton	14.5	1.01	1.0070	15.9
True	7.3	0.61	0.9806	8.5
ViceroyRichLight	8.6	0.69	0.9693	10.6
VirginiaSlims	15.2	1.02	0.9496	13.9
WinstonLights	12.0	0.82	1.1184	14.9

a. Examine correlations between variables in the study and hence assess the possibility of problems of multicollinearity affecting any subsequent regression model involving independent variables Tar and Nicotine.

b. Thus develop an estimated multiple regression equation using an appropriate number of the independent variables featured in the study.

c. Are your predictors statistically significant? Use $\alpha = 0.05$. What explanation can you give for the results observed?

25 The data below (Dunn, 2007) come from a study investigating a new method of measuring body composition. Body fat percentage, age and gender is given for 18 adults aged between 23 and 61.

BODYFAT

Age	Percent.Fat	Gender
23	9.5	M
23	27.9	F
27	7.8	M
27	17.8	M
39	31.4	F
41	25.9	F
45	27.4	M
49	25.2	F
50	31.1	F
53	34.7	F
53	42	F
54	29.1	F
56	32.5	F
57	30.3	F
58	33	F
58	33.8	F
60	41.1	F
61	34.5	F

a. Develop an estimated regression equation that relates Age and Gender to Percent.Fat
b. Is Age a significant factor in predicting Percent.Fat? Explain. Use $\alpha = 0.05$.
c. What is the estimated body fat percentage for a female aged 45?

15.8 Residual analysis

In Chapter 14 we pointed out that standardized residuals were frequently used in residuals plots and in the identification of outliers. The general formula for the standardized residual for observation *i* follows.

Standardized residual for observation *i*

$$\frac{y_i - \hat{y}_i}{s_{y_i - \hat{y}_i}}$$

where

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

The general formula for the standard deviation of residual *i* is defined as follows.

Standard deviation of residual *i*

$$s_{y_i - \hat{y}_i} = s\sqrt{1 - h_i} \qquad \textbf{(15.25)}$$

where

s = standard error of the estimate
h_i = leverage of observation *i*

The **leverage** of an observation is determined by how far the values of the independent variables are from their means. The computation of h_i, $s_{y_i - \hat{y}_i}$ and hence the standardized residual for observation *i* in multiple regression analysis is too complex to be done by hand. However, the standardized residuals can be easily obtained as part of the output from statistical software packages. Table 15.7 lists the predicted values, the residuals, and the standardized residuals for the Eurodistributor example presented previously in this chapter; we obtained these values by using the MINITAB statistical software package. The predicted values in the table are based on the estimated regression equation

$$\hat{y} = -0.869 + 0.0611x_1 + 0.923x_2$$

The standardized residuals and the predicted values of *Y* from Table 15.7 are used in the standardized residual plot in Figure 15.10.

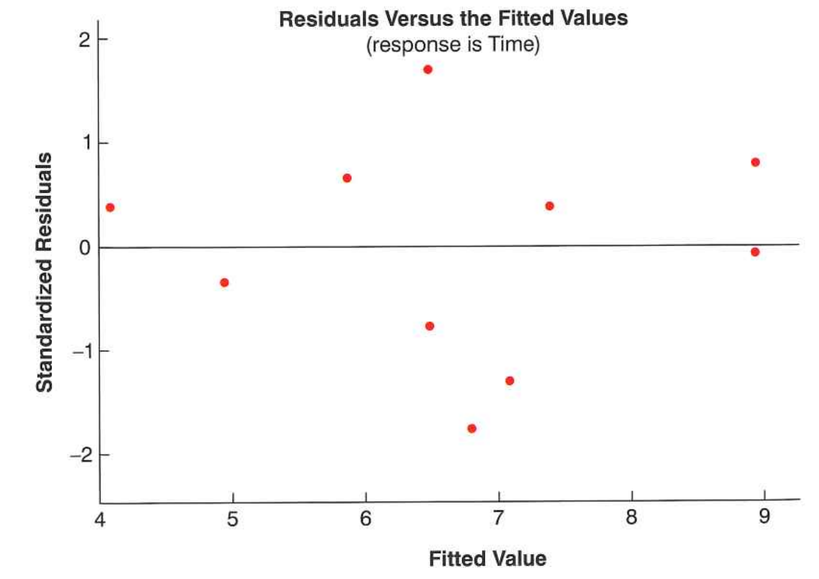
This standardized residual plot does not indicate any unusual abnormalities. Also, all of the standardized residuals are between −2 and +2; hence, we have no reason to question the assumption that the error term ε is normally distributed. We conclude that the model assumptions are reasonable.

Table 15.7 Residuals and standardized residuals for the Eurodistributor regression analysis

Distance travelled (X_1)	Deliveries (X_2)	Travel time (Y)	Predicted value ($\hat{y}$)	Residual ($y - \hat{y}$)	Standardized residual
100	4	9.3	8.93846	0.361540	0.78344
50	3	4.8	4.95830	−0.158305	−0.34962
100	4	8.9	8.93846	−0.038460	−0.08334
100	2	6.5	7.09161	−0.591609	−1.30929
50	2	4.2	4.03488	0.165121	0.38167
80	2	6.2	5.86892	0.331083	0.65431
75	3	7.4	6.48667	0.913330	1.68917
65	4	6.0	6.79875	−0.798749	−1.77372
90	3	7.6	7.40369	0.196311	0.36703
90	2	6.1	6.48026	−0.380263	−0.77639

A normal probability plot also can be used to determine whether the distribution of ε appears to be normal. The procedure and interpretation for a normal probability plot were discussed in Section 14.8. The same procedure is appropriate for multiple regression. Again, we would use a statistical software package to perform the computations and provide the normal probability plot.

Figure 15.10 Standardized residual plot for the Eurodistributor multiple regression analysis



Detecting outliers

An **outlier** is an observation that is unusual in comparison with the other data; in other words, an outlier does not fit the pattern of the other data. In Chapter 14 we showed an example of an outlier and discussed how standardized residuals can be used to detect outliers.

MINITAB classifies an observation as an outlier if the value of its standardized residual is less than −2 or greater than +2. Applying this rule to the standardized residuals for the Eurodistributor example (see Table 15.7), we do not detect any outliers in the data set.

In general, the presence of one or more outliers in a data set tends to increase s, the standard error of the estimate, and hence increase $s_{y_i - \hat{y}_i}$, the standard deviation of residual i. Because $s_{y_i - \hat{y}_i}$ appears in the denominator of the formula for the standardized residual (15.24), the size of the standardized residual will decrease as s increases.

As a result, even though a residual may be unusually large, the large denominator in expression (15.24) may cause the standardized residual rule to fail to identify the observation as being an outlier. We can circumvent this difficulty by using a form of standardized residuals called **studentized deleted residuals**.

Studentized deleted residuals and outliers

Suppose the ith observation is deleted from the data set and a new estimated regression equation is developed with the remaining $n - 1$ observations. Let $s_{(i)}$ denote the standard error of the estimate based on the data set with the ith observation deleted. If we compute the standard deviation of residual i (15.25) using $s_{(i)}$ instead of s, and then compute the standardized residual for observation i (15.24) using the revised value, the resulting standardized residual is called a studentized deleted residual.

If the ith observation is an outlier, $s_{(i)}$ will be less than s. The absolute value of the ith studentized deleted residual therefore will be larger than the absolute value of the standardized residual. In this sense, studentized deleted residuals may detect outliers that standardized residuals do not detect. Many statistical software packages provide an option for obtaining studentized deleted residuals. Using MINITAB, we obtained the studentized deleted residuals for the Eurodistributor example; the results are reported in Table 15.8. The t distribution can be used to determine whether the studentized deleted residuals indicate the presence of outliers. Recall that p denotes the number of independent variables and n denotes the number of observations. Hence, if we delete the ith observation, the number of observations in the reduced data set is $n - 1$; in this case the error sum of squares has $(n - 1) - p - 1$ degrees of freedom. For the Eurodistributor example with $n = 10$ and $p = 2$, the degrees of freedom for the error sum of squares with the ith observation deleted is $9 - 2 - 1 = 6$. At a 0.05 level of significance, the t distribution (Table 2 of Appendix B) shows that with six degrees of freedom, $t_{0.025} = 2.447$. If the value of the ith studentized deleted residual is less than −2.447 or greater than +2.447, we can conclude that the ith observation is an outlier. The studentized deleted residuals in Table 15.8 do not exceed those limits; therefore, we conclude that outliers are not present in the data set.

Table 15.8 Studentized deleted residuals for Eurodistributor

Distance travelled (X_1)	Deliveries (X_2)	Travel time (Y)	Standardized residual	Studentized deleted residual
100	4	9.3	0.78344	0.75938
50	3	4.8	−0.34962	−0.32654
100	4	8.9	−0.08334	−0.0772
100	2	6.5	−1.30929	−1.39494
50	2	4.2	0.38167	0.35709
80	2	6.2	0.65431	0.62519
75	3	7.4	1.68917	2.03187
65	4	6.0	−1.77372	−2.21314
90	3	7.6	0.36703	0.34312
90	2	6.1	−0.77639	−0.7519

Influential observations

In Section 14.9 we discussed how the leverage of an observation can be used to identify observations for which the value of the independent variable may have a strong influence on the regression results. As we acknowledged, the leverage (h_i) of an observation, measures how far the values of the independent variables are from their mean values. The leverage values are easily obtained as part of the output from statistical software packages. MINITAB computes the leverage values and uses the rule of thumb

$$h_i > 3(p + 1)/n$$

to identify **influential observations**. For the Eurodistributor example with $p = 2$ independent variables and $n = 10$ observations, the critical value for leverage is $3(2 + 1)/10 = 0.9$. The leverage values for the Eurodistributor example obtained by using MINITAB are reported in Table 15.9. As h_i does not exceed 0.9, no influential observations in the data set are detected.

Using Cook's distance measure to identify influential observations

A problem that can arise in using leverage to identify influential observations is that an observation can be identified as having high leverage and not necessarily be influential in terms of the resulting estimated regression equation. For example, Table 15.10 shows a

Table 15.9 Leverage and Cook's distance measures for Eurodistributor

Distance travelled (X_1)	Deliveries (X_2)	Travel time (Y)	Leverage (h_i)	Cook's D (D_i)
100	4	9.3	0.351704	0.110994
50	3	4.8	0.375863	0.024536
100	4	8.9	0.351704	0.001256
100	2	6.5	0.378451	0.347923
50	2	4.2	0.430220	0.036663
80	2	6.2	0.220557	0.040381
75	3	7.4	0.110009	0.117561
65	4	6.0	0.382657	0.650029
90	3	7.6	0.129098	0.006656
90	2	6.1	0.269737	0.074217

Table 15.10 Data set illustrating potential problem using the leverage criterion

x_i	y_i	Leverage h_i
1	18	0.204170
1	21	0.204170
2	22	0.164205
3	21	0.138141
4	23	0.125977
4	24	0.125977
5	26	0.127715
15	39	0.909644

data set consisting of eight observations and their corresponding leverage values (obtained by using MINITAB). Because the leverage for the eighth observation is 0.91 > 0.75 (the critical leverage value), this observation is identified as influential. Before reaching any final conclusions, however, let us consider the situation from a different perspective.

Figure 15.11 shows the scatter diagram and the estimated regression equation corresponding to the data set in Table 15.10. We used MINITAB to develop the following estimated regression equation for these data.

$$\hat{y} = 18.2 + 1.39\,x$$

The straight line in Figure 15.11 is the graph of this equation. Now, let us delete the observation $X = 15$, $Y = 39$ from the data set and fit a new estimated regression equation to the remaining seven observations; the new estimated regression equation is

$$\hat{y} = 18.1 + 1.42\,x$$

We note that the y-intercept and slope of the new estimated regression equation are not fundamentally different from the values obtained by using all the data. Although the leverage criterion identified the eighth observation as influential, this observation clearly had little influence on the results obtained. Thus, in some situations using only leverage to identify influential observations can lead to wrong conclusions.

Cook's distance measure uses both the leverage of observation i, h_i, and the residual for observation i, $(y_i - \hat{y}_i)$, to determine whether the observation is influential.

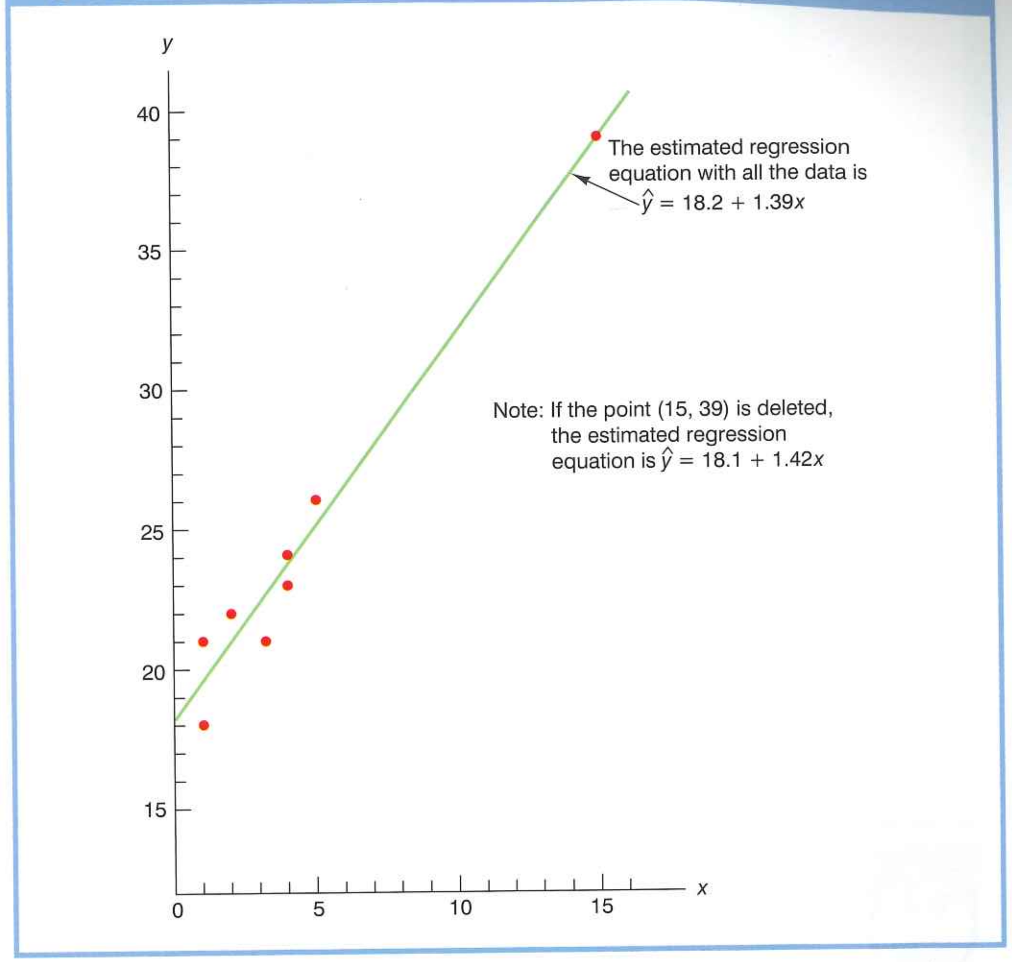
Cook's distance measure

$$D_i = \frac{(y_i - \hat{y}_i)^2\, h_i}{(p-1)s^2\,(1-h_i)^2} \qquad \textbf{(15.26)}$$

where

D_i = Cook's distance measure for observation i
$y_i - \hat{y}_i$ = the residual for observation i
h_i = the leverage for observation i
p = the number of independent variables
s = the standard error of the estimate

Figure 15.11 Scatter diagram for the data set in Table 15.10



The estimated regression equation with all the data is $\hat{y} = 18.2 + 1.39x$

Note: If the point (15, 39) is deleted, the estimated regression equation is $\hat{y} = 18.1 + 1.42x$

The value of Cook's distance measure will be large and indicate an influential observation if the residual or the leverage is large. As a rule of thumb, values of $D_i > 1$ indicate that the ith observation is influential and should be studied further. The last column of Table 15.9 provides Cook's distance measure for the Eurodistributor problem as given by MINITAB. Observation 8 with $D_i = 0.650029$ has the most influence. However, applying the rule $D_i > 1$, we should not be concerned about the presence of influential observations in the Eurodistributor data set.

Exercises

Methods

26 Data for two variables, X and Y, follow.

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

a. Develop the estimated regression equation for these data.

b. Plot the standardized residuals versus $\hat{y}$. Do there appear to be any outliers in these data? Explain.

c. Compute the studentized deleted residuals for these data. At the 0.05 level of significance, can any of these observations be classified as an outlier? Explain.

27 Data for two variables, X and Y, follow.

x_i	22	24	26	28	40
y_i	12	21	31	35	70

a. Develop the estimated regression equation for these data.

b. Compute the studentized deleted residuals for these data. At the 0.05 level of significance, can any of these observations be classified as an outlier? Explain.

c. Compute the leverage values for these data. Do there appear to be any influential observations in these data? Explain.

d. Compute Cook's distance measure for these data. Are any observations influential? Explain.

Applications

GIGARETTES

28 Exercise 5 gave data on weekly gross revenue, television advertising, and newspaper advertising for Toulon theatres.

a. Find an estimated regression equation relating weekly gross revenue to television and newspaper advertising.

b. Plot the standardized residuals against $\hat{y}$. Does the residual plot support the assumptions about ε? Explain.

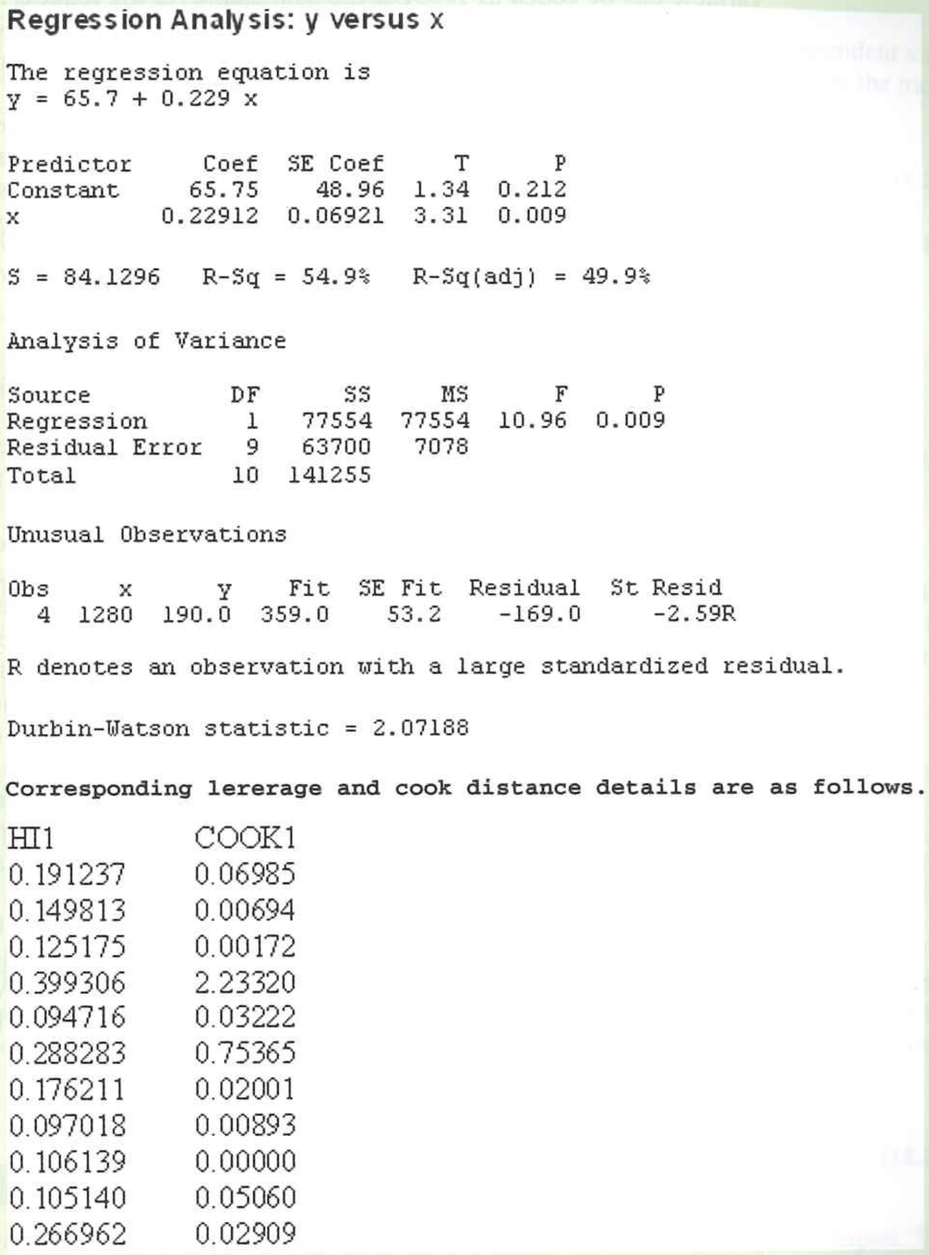
c. Check for any outliers in these data. What are your conclusions?

d. Are there any influential observations? Explain.

29 Data (Tufte, 1974) on male deaths per million in 1950 for lung cancer (Y) and *per capita* cigarette consumption in 1930 (X) are given below:

Country	y	x	Country	y	x
Ireland	58	220	Norway	90	250
Sweden	115	310	Canada	150	510
Denmark	165	380	Australia	170	455
USA	190	1280	Holland	245	460
Switzerland	250	530	Finland	350	1115
GB	465	1145			

Results from a simple regression analysis of this information are as follows:



```
Regression Analysis: y versus x

The regression equation is
y = 65.7 + 0.229 x

Predictor      Coef   SE Coef      T      P
Constant      65.75     48.96   1.34  0.212
x           0.22912   0.06921   3.31  0.009

S = 84.1296   R-Sq = 54.9%   R-Sq(adj) = 49.9%

Analysis of Variance

Source          DF      SS     MS      F      P
Regression       1   77554  77554  10.96  0.009
Residual Error   9   63700   7078
Total           10  141255

Unusual Observations

Obs     x      y    Fit  SE Fit  Residual  St Resid
  4  1280  190.0  359.0    53.2    -169.0    -2.59R

R denotes an observation with a large standardized residual.

Durbin-Watson statistic = 2.07188

Corresponding lererage and cook distance details are as follows.

HI1        COOK1
0.191237   0.06985
0.149813   0.00694
0.125175   0.00172
0.399306   2.23320
0.094716   0.03222
0.288283   0.75365
0.176211   0.02001
0.097018   0.00893
0.106139   0.00000
0.105140   0.05060
0.266962   0.02909
```

Carry out any further statistical tests you deem appropriate, otherwise comment on the effectiveness of the linear modes.

15.9 Logistic regression

In many regression applications the dependent variable may only assume two discrete values. For instance, a bank might like to develop an estimated regression equation for predicting whether a person will be approved for a credit card. The dependent variable can be coded as $Y = 1$ if the bank approves the request for a credit card and $Y = 0$ if the bank rejects the request for a credit card. Using logistic regression we can estimate the probability that the bank will approve the request for a credit card given a particular set of values for the chosen independent variables.

Consider an application of logistic regression involving a direct mail promotion being used by Stamm Stores. Stamm owns and operates a national chain of women's fashion stores. Five thousand copies of an expensive four-colour sales catalogue have been printed, and each catalogue includes a coupon that provides a €50 discount on purchases of €200 or more.

The catalogues are expensive and Stamm would like to send them to only those customers who have the highest probability of making a €200 purchase using the discount coupons.

Management thinks that annual spending at Stamm Stores and whether a customer has a Stamm credit card are two variables that might be helpful in predicting whether a customer who receives the catalogue will use the coupon to make a €200 purchase. Stamm conducted a pilot study using a random sample of 50 Stamm credit card customers and 50 other customers who do not have a Stamm credit card. Stamm sent the catalogue to each of the 100 customers selected. At the end of a test period, Stamm noted whether the customer made a purchase (coded 1 if the customer made a purchase and 0 if not). The sample data for the first ten catalogue recipients are shown in Table 15.11. The amount each customer spent last year at Stamm is shown in thousands of euros and the credit card information has been coded as 1 if the customer has a Stamm credit card and 0 if not. In the Purchase column, a 1 is recorded if the sampled customer used the €50 discount coupon to make a purchase of €200 or more.

We might think of building a multiple regression model using the data in Table 15.11 to help Stamm predict whether a catalogue recipient will make a purchase. We would use Annual Spending and Stamm Card as independent variables and Purchase as the dependent variable.

Because the dependent variable may only assume the values of 0 or 1, however, the ordinary multiple regression model is not applicable. This example shows the type of situation for which logistic regression was developed. Let us see how logistic regression can be used to help Stamm predict which type of customer is most likely to take advantage of their promotion.

Table 15.11 Sample data for Stamm Stores

Customer	Annual spending (€000s)	Stamm card	Purchase
1	2.291	1	0
2	3.215	1	0
3	2.135	1	0
4	3.924	0	0
5	2.528	1	0
6	2.473	0	1
7	2.384	0	0
8	7.076	0	0
9	1.182	1	1
10	3.345	0	0

Logistic regression equation

In many ways logistic regression is like ordinary regression. It requires a dependent variable, Y, and one or more independent variables. In multiple regression analysis, the mean or expected value of Y, is referred to as the multiple regression equation.

$$E(Y) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots\cdots + \beta_px_p \quad (15.27)$$

In logistic regression, statistical theory as well as practice has shown that the relationship between $E(Y)$ and $X_1, X_2, \ldots X_p$ is better described by the following nonlinear equation.

Logistic regression equation

$$E(Y) = \frac{e^{\beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_px_p}}{1 + e^{\beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_px_p}} \quad (15.28)$$

If the two values of the dependent variable Y are coded as 0 or 1, the value of $E(Y)$ in equation (15.28) provides the *probability* that $Y = 1$ given a particular set of values for the independent variables $X_1, X_2, \ldots X_p$. Because of the interpretation of $E(Y)$ as a probability, the **logistic regression equation** is often written as follows.

Interpretation of *E*(*Y*) as a probability in logistic regression

$$E(Y) = P(y = 1|x_1, x_2, \ldots x_p) \quad (15.29)$$

To provide a better understanding of the characteristics of the logistic regression equation, suppose the model involves only one independent variable X and the values of the model parameters are $\beta_0 = -7$ and $\beta_1 = 3$. The logistic regression equation corresponding to these parameter values is

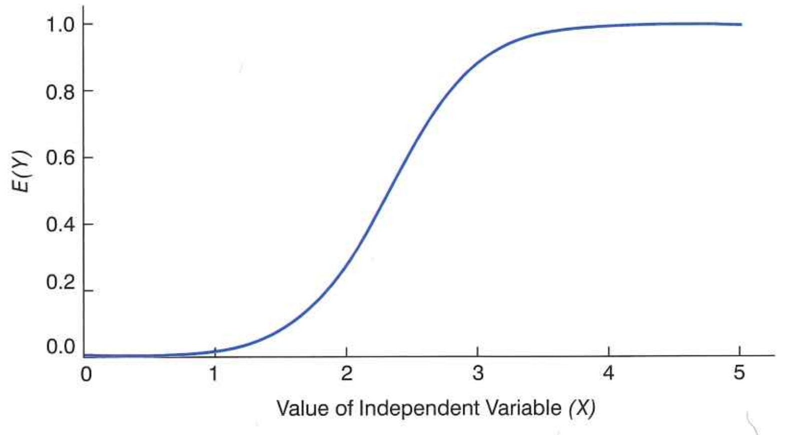
$$E(Y) = P(Y = 1 \mid x) = \frac{e^{\beta_0 + \beta_1x}}{1 + e^{\beta_0 + \beta_1x}} = \frac{e^{-7+3x}}{1 + e^{-7+3x}} \quad (15.30)$$

Figure 15.12 shows a graph of equation (15.30). Note that the graph is S-shaped. The value of $E(Y)$ ranges from 0 to 1, with the value of $E(Y)$ gradually approaching 1 as the value of X becomes larger and the value of $E(Y)$ approaching 0 as the value of X becomes smaller. Note also that the values of $E(Y)$, representing probability, increase fairly rapidly as X increases from 2 to 3. The fact that the values of $E(Y)$ range from 0 to 1 and that the curve is S-shaped makes equation (15.30) ideally suited to model the probability the dependent variable is equal to 1.

Estimating the logistic regression equation

In simple linear and multiple regression the least squares method is used to compute $b_0, b_1, \ldots, b_p$ as estimates of the model parameters $(\beta_0, \beta_1, \ldots, \beta_p)$. The nonlinear form of the logistic regression equation makes the method of computing estimates

Figure 15.12 Logistic regression equation for $\beta_0 = -7$ and $\beta_1 = 3$



more complex and beyond the scope of this text. We will use computer software to provide the estimates. The **estimated logistic regression equation** is

Estimated logistic regression equation

$$\hat{y} = \text{estimate of } P(Y = 1 \mid x_1, x_2, \ldots x_p) = \frac{e^{b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p}}{1 + e^{b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p}} \tag{15.31}$$

Here $\hat{y}$ provides an estimate of the probability that $Y = 1$, given a particular set of values for the independent variables.

Let us now return to the Stamm Stores example. The variables in the study are defined as follows:

$$Y = \begin{cases} 0 \text{ if the customer made no purchase during the test period} \\ 1 \text{ if the customer made a purchase during the test period} \end{cases}$$

$$X_1 = \text{annual spending at Stamm Stores (€000s)}$$

$$X_2 = \begin{cases} 0 \text{ if the customer does not have a Stamm credit card} \\ 1 \text{ if the customer has a Stamm credit card} \end{cases}$$

Therefore, we choose a logistic regression equation with two independent variables.

$$E(Y) = \frac{e^{\beta_0 + \beta_1x_1 + \cdots + \beta_px_p}}{1 + e^{\beta_0 + \beta_1x_1 + \cdots + \beta_px_p}} \tag{15.32}$$

Using the sample data (see Table 15.11), MINITAB's binary logistic regression procedure was used to compute estimates of the model parameters β_0, β_1, and β_2. A portion of the output obtained is shown in Figure 15.13. We see that $b_0 = -2.1464$, $b_1 = 0.3416$, and $b_2 = 1.0987$. Thus, the estimated logistic regression equation is

$$\hat{y} = \frac{e^{b_0 + b_1x_1 + \cdots + b_px_p}}{1 + e^{b_0 + b_1x_1 + \cdots + b_px_p}} = \frac{e^{-2.1464 + 0.3416x_1 + 1.0987x_2}}{1 + e^{-2.1464 + 0.3416x_1 + 1.0987x_2}} \tag{15.33}$$

Figure 15.13 Partial logistic regression output for the Stamm Stores example

```
Logistic Regression Table

                                                Odds     95% CI
Predictor       Coef   SE Coef      Z      P   Ratio   Lower  Upper
Constant    -2.14637  0.577245  -3.72  0.000
Spending    0.341643  0.128672   2.66  0.008    1.41    1.09   1.81
Card         1.09873  0.444696   2.47  0.013    3.00    1.25   7.17


Log-Likelihood = -60.487
Test that all slopes are zero: G = 13.628, DF = 2, P-Value = 0.001
```

We can now use equation (15.33) to estimate the probability of making a purchase for a particular type of customer. For example, to estimate the probability of making a purchase for customers that spend €2000 annually and do not have a Stamm credit card, we substitute $X_1 = 2$ and $X_2 = 0$ into equation (15.33).

$$\hat{y} = \frac{e^{-2.1464 + 0.3416(2) + 1.0987(0)}}{1 + e^{-2.1464 + 0.3416(2) + 1.0987(0)}} = \frac{e^{-1.4632}}{1 + e^{-1.4632}} = \frac{0.2315}{1.2315} = 0.1880$$

Thus, an estimate of the probability of making a purchase for this particular group of customers is approximately 0.19. Similarly, to estimate the probability of making a purchase for customers that spent €2000 last year and have a Stamm credit card, we substitute $X_1 = 2$ and $X_2 = 1$ into equation (15.33).

$$\hat{y} = \frac{e^{-2.1464 + 0.3416(2) + 1.0987(1)}}{1 + e^{-2.1464 + 0.3416(2) + 1.0987(1)}} = \frac{e^{-0.3645}}{1 + e^{-0.3645}} = \frac{0.6945}{1.6945} = 0.4099$$

Thus, for this group of customers, the probability of making a purchase is approximately 0.41. It appears that the probability of making a purchase is much higher for customers with a Stamm credit card. Before reaching any conclusions, however, we need to assess the statistical significance of our model.

Testing for significance

Testing for significance in logistic regression is similar to testing for significance in multiple regression. First we conduct a test for overall significance. For the Stamm Stores example, the hypotheses for the test of overall significance follow:

$$H_0\colon \beta_1 = \beta_2 = 0$$
$$H_1\colon \beta_1 \text{ and/or } \beta_2 \text{ is not equal to zero}$$

The test for overall significance is based upon the value of a G test statistic. This is commonly referred to as the 'Deviance Statistic'. If the null hypothesis is true, the sampling distribution of G follows a chi-square distribution with degrees of freedom equal to the number of independent variables in the model. Although the computation of G is beyond the scope of the book, the value of G and its corresponding p-value are provided as part of MINITAB's binary logistic regression output. Referring to the last line in Figure 15.13, we see that the value of G is 13.628, its degrees of freedom are 2, and its p-value is 0.001. Thus, at any level of significance $\alpha \geq 0.001$, we would reject the null hypothesis and conclude that the overall model is significant.

If the G test shows an overall significance, a z test can be used to determine whether each of the individual independent variables is making a significant contribution to the overall model. For the independent variables X_i, the hypotheses are

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

If the null hypothesis is true, the value of the estimated coefficient divided by its standard error follows a standard normal probability distribution. The column labelled Z in the MINITAB output contains the values of $z_i = b_i / s_{b_i}$ for each of the estimated coefficients and the column labelled p contains the corresponding p-values. The z_i ratio is also known as a 'Wald Statistic'. Suppose we use $\alpha = 0.05$ to test for the significance of the independent variables in the Stamm model. For the independent variable X_1 the z value is 2.66 and the corresponding p-value is 0.008. Thus, at the 0.05 level of significance we can reject H_0: $\beta_1 = 0$. In a similar fashion we can also reject H_0: $\beta_2 = 0$ because the p-value corresponding to $z = 2.47$ is 0.013. Hence, at the 0.05 level of significance, both independent variables are statistically significant.

Managerial use

We now use the estimated logistic regression equation to make a decision recommendation concerning the Stamm Stores catalogue promotion. For Stamm Stores, we already computed

$$P(Y = 1 | X_1 = 2, X_2 = 1) = 0.4099 \quad \text{and} \quad P(Y = 1 | X_1 = 2, X_2 = 0) = 0.1880$$

These probabilities indicate that for customers with annual spending of €2000 the presence of a Stamm credit card increases the probability of making a purchase using the discount coupon. In Table 15.12 we show estimated probabilities for values of annual spending ranging from €1000 to €7000 for both customers who have a Stamm credit card and customers who do not have a Stamm credit card. How can Stamm use this information to better target customers for the new promotion? Suppose Stamm wants to send the promotional catalogue only to customers who have a 0.40 or higher probability of making a purchase. Using the estimated probabilities in Table 15.12, Stamm promotion strategy would be:

Customers who have a Stamm credit card: Send the catalogue to every customer that spent €2000 or more last year.

Customers who do not have a Stamm credit card: Send the catalogue to every customer that spent €6000 or more last year.

Looking at the estimated probabilities further, we see that the probability of making a purchase for customers who do not have a Stamm credit card, but spend €5000 annually is 0.3921. Thus, Stamm may want to consider revising this strategy by including those customers who do not have a credit card as long as they spent €5000 or more last year.

Table 15.12 Estimated probabilities for Stamm Stores

		Annual spending						
		€1000	€2000	€3000	€4000	€5000	€6000	€7000
Credit card	Yes	0.3305	0.4099	0.4943	0.5790	0.6593	0.7314	0.7931
	No	0.1413	0.1880	0.2457	0.3143	0.3921	0.4758	0.5609

Interpreting the logistic regression equation

Interpreting a regression equation involves relating the independent variables to the business question that the equation was developed to answer. With logistic regression, it is difficult to interpret the relation between the independent variables and the probability that $Y = 1$ directly because the logistic regression equation is nonlinear. However, statisticians have shown that the relationship can be interpreted indirectly using a concept called the odds ratio.

The **odds in favour of an event occurring** is defined as the probability the event will occur divided by the probability the event will not occur. In logistic regression the event of interest is always $Y = 1$. Given a particular set of values for the independent variables, the odds in favour of $Y = 1$ can be calculated as follows:

$$\text{Odds} = \frac{P(Y = 1 | X_1, X_2, \ldots X_y)}{P(Y = 0 | X_1, X_2, \ldots X_y)} = \frac{P(Y = 1 | X_1, X_2, \ldots X_y)}{1 - P(Y = 1 | X_1, X_2, \ldots X_y)} \qquad \textbf{(15.34)}$$

$$\ldots$$

The **odds ratio** measures the impact on the odds of a one-unit increase in only one of the independent variables. The odds ratio is the odds that $Y = 1$ given that one of the independent variables has been increased by one unit (odds$_1$) divided by the odds that $Y = 1$ given no change in the values for the independent variables (odds$_0$).

Odds ratio

$$\text{Odds ratio} = \frac{\text{odds}_1}{\text{odds}_0} \qquad \textbf{(15.35)}$$

For example, suppose we want to compare the odds of making a purchase for customers who spend 2000 annually and have a Stamm credit card ($X_1 = 2$ and $X_2 = 1$) to the odds of making a purchase for customers who spend €2000 annually and do not have a Stamm credit card ($X_1 = 2$ and $X_2 = 0$). We are interested in interpreting the effect of a one-unit increase in the independent variable X_2. In this case

$$\text{Odds}_1 = \frac{P(Y = 1 | X_1 = 2, X_2 = 1)}{1 - P(Y = 1 | X_1 = 2, X_2 = 1)}$$

and

$$\text{Odds}_0 = \frac{P(Y = 1 | X_1 = 2, X_2 = 0)}{1 - P(Y = 1 | X_1 = 2, X_2 = 0)}$$

Previously we showed that an estimate of the probability that $Y = 1$ given $X_1 = 2$ and $X_2 = 1$ is 0.4099, and an estimate of the probability that $Y = 1$ given $X_1 = 2$ and $X_2 = 0$ is 0.1880. Thus,

$$\text{Estimate of odds}_1 = \frac{0.4099}{1 - 0.4099} = 0.6946$$

and

$$\text{Estimate of odds}_0 = \frac{0.1880}{1 - 0.1880} = 0.2315$$

The estimated odds ratio is

$$\text{Estimated odds ratio} = \frac{0.6946}{0.2315} = 3.00$$

Thus, we can conclude that the estimated odds in favour of making a purchase for customers who spent €2000 last year and have a Stamm credit card are three times greater than the estimated odds in favour of making a purchase for customers who spent €2000 last year and do not have a Stamm credit card.

The odds ratio for each independent variable is computed while holding all the other independent variables constant. But it does not matter what constant values are used for the other independent variables. For instance, if we computed the odds ratio for the Stamm credit card variable (X_2) using €3000, instead of €2000, as the value for the annual spending variable (X_1), we would still obtain the same value for the estimated odds ratio (3.00). Thus, we can conclude that the estimated odds of making a purchase for customers who have a Stamm credit card are three times greater than the estimated odds of making a purchase for customers who do not have a Stamm credit card.

The odds ratio is standard output for logistic regression software packages. Refer to the MINITAB output in Figure 15.13. The column with the heading Odds Ratio contains the estimated odds ratios for each of the independent variables. The estimated odds ratio for X_1 is 1.41 and the estimated odds ratio for X_2 is 3.00. We already showed how to interpret the estimated odds ratio for the binary independent variable X_2. Let us now consider the interpretation of the estimated odds ratio for the continuous independent variable X_1.

The value of 1.41 in the Odds Ratio column of the MINITAB output tells us that the estimated odds in favour of making a purchase for customers who spent €3000 last year is 1.41 times greater than the estimated odds in favour of making a purchase for customers who spent €2000 last year. Moreover, this interpretation is true for any one-unit change in X_1.

For instance, the estimated odds in favour of making a purchase for someone who spent €5000 last year is 1.41 times greater than the odds in favour of making a purchase for a customer who spent €4000 last year. But suppose we are interested in the change in the odds for an increase of more than one unit for an independent variable. Note that X_1 can range from 1 to 7. The odds ratio as printed by the MINITAB output does not answer this question.

To answer this question we must explore the relationship between the odds ratio and the regression coefficients.

A unique relationship exists between the odds ratio for a variable and its corresponding regression coefficient. For each independent variable in a logistic regression equation it can be shown that

$$\text{Odds ratio} = e^{b_i}$$

To illustrate this relationship, consider the independent variable X_1 in the Stamm example. The estimated odds ratio for X_1 is

$$\text{Estimated odds ratio} = e^{b_1} = e^{0.3416} = 1.41$$

Similarly, the estimated odds ratio for X_2 is

$$\text{Estimated odds ratio} = e^{b_2} = e^{1.0987} = 3.00$$

This relationship between the odds ratio and the coefficients of the independent variables makes it easy to compute estimates of the odds ratios once we develop estimates of the model parameters. Moreover, it also provides us with the ability to investigate changes in the odds ratio of more than or less than one unit for a continuous independent variable.

The odds ratio for an independent variable represents the change in the odds for a one unit change in the independent variable holding all the other independent variables constant. Suppose that we want to consider the effect of a change of more than one unit, say c units. For instance, suppose in the Stamm example that we want to compare the odds of making a purchase for customers who spend €5000 annually ($X_1 = 5$) to the odds of making a purchase for customers who spend €2000 annually ($X_1 = 2$). In this case $c = 5 - 2 = 3$ and the corresponding estimated odds ratio is

$$e^{cb_i} = e^{3(0.3416)} = e^{1.0248} = 2.79$$

This result indicates that the estimated odds of making a purchase for customers who spend €5000 annually is 2.79 times greater than the estimated odds of making a purchase for customers who spend €2000 annually. In other words, the estimated odds ratio for an increase of €3000 in annual spending is 2.79.

In general, the odds ratio enables us to compare the odds for two different events. If the value of the odds ratio is 1, the odds for both events are the same. Thus, if the independent variable we are considering (such as Stamm credit card status) has a positive impact on the probability of the event occurring, the corresponding odds ratio will be greater than 1. Most logistic regression software packages provide a confidence interval for the odds ratio. The MINITAB output in Figure 15.13 provides a 95 per cent confidence interval for each of the odds ratios. For example, the point estimate of the odds ratio for X_1 is 1.41 and the 95 per cent confidence interval is 1.09 to 1.81. Because the confidence interval does not contain the value of 1, we can conclude that X_1, has a significant effect on the odds ratio. Similarly, the 95 per cent confidence interval for the odds ratio for X_2 is 1.25 to 7.17. Because this interval does not contain the value of 1, we can also conclude that X_2 has a significant effect on the odds ratio.

Logit transformation

An interesting relationship can be observed between the odds in favour of $Y = 1$ and the exponent for e in the logistic regression equation. It can be shown that

$$\ln(\text{odds}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

This equation shows that the natural logarithm of the odds in favour of $Y = 1$ is a linear function of the independent variables. This linear function is called the **logit**. We will use the notation $g(x_1, x_2, \ldots x_p)$ to denote the logit.

Logit

$$g(x_1, x_2, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad \textbf{(15.36)}$$

Substituting $g(x_1, x_2, \ldots x_p)$ for $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$ in equation (15.28), we can write the logistic regression equation as

$$E(Y) = \frac{e^{g(x_1, x_2, \ldots, x_p)}}{1 + e^{g(x_1, x_2, \ldots, x_p)}} \qquad (15.37)$$

Once we estimate the parameters in the logistic regression equation, we can compute an estimate of the logit. Using $\hat{g}\,(x_1, x_2 \ldots x_p)$ to denote the **estimated logit**, we obtain

Estimated logit

$$\hat{g}\,(x_1, x_2, \ldots x_p) = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p \qquad (15.38)$$

Therefore, in terms of the estimated logit, the estimated regression equation is

$$\hat{y} = \frac{e^{b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p}}{1 + e^{b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p}} = \frac{e^{\hat{g}\,(x_1, x_2, \ldots, x_p)}}{1 + e^{\hat{g}(x_1, x_2, \ldots, x_p)}} \qquad (15.39)$$

For the Stamm Stores example, the estimated logit is

$$\hat{g}\,(x_1, x_2) = -\,2.1464 + 0.3416x_1 + 1.0987x_2$$

and the estimated regression equation is

$$y = \frac{e^{\hat{g}(x_1, x_2)}}{1 + e^{\hat{g}(x_1, x_2)}} = \frac{e^{-2.1464 + 0.3416x_1 + 1.0987x_2}}{1 + e^{-2.1464 + 0.3416x_1 + 1.0987x_2}}$$

Therefore, because of the unique relationship between the estimated logit and the estimated logistic regression equation, we can compute the estimated probabilities for Stamm Stores by dividing $e^{\hat{g}\,(x_1, x_2)}$ by $1 + e^{\hat{g}\,(x_1, x_2)}$.

Exercises

Applications

30 Refer to the Stamm Stores example introduced in this section. The dependent variable is coded as $Y = 1$ if the customer makes a purchase and 0 if not.

Suppose that the only information available to help predict whether the customer will make a purchase is the customer's credit card status, coded as $X = 1$ if the customer has a Stamm credit card and $X = 0$ if not.

a. Write the logistic regression equation relating X to Y.

b. What is the interpretation of $E(Y)$ when $X = 0$?

c. For the Stamm data in Table 15.11, use MINITAB to compute the estimated logit.

d. Use the estimated logit computed in part (c) to compute an estimate of the probability of making a purchase for customers who do not have a Stamm credit card and an estimate of the probability of making a purchase for customers who have a Stamm credit card.

e. What is the estimate of the odds ratio? What is its interpretation?

31 In Table 15.12 we provided estimates of the probability of a purchase in the Stamm Stores catalogue promotion. A different value is obtained for each combination of values for the independent variables.

a. Compute the odds in favour of a purchase for a customer with annual spending of €4000 who does not have a Stamm credit card ($X_1 = 4, X_2 = 0$).

b. Use the information in Table 15.12 and part (a) to compute the odds ratio for the Stamm credit card variable X_2 holding annual spending constant at $X_1 = 4$.

c. In the text, the odds ratio for the credit card variable was computed using the information in the €2000 column of Table 15.12. Did you get the same value for the odds ratio in part (b)?

32 Community Bank would like to increase the number of customers who use payroll direct deposit. Management is considering a new sales campaign that will require each branch manager to call each customer who does not currently use payroll direct deposit. As an incentive to sign up for payroll direct deposit, each customer contacted will be offered free banking for two years. Because of the time and cost associated with the new campaign, management would like to focus their efforts on customers who have the highest probability of signing up for payroll direct deposit. Management believes that the average monthly balance in a customer's current account may be a useful predictor of whether the customer will sign up for direct payroll deposit. To investigate the relationship between these two variables, Community Bank tried the new campaign using a sample of 50 current account customers that do not currently use payroll direct deposit. The sample data show the average monthly current account balance (in hundreds of euros) and whether the customer contacted signed up for payroll direct deposit (coded 1 if the customer signed up for payroll direct deposit and 0 if not). The data are contained in the data set named Bank; a portion of the data follows.

BANK

Customer	X Monthly balance	Y Direct deposit
1	1.22	0
2	1.56	0
3	2.10	0
4	2.25	0
5	2.89	0
6	3.55	0
7	3.56	0
8	3.65	1
.	.	.
.	.	.
.	.	.
48	18.45	1
49	24.98	0
50	26.05	1

a. Write the logistic regression equation relating X to Y.

b. For the Community Bank data, use MINITAB to compute the estimated logistic regression equation.

c. Conduct a test of significance using the G test statistic. Use $\alpha = 0.05$.

d. Estimate the probability that customers with an average monthly balance of €1000 will sign up for direct payroll deposit.

e. Suppose Community Bank only wants to contact customers who have a 0.50 or higher probability of signing up for direct payroll deposit. What is the average monthly balance required to achieve this level of probability?

f. What is the estimate of the odds ratio? What is its interpretation?

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

Summary

In this chapter, we introduced multiple regression analysis as an extension of simple linear regression analysis presented in Chapter 14. Multiple regression analysis enables us to understand how a dependent variable is related to two or more independent variables. The regression equation $E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$ shows that the expected value or mean value of the dependent variable Y is related to the values of the independent variables $X_1, X_2, \ldots, X_p$. Sample data and the least squares method are used to develop the estimated regression equation $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$. In effect $b_0, b_1, b_2, \ldots, b_p$ are sample statistics used to estimate the unknown model parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ Computer printouts were used throughout the chapter to emphasize the fact that statistical software packages are the only realistic means of performing the numerous computations required in multiple regression analysis.

The multiple coefficient of determination was presented as a measure of the goodness of fit of the estimated regression equation. It determines the proportion of the variation of Y that can be explained by the estimated regression equation. The adjusted multiple coefficient of determination is a similar measure of goodness of fit that adjusts for the number of independent variables and thus avoids overestimating the impact of adding more independent variables. Model assumptions for multiple regression are shown to parallel those for simple regression analysis.

An F test and a t test were presented as ways of determining statistically whether the relationship among the variables is significant. The F test is used to determine whether there is a significant overall relationship between the dependent variable and the set of all independent variables. The t test is used to determine whether there is a significant relationship between the dependent variable and an individual independent variable given the other independent variables in the regression model. Correlation among the independent variables, known as multicollinearity, was discussed.

The section on qualitative independent variables showed how dummy variables can be used to incorporate qualitative data into multiple regression analysis. The section on residual analysis showed how residual analysis can be used to validate the model assumptions, detect outliers and identify influential observations. Standardized residuals, leverage, studentized deleted residuals and Cook's distance measure were discussed. The chapter concluded with a section on how logistic regression can be used to model situations in which the dependent variable may only assume two values.

Key terms

Adjusted multiple coefficient of determination
Cook's distance measure
Dummy variable
Estimated logistic regression equation
Estimated logit
Estimated multiple regression equation
Influential observation
Least squares method
Leverage
Logistic regression equation
Logit
Multicollinearity
Multiple coefficient of determination
Multiple regression analysis
Multiple regression equation
Multiple regression model
Odds in favour of an event occurring
Odds ratio
Outlier
Qualitative independent variable
Studentized deleted residuals
Variance inflation factor

Key formulae

Multiple regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \qquad (15.1)$$

Multiple regression equation

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad (15.2)$$

Estimated multiple regression equation

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p \qquad (15.3)$$

Least squares criterion

$$\min \Sigma (y_i - \hat{y}_i)^2 \qquad (15.4)$$

Relationship among SST, SSR and SSE

$$SST = SSR + SSE \qquad (15.7)$$

Multiple coefficient of determination

$$R^2 = \frac{SSR}{SST} \qquad (15.8)$$

Adjusted multiple coefficient of determination

$$\text{adj } R^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1} \qquad (15.9)$$

Mean square regression

$$MSR = \frac{SSR}{p} \qquad (15.12)$$

Mean square error

$$MSE = s^2 = \frac{SSE}{n-p-1} \qquad (15.13)$$

***F* test statistic**

$$F = \frac{MSR}{MSE} \qquad (15.14)$$

***t* test statistic**

$$t = \frac{b_i}{s_{b_i}} \qquad (15.15)$$

Variance Inflation Factor

$$VIF(X_j) \quad \frac{1}{1 - R^2_j} \qquad (15.16)$$

Standardized residual for observation *i*

$$\frac{y_i - \hat{y}_i}{s_{y_i - \hat{y}_i}} \qquad (15.24)$$

Standard deviation of residual *i*

$$s_{y_i - \hat{y}_i} = s\sqrt{1 - h_i} \qquad (15.25)$$

Cook's distance measure

$$D_i = \frac{(y_i - \hat{y}_i)^2 h_i}{(p-1)s^2 (1 - h_i)^2} \qquad (15.26)$$

Logistic regression equation

$$E(Y) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p}} \qquad (15.28)$$

Interpretation of *E*(Y) as a probability in logistic regression

$$E(Y) = P(Y = 1 | x_1, x_2, \ldots x_p) \qquad (15.29)$$

Estimated logistic regression equation

$$\hat{y} = \text{estimate of } P(Y = 1 \mid x_1, x_2, \ldots x_p) = \frac{e^{b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p}}{1 + e^{b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p}} \qquad (15.31)$$

Odds ratio

$$\text{Odds ratio} = \frac{\text{odds}_1}{\text{odds}_0} \qquad (15.35)$$

Logit

$$g(x_1, x_2, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad (15.36)$$

Estimated logits

$$\hat{g}\,(x_1, x_2, \ldots x_p) = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p \qquad (15.38)$$

Case problem Consumer Research

Consumer Research is an independent agency that conducts research on consumer attitudes and behaviours for a variety of firms. In one study, a client asked for an investigation of consumer characteristics that tend to be used to predict the amount charged by credit card users. Data were collected on annual income, household size and annual credit card charges for a sample of 50 consumers. The following data are on the CD accompanying the text in the data set named Consumer.

Managerial report

1 Use methods of descriptive statistics to summarize the data. Comment on the findings.

2 Develop estimated regression equations, first using annual income as the independent variable and then using household size as the independent variable. Which variable is the better predictor of annual credit card charges? Discuss your findings.

3 Develop an estimated regression equation with annual income and household size as the independent variables. Discuss your findings.

4 What is the predicted annual credit card charge for a three-person household with an annual income of €40 000?

5 Discuss the need for other independent variables that could be added to the model.

What additional variables might be helpful?

CONSUMER

Income (€000s)	Household size	Amount charged (€)	Income (€000s)	Household size	Amount charged (€)
54	3	4016	54	6	5573
30	2	3159	30	1	2583
32	4	5100	48	2	3866
50	5	4742	34	5	3586
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
42	2	3020	46	5	4820
41	7	4828	66	4	5149

Shopper paying for purchase with a credit card. © Marcus Clackson.

Software Section for Chapter 15

Multiple regression using MINITAB

In this section we show how MINITAB can be used to model multiple regression problems using data for the Eurodistributor Company. First, the data must be entered in a MINITAB worksheet. The distances are entered in column C1, the number of deliveries are entered in column C2, and the travel times (hours) are entered in column C3. The variable names Distance, Deliveries and Time are entered as the column headings on the worksheet. In subsequent steps, we refer to the data by using the variable names Distance, Deliveries and Time. The steps involved in using MINITAB to produce the regression results shown in Figure 15.4 follow.

EURODIS-TRIBUTOR

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

Step 2 Enter Time in the **Response** box [**Regression** panel]
Enter Distance and Deliveries in the **Predictors** box
Click **OK**

Step 3 Click **OK** [**Regression** panel]

Logistic regression using MINITAB

MINITAB calls logistic regression with a dependent variable that can only assume the values 0 and 1 Binary Logistic Regression. In this section we describe the steps required to use MINITAB's Binary Logistic Regression procedure to generate the computer output for the Stamm Stores problem shown in Figure 15.13. First, the data must be entered in a MINITAB worksheet. The amounts customers spent last year at Stamm (in thousands of euros) are entered into column C2, the credit card data (1 if a Stamm card; 0 otherwise) are entered into column C3, and the purchase data (1 if the customer made a purchase; 0 otherwise) are entered in column C4. The variable names Spending, Card and Purchase are entered as the column headings on the worksheet. In subsequent steps, we refer to the data by using the variable names Spending, Card and Purchase. The steps involved in using MINITAB to generate the logistic regression output follow.

STAMM

Step 1 **Regression > Binary Logistic Regression** [Main menu bar]

Step 2 Enter Purchase in the **Response** box [**Binary Logistic Regression** panel]
Enter Spending and Card in the **Model** box
Click **OK**

Step 3 Click **OK** [**Regression** panel]

Figure 15.4 PASW output for Eurodistributor with two independent variables

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.951[a]	.904	.876	.5731

a. Predictors: (Constant), Deliveries, Distance

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	21.601	2	10.800	32.878	.000[a]
	Residual	2.299	7	.328		
	Total	23.900	9			

a. Predictors: (Constant), Deliveries, Distance

b. Dependent Variable: Time

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.869	.952		-.913	.392
	Distance	.061	.010	.735	6.182	.000
	Deliveries	.923	.221	.496	4.176	.004

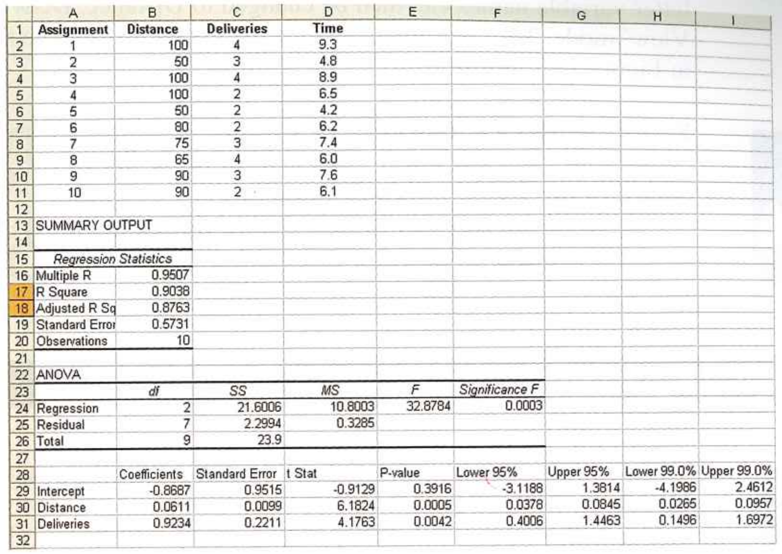
a. Dependent Variable: Time

Multiple regression using EXCEL

EXCEL

In Section 15.2 we discussed the computer solution of multiple regression problems by showing MINITAB's output for the Eurodistributor Company problem. In this section we describe how to use EXCEL's Regression tool to develop the estimated multiple regression equation for the Eurodistributor problem. Refer to Figure 15.14 as we describe the tasks involved. First, the labels Assignment, Distance, Deliveries and Time are entered into cells A1:D1 of the worksheet, and the sample data into cells B2:D11. The numbers 1–10 in cells A2:A11 identify each observation. The steps involved in using the Regression tool for multiple regression analysis follow.

Figure 15.15 EXCEL output for Eurodistributor with two independent variables



	A	B	C	D	E	F	G	H	I
1	Assignment	Distance	Deliveries	Time					
2	1	100	4	9.3					
3	2	50	3	4.8					
4	3	100	4	8.9					
5	4	100	2	6.5					
6	5	50	2	4.2					
7	6	80	2	6.2					
8	7	75	3	7.4					
9	8	65	4	6.0					
10	9	90	3	7.6					
11	10	90	2	6.1					
12									
13	SUMMARY OUTPUT								
14									
15	*Regression Statistics*								
16	Multiple R	0.9507							
17	R Square	0.9038							
18	Adjusted R Sq	0.8763							
19	Standard Error	0.5731							
20	Observations	10							
21									
22	ANOVA								
23		*df*	*SS*	*MS*	*F*	*Significance F*			
24	Regression	2	21.6006	10.8003	32.8784	0.0003			
25	Residual	7	2.2994	0.3285					
26	Total	9	23.9						
27									
28		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 99.0%	Upper 99.0%
29	Intercept	-0.8687	0.9515	-0.9129	0.3916	-3.1188	1.3814	-4.1986	2.4612
30	Distance	0.0611	0.0099	6.1824	0.0005	0.0378	0.0845	0.0265	0.0957
31	Deliveries	0.9234	0.2211	4.1763	0.0042	0.4006	1.4463	0.1496	1.6972
32									

EURODIS-TRIBUTOR

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

Step 2 Enter D1:D11 in the **Input Y Range** box [**Regression** panel]
Enter B1:C11 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**

In the EXCEL output shown in Figure 15.14 the label for the independent variable X_1 is Distance (see cell A30), and the label for the independent variable X_2 is Deliveries (see cell A31). The estimated regression equation is

$$\hat{y} = -0.8687 + 0.0611x_1 + 0.9234x_2$$

Note that using EXCEL's Regression tool for multiple regression is almost the same as using it for simple linear regression. The major difference is that in the multiple regression case a larger range of cells is required in order to identify the independent variables.

Note that Logistic Regression is not a standard analysis feature of EXCEL.

Multiple regression using PASW

PASW

First, the data must be entered in a PASW worksheet. In 'Data View' mode, distances are entered in rows 1 to 10 of the leftmost column. This is automatically labelled by the system V1. Similarly the number of deliveries and travel times 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 Distance, Deliveries and Time in 'Variable View' mode. The steps involved in using PASW to produce the regression results shown in Figure 15.4 follow.

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

Step 2 Enter Time in the **Dependent** box [**Linear** panel]
Enter Distance and Deliveries in the **Independent(s)** box
Click **OK**

Logistic regression using PASW

First, the data must be entered in a PASW worksheet in Data View mode. The amounts customers spent last year at Stamm (in thousands of euros) are entered into rows 1 to 100 of the leftmost column. Corresponding credit card details (1 if a Stamm card; 0 otherwise) and purchase data (1 if the customer made a purchase; 0 otherwise) are entered into the immediately adjacent columns to the right. The system automatically assigns headings V1, V2 and V3 to these columns but these can be easily changed to Spending, Card and Purchase in Variable View mode. The following command sequence will generate the logistic regression output.

Step 1 **Analyze > Regression > Binary Logistic** Select the menubar item

Step 2 Enter Purchase in the **Dependent** box [**Logistic Regression** panel]
Enter Spending and Card in the **Covariates** box
Click **OK**
Click **OK**

Selective output is as follows:

Classification Table[a]

	Observed		Predicted		
			Purchase		Percentage Correct
			0	1	
Step 1	Purchase	0	52	8	86.7
		1	20	20	50.0
	Overall Percentage				72.0

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Spending	.342	.129	7.050	1	.008	1.407
	Card	1.099	.445	6.105	1	.013	3.000
	Constant	-2.146	.577	13.826	1	.000	.117

a. Variable(s) entered on step 1: Spending, Card.

Chapter 16

Regression Analysis: Model Building