MODULE A

Multiple Regression Analysis

Dennis L. Young

ARIZONA STATE UNIVERSITY

TO ACCOMPANY

INTRODUCTORY STATISTICS

SIXTH EDITION

Neil A. Weiss

Addison Wesley

Boston San Francisco New York
London Toronto Sydney Tokyo Singapore Madrid
Mexico City Munich Paris Cape Town Hong Kong Montreal

Composition: Windfall Software

CONTENTS

module A

Multiple Regression Analysis

MODULE OUTLINE

GENERAL OBJECTIVES Chapters 14 and 15 examined descriptive and inferential methods for **simple linear regression** in which one predictor variable is used to predict a response variable. There we learned how to estimate population regression equations, test hypotheses about regression parameters, and analyze residuals to determine whether the regression model assumptions are valid.

In this module we will discuss extending the regression model to include more than one predictor variable. We will learn how to estimate the population regression equation, conduct tests of hypotheses about regression parameters, and assess model assumptions using residuals. We will also extend the coefficient of determination, r^2, to the case of more than one predictor. To illustrate our discussion of multiple regression analysis, we will return to the Orion data and add an additional predictor variable, mileage, in an effort to predict the price of a used Orion.

case study

AUTOMOBILE INSURANCE RATES

What affects automobile insurance rates? Responsible drivers carry automobile insurance to protect themselves financially if involved in an automobile accident or in case of theft. When we purchase automobile insurance we know that various factors affect the rate we pay. Such factors as our age, previous driving record, and type of car insured have an effect on our insurance premium.

However, there are other factors that play a role in determining automobile insurance rates that are not specifically associated with an individual driver. For example, how often are people killed in automobile accidents in the area in which we live? What is the rate of automobile theft in our area? How expensive is it to be cared for in a hospital? How much time is spent driving to and from work? How densely populated is the area in which we live?

In order to investigate the effect of these factors on automobile insurance rates, information on each of the 50 states was obtained from the 117th edition (1997) of the *Statistical Abstract of the United States* on the following six variables:

1. Average automobile insurance rate (1995)
2. Population density (1996)
3. Automobile theft rate (1995)
4. Automobile deaths per 100 million miles driven (1995)
5. Average drive time to work (1995)
6. Average cost of a day's stay in a hospital (1995)

The data for all the states can be found in Table A.3 on page A-83.

After we study multiple linear regression analysis in this module, we will be able to provide an equation that can be used to predict the average automobile insurance rate based on the five variables listed above in 2–6. We will be able to determine whether there are useful associations between the average automobile insurance rate and the five predictor variables. Finally, we will be able to assess how well our equation fits the data.

A.1 THE MULTIPLE LINEAR REGRESSION MODEL

In simple linear regression, we considered the problem of predicting a **response variable** with a single **predictor variable.** Now we will develop the method of **multiple linear regression** for the situation in which there are two or more predictor variables.

LINEAR EQUATIONS IN TWO INDEPENDENT VARIABLES

In order to understand multiple linear regression, we need to give a brief discussion of **linear equations** in two or more independent variables. We consider the case of two independent variables so that we can graphically display results. The general form of a linear equation with two independent variables can be written as

$$y = b_0 + b_1x_1 + b_2x_2,$$

where b_0, b_1, and b_2 are constants, x_1 and x_2 are the independent variables, and y is the dependent variable.

The graph of a linear equation with two independent variables is a **plane** in three-dimensional space. Linear equations in two (or more) independent variables have many applications in business and in the social, behavioral, and physical sciences. We illustrate the use of a linear equation in two independent variables by extending Example 14.1.

Example A.1 *Linear Equations with Two Independent Variables: CJ^2 Business Services*

Recall Example 14.1. CJ^2 Business Services does word processing at a rate of \$20/hr plus a one-time \$25 disk charge. Suppose also that it charges an additional \$10 for each graphical illustration. The total cost of a job depends on the amount of time (in hours) it takes to finish the job and the number of illustrations. Find the equation that gives the total cost for a job.

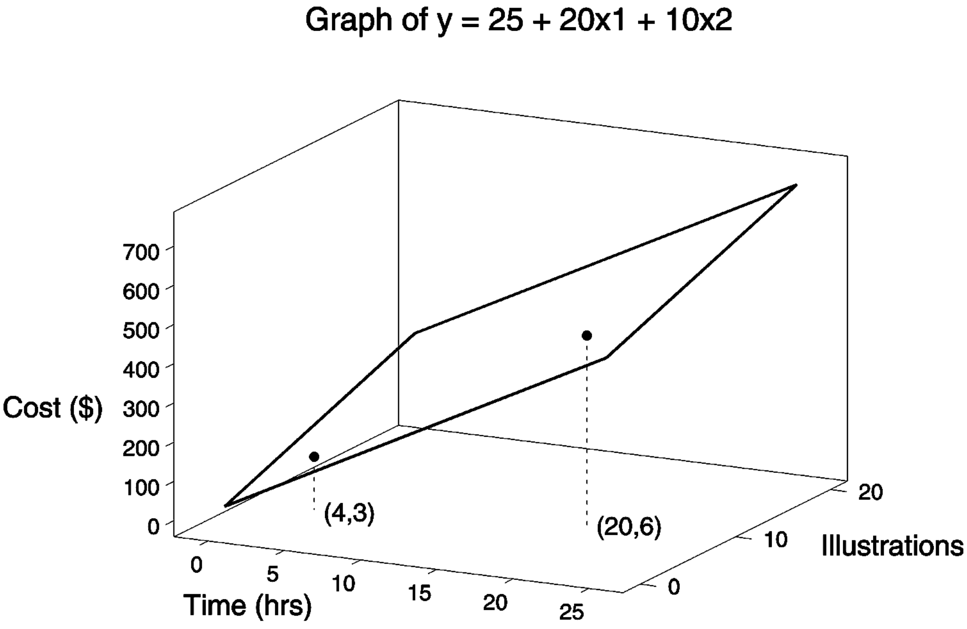
Solution Let x_1 denote the number of hours required to complete the job, x_2 denote the number of illustrations, and y denote the total cost to the customer. A job that takes x_1 hours with x_2 illustrations will cost \$20·x_1 plus \$10·x_2 plus the \$25 disk charge. So the total cost is

$$y = 25 + 20x_1 + 10x_2.$$

This equation for total cost, $y = 25 + 20x_1 + 10x_2$, is a linear equation. Here $b_0 = 25$, $b_1 = 20$, and $b_2 = 10$. This equation gives us the exact cost if we are given the hours required and number of illustrations for a job. For example, if a job takes 4 hours and has 3 illustrations, the cost is $y = 25 + 20 \cdot 4 + 10 \cdot 3 = \$135$; a job that takes 20 hours with 6 illustrations costs $y = 25 + 20 \cdot 20 + 10 \cdot 6 = \$485$.

The graph of the equation $y = 25 + 20x_1 + 10x_2$ is given in Fig. A.1. Values of time and number of illustrations are shown in the plane making up the "floor" of the graph. Time (x_1) is along the axis in the front and number of illustrations

Figure A.1
Graph of the plane
$y = 25 + 20x_1 + 10x_2$



(x_2) is along the axis going to the rear of the "floor." This "floor" is the time-illustrations (or x_1-x_2) plane. The value of cost is on the vertical axis rising perpendicularly from the time-illustrations (x_1-x_2) plane. The height above the "floor" represents the cost (y) of the job.

For example, for a job taking 4 hours with 3 illustrations, we find the point $(4, 3)$ in the "floor" and then go up a vertical distance (measured in dollars) to a height of $135 to graph the point on the plane giving the cost (y) for this job. The point is illustrated by the solid dot positioned directly above the point $(4, 3)$ on the "floor." In a similar fashion, a job that takes 20 hours with 6 illustrations for a cost of $485 is illustrated by the solid dot rising above the point $(20, 6)$ at a height of $485. All the points representing the cost of a job, as determined by $y = 25 + 20x_1 + 10x_2$, lie on a plane that is partially shown in Fig. A.1 as the region bounded by the heavy solid lines. ◆

INTERCEPT AND SLOPES

For the linear equation in two independent variables, $y = b_0 + b_1x_1 + b_2x_2$, the constants b_0, b_1, and b_2 have useful geometric interpretations. The number b_0 is the value of y when $x_1 = 0$ and $x_2 = 0$, and it gives the value of y where the plane intersects the y-axis. The number b_1 measures the steepness of the plane along the x_1 direction, and b_2 measures the steepness of the plane along the x_2 direction. The number b_1 tells us how much y changes when x_1 increases by 1 unit, while x_2 is held fixed. The number b_2 tells us how much y changes when x_2 increases by 1 unit, while x_1 is held fixed.

DEFINITION A.1 y-Intercept and Partial Slopes

For a linear equation, $y = b_0 + b_1x_1 + b_2x_2$, the number b_0 is called the ***y-intercept,*** and the numbers b_1 and b_2 are called the ***partial slopes*** for the independent variables x_1 and x_2, respectively. ■

Example A.2 Interprets y-Intercept and Partial Slopes: CJ^2 Business Services

In Example A.1 the linear equation that gives the total cost, y, of a word-processing job in terms of the number of hours, x_1, and the number of illustrations, x_2, is

$$y = 25 + 20x_1 + 10x_2.$$

a. Find the y-intercept and partial slopes of this linear equation.

b. Interpret the y-intercept and partial slopes in terms of the graph of the equation.

c. Interpret the y-intercept and partial slopes in terms of word-processing costs.

Solution

a. The y-intercept for the equation is $b_0 = 25$, the partial slope for x_1 is $b_1 = 20$, and the partial slope for x_2 is $b_2 = 10$.

b. The y-intercept $b_0 = 25$ is the value of y at which the plane $y = 25 + 20x_1 + 10x_2$ intersects the y-axis. The partial slope $b_1 = 20$ indicates that y increases by 20 units for every increase in x_1 of 1 unit, while x_2 is held constant. The partial slope $b_2 = 10$ indicates that y increases by 10 units for every increase in x_2 of 1 unit, while x_1 is held constant.

c. The y-intercept $b_0 = 25$ is the total cost of a job that takes 0 hours and has 0 illustrations. Thus the y-intercept of $25 is the fixed cost that is charged no matter how long a job takes or how many illustrations it has. The partial slope $b_1 = 20$ represents the increase in total cost of a job for an increase of 1 hour in time, provided the number of illustrations does not change. The partial slope $b_2 = 10$ represents the increase in the total cost of a job for an increase of one illustration, provided the time for the job does not change. ◆

LINEAR EQUATION—MORE THAN TWO INDEPENDENT VARIABLES

In our word processing example, it is not difficult to think of other services that would increase the cost of a job. There would be additional cost if special paper were used or if extra copies were desired. One could easily think of a **linear equation in** k **independent variables,** $x_1, x_2, \ldots, x_k$, to describe the total cost of a job. We would write

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k.$$

The graph of this linear equation is a (hyper-) plane in $k + 1$ dimensions. The number b_0 represents the y-intercept, that is, the value of y when $x_1 = x_2 = \cdots = x_k = 0$. The coefficient b_i ($i = 1, \ldots, k$) is the partial slope of x_i, holding all other xs fixed. So b_i tells us the change in y for a unit increase in x_i, holding all other xs fixed.

It is not possible to draw the graph of the plane describing the values of y when there are three or more independent variables. The inclusion of more independent variables results in additional complexity in our understanding of how y is related to the independent variables. This will become evident when we attempt to visualize the relationship between a response variable y and more than two predictor variables.

MULTIPLE LINEAR REGRESSION MODEL

In simple linear regression, where there is only one predictor variable, x, the regression model specifies that the **conditional mean** of y at a given value of the predictor, x, can be represented by $\beta_0 + \beta_1 x$, where β_0 and β_1 are regression parameters. This means that, on the average, the relationship between y and x is described by a straight line. The parameter β_0 is the y-intercept of the line, and β_1 is the slope, or the change in the conditional mean of y for a unit change in x. We saw in Chapter 14 that this model was useful in predicting the price of a car (the Orion) based on its age.

In many cases there are two or more variables that are useful in predicting a response variable y. For example, the price of a used car might depend on its mileage in addition to its age.

Suppose there are two predictor variables, x_1 and x_2. The **multiple linear regression model** incorporates the two predictors by adding together the terms $\beta_1 x_1$ and $\beta_2 x_2$. This model states that the conditional mean of y at given values of x_1 and x_2 is $\beta_0 + \beta_1 x_1 + \beta_2 x_2$, for regression parameters β_0, β_1, and β_2. It means that, on the average, the relationship between y and x_1 and x_2 is described by a plane in three dimensions. That is, the conditional means of y fall on a plane given by the equation $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

The parameter β_0 is the y-intercept for the plane. The parameter β_1 gives the change in the conditional mean of y for a unit change in x_1, while x_2 is held fixed. Likewise, the parameter β_2 gives the change in the conditional mean of y for a unit change in x_2, while x_1 is held fixed.

The parameters β_1 and β_2 are sometimes called the *partial regression coefficients* since they are like the partial slopes for linear equations. The equation $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ is called the **population multiple linear regression equation** or the **population regression equation.**

Example A.3 *Illustrates the Multiple Linear Regression Model: Orion Prices*

Consider the problem of predicting the price of an Orion. In Chapters 14 and 15, the age (in years) of an Orion is used to predict its price. Suppose we have additional information about each Orion in the form of the number of miles that the car has been driven. We will measure mileage in 1000 mile units. Develop the multiple linear regression model to predict the price of an Orion as a function of the two predictor variables age and mileage.

Solution Let the price of an Orion be denoted by y, the age by x_1, and the mileage by x_2. Then the multiple linear regression model expresses the conditional mean of y at an age of x_1 and a mileage of x_2 as $\beta_0 + \beta_1 x_1 + \beta_2 x_2$, where β_0 is the y-intercept, β_1 is the partial regression coefficient for x_1, and β_2 is the partial regression coefficient for x_2.

The parameter β_0 represents the mean value of price when both age and mileage are zero, that is, for a new Orion. The parameter β_1 gives the change in the conditional mean of the price for a change of one year in the age (x_1) of an Orion, holding the value of mileage (x_2) constant. The parameter β_2 gives the change in the conditional mean of the price for a change of one unit (1000 miles) in the mileage (x_2), holding the value of age (x_1) constant. Geometrically, this model specifies that conditional mean values of price (y) lie on a plane given by the equation $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. ◆

There may be more than two predictor variables in a multiple linear regression model. When there are k predictor variables $x_1, \ldots, x_k$, the multiple linear regression model gives the conditional mean of y at given values of $x_1, \ldots, x_k$ as $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$. In this case the conditional means of y lie on a (hyper-) plane in $k+1$ dimensions. For $i = 1, 2, \ldots, k$, the population regression coefficient β_i is the change in the conditional mean of y for an increase in x_i of one unit while all other predictor variables remain fixed. β_0 is the y-intercept.

Exercises A.1

Statistical Concepts and Skills

A.1 Regarding linear equations in two or more independent variables:

a. Write the general form of such an equation.
b. In your expression in part (a), which letters represent constants and which represent variables?
c. In your expression in part (a), which letters represent the independent variables and which letter represents the dependent variable?

A.2 Fill in the blanks.

a. The graph of a linear equation with two independent variables is a ______.
b. The graph of a linear equation with $k > 2$ independent variables is a ______.

A.3 Consider a linear equation $y = b_0 + b_1 x_1 + b_2 x_2$.

a. Identify and give the geometric interpretation of b_0.
b. Identify and then give the geometric interpretation of b_1 and b_2.

A.4 Answer true or false to each of the following statements and explain your answers.

a. The plane representing the graph of a linear equation in two independent variables slopes upward unless both partial slopes are zero.
b. The value of the y-intercept has no effect on how the value of y changes as x_1 and x_2 increase.

A.5 Car Rental Cost. To rent a car, the Acme Rental Company charges \$39.95 per day plus \$0.20 per mile. There is also a \$20 surcharge if the car is rented at the airport. Let y be the total cost of a rental, x_1 the number of days rented, and x_2 the number of miles driven.

a. If the car is rented from Acme at the airport, obtain the equation that expresses y in terms of x_1 and x_2.
b. What are the values of b_0, b_1, and b_2?

c. How much does the total cost change if a car is rented an additional day, but the mileage driven does not change?

d. What is the cost of renting an Acme car for 5 days and driving 500 miles?

A.6 Air Conditioning Service. Encore Air Conditioning charges $36 per hour plus a $30 service charge. If refrigerant is needed in the air conditioner, there is a $20 per pound charge. Let y denote the total cost to the customer, x_1 the number of hours it takes for the job, and x_2 the amount of refrigerant needed (in pounds).

a. Obtain the equation that expresses y in terms of x_1 and x_2.

b. Find b_0, b_1, and b_2.

c. What is the change in total cost for a job if the number of hours for the job does not change, but the amount of refrigerant increases by one pound?

d. What is the cost of a job that lasts 2 hours and requires 5 pounds of refrigerant?

A.7 Banquet Room Rental. The banquet room at the Saguaro Steak House can be rented at a rate of $40 per hour with a one time $75 clean up charge. In addition there is a $16 per person charge for the restaurant's buffet dinner, and a $5 per drink beverage charge. Let x_1 be the number of hours the banquet room is rented, x_2 the number of people who have the buffet dinner, x_3 the number of drinks ordered, and y the total cost of holding a reception.

a. Obtain the equation that expresses y in terms of x_1, x_2, and x_3.

b. Find b_0, b_1, b_2, and b_3.

c. What is the change in the total cost of a reception if the number of hours and the number of drinks ordered does not change, but the number of people having the buffet dinner increases by one?

d. What is the cost of a reception that lasts 4 hours and is attended by 50 people who have the buffet dinner and order 100 drinks?

A.8 Bakery Products. The Mill Avenue Bakery packages its products in three sizes: one-half pound, one pound, and two pound packages. In determining the total weight of the bakery's products loaded on a delivery truck, let x_1 denote the number of one-half-pound packages, x_2 the number of one-pound packages, and x_3 the number of two-pound packages. Let y denote the total weight of bakery packages on the truck.

a. Obtain the equation that expresses y in terms of x_1, x_2, and x_3.

b. Find b_0, b_1, b_2, and b_3.

c. What is the change in the total weight for an increase of one two-pound package, provided the number of one-half- and one-pound packages does not change?

d. Find the total weight of the bakery's product on the delivery truck if there are 100 one-half-pound packages, 200 one-pound packages, and 30 two-pound packages.

In each of Exercises A.9–A.12,

a. determine the y-intercept and partial slopes of the specified linear equation.

b. explain what the y-intercept and partial slopes represent in terms of the graph of the equation.

c. explain what the y-intercept and partial slopes represent in terms relating to the application.

A.9 $y = 20 + 39.95x_1 + 0.20x_2$ (from Exercise A.5)

A.10 $y = 30 + 36x_1 + 20x_2$ (from Exercise A.6)

A.11 $y = 75 + 40x_1 + 16x_2 + 5x_3$ (from Exercise A.7)

A.12 $y = 0.5x_1 + x_2 + 2x_3$ (from Exercise A.8)

In each of Exercises A.13–A.22, you are given the linear equation of a plane. For each exercise,

a. find the y-intercept and the partial slopes.

b. determine whether the plane slopes upward, slopes downward, or is horizontal in the x_1 axis direction, when x_2 is fixed at a specified value.

A.13 $y = 3 + 4x_1 + 7x_2$

A.14 $y = -1 + 2x_1 - 5x_2$

A.15 $y = 6 - 7x_1 + 10x_2$

A.16 $y = -8 - 4x_1 + 3x_2$

A.17 $y = -2 + 3x_1$

A.18 $y = 15 - 6x_1$

A.19 $y = 7 + 3x_2$

A.20 $y = 12 - 4x_2$

A.21 $y = 9$

A.22 $y = -16$

In each of Exercises A.23–A.30, we have identified the y-intercept and the partial slopes of a plane. For each exercise,

a. determine whether the plane slopes upward, slopes downward, or is horizontal in the x_2 axis direction, when x_1 is fixed at a specified value.

b. find the equation of the plane.

A.23 $b_0 = 5, b_1 = 2, b_2 = 7$

A.24 $b_0 = -3, b_1 = 4, b_2 = 6$

A.25 $b_0 = -2, b_1 = 3, b_2 = -7$

A.26 $b_0 = 6, b_1 = -2, b_2 = -1$

A.27 $b_0 = 0, b_1 = -0.05, b_2 = 1.5$

A.28 $b_0 = -6, b_1 = 5, b_2 = 0$

A.29 $b_0 = 0, b_1 = 0, b_2 = 3$

A.30 $b_0 = 12, b_1 = 0, b_2 = -5$

Extending the Concepts and Skills

A.31 Why is it often preferable to use more than one predictor variable in a regression analysis?

A.32 **Grade Prediction.** The Statistics Department at a large university would like to predict the final grades of students in its introductory statistics course.

a. If the final grade is measured as a percentage from 0% to 100%, suggest five predictor variables that might be useful in predicting final grade.
b. Write the multiple linear regression model for the conditional mean of the response variable, final grade.

A.33 **Gasoline Mileage.** An engineer wants to determine factors that affect the EPA miles per gallon rating of new automobiles available for sale in the United States.

a. Suggest five predictor variables that might be useful in predicting the miles per gallon rating.
b. Write the multiple linear regression model for the conditional mean of the response variable, miles per gallon.

A.34 **Blood Pressure Medication.** A medical researcher wants to determine factors that affect a patient's response to a medication that is designed to reduce blood pressure.

a. What would you select as the response variable in this study?
b. Suggest five predictor variables that might be useful in prediction of the response variable.
c. Write the multiple linear regression model for the conditional mean of the response variable.

A.35 **Infant Mortality Rate.** A social scientist wants to predict the infant mortality rate in cities in the United States.

a. Suggest five predictor variables (characteristics associated with a city) that might be useful in predicting infant mortality rate.
b. Write the multiple linear regression model for the conditional mean of the response variable, infant mortality rate.

A.2 ESTIMATION OF THE REGRESSION PARAMETERS

Usually, for real-life applications, we are not able to determine exactly the value of the variable we are trying to predict even when we know the values of the predictor variables. For example, in the case of the price of an Orion, there are many factors that affect the price that are not accounted for by age and mileage. We do not have an exact relationship as we did for the cost of a word-processing job, as described in Examples A.1 and A.2.

When we attempt to understand the relationship between a response variable y and its predictor variables, it is useful to try to visualize the relationship by plotting the data. In simple linear regression, we plot the response variable against the single predictor variable in a scatterplot or scatter diagram. In multiple linear regression, we plot the response variable against each of the predictor variables. It is also helpful to plot each predictor variable against all other predictor variables. The collection of all of these scatterplots is usually displayed in an array called the **scatterplot matrix.**

We will see later that the plots of the predictors versus other predictors are useful in identifying relationships that might exist among the predictors. In Module B we will discover that relationships among the predictor variables can make it difficult to interpret the regression coefficients.

Example A.4 *Scatterplot Matrix: Orion Prices*

Table A.1 displays the price, age, and mileage for a sample of 11 Orions. As mentioned in Section 14.2, these data were obtained from the *Auto Trader* magazine. Price is in hundreds of dollars, age is in years, and mileage is in thousands of miles. Investigate the relationship among the variables by constructing the scatterplot matrix of the data.

Solution For this data set, the scatterplot matrix is shown in Fig. A.2 on the next page. Here we have in the first row of the array (going from left to right) the plot of price versus age and the plot of price versus mileage. In the second row, we have the plot of age versus price and the plot of age versus mileage. The third row contains the plot of mileage versus price and the plot of mileage versus age. Each diagonal entry of the array contains the name of the variable associated with the vertical axis of the plots in the row (left and right) in which it appears, and with the horizontal axis of the column (up and down) in which it appears.

The scatterplot matrix provides the plot of price versus age that we saw in Fig. 14.7. This plot shows the fairly linear relationship between price and age. The plot of price versus miles also shows a fairly linear relationship between price and mileage. The point in the upper left of these two plots might not fit the linear relationship as well as we would like. This point was discussed in Chapter 14 as being an influential observation in trying to predict price from the age of the car. We will investigate this point later.

The plot of age versus miles also shows a fairly linear relationship. This is not unexpected since the older a car is, the more miles it will have been driven, on the average.

◆

In the case of the Orion example, there are only two predictor variables (age and mileage), and we are able to construct a three-dimensional scatterplot of the data points. This plot is shown in Fig. A.3.

Table A.1 Data on age, miles driven, and price for 11 Orions

Car	Age (yrs)	Miles (1000)	Price ($100s)
	x_1	x_2	y
1	5	57	85
2	4	40	103
3	6	77	70
4	5	60	82
5	5	49	89
6	5	47	98
7	6	58	66
8	6	39	95
9	2	8	169
10	7	69	70
11	7	89	48

Figure A.2
Scatterplot matrix of the Orion data

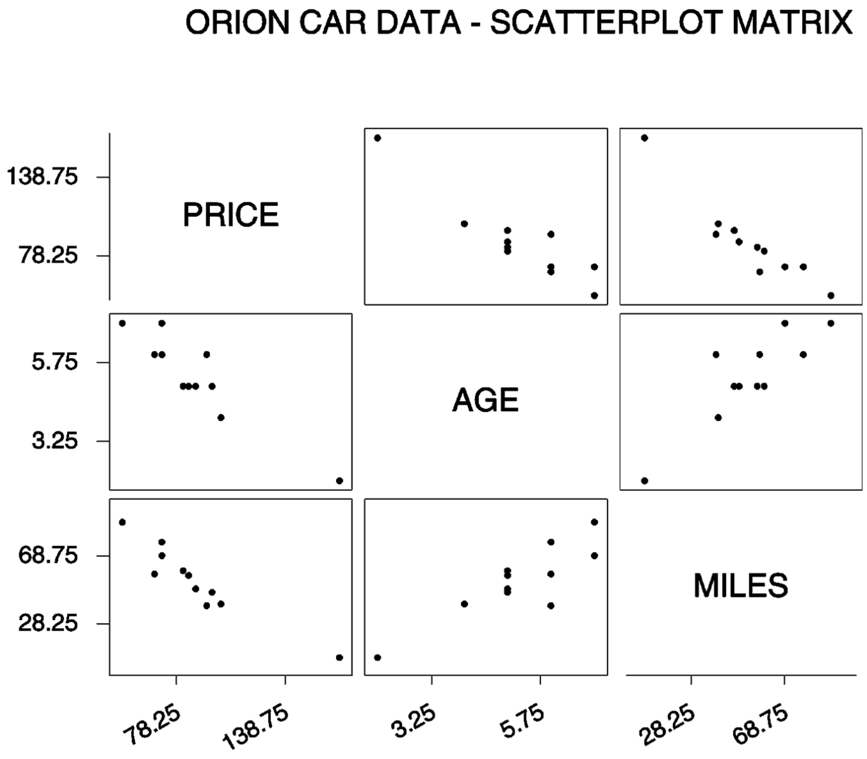
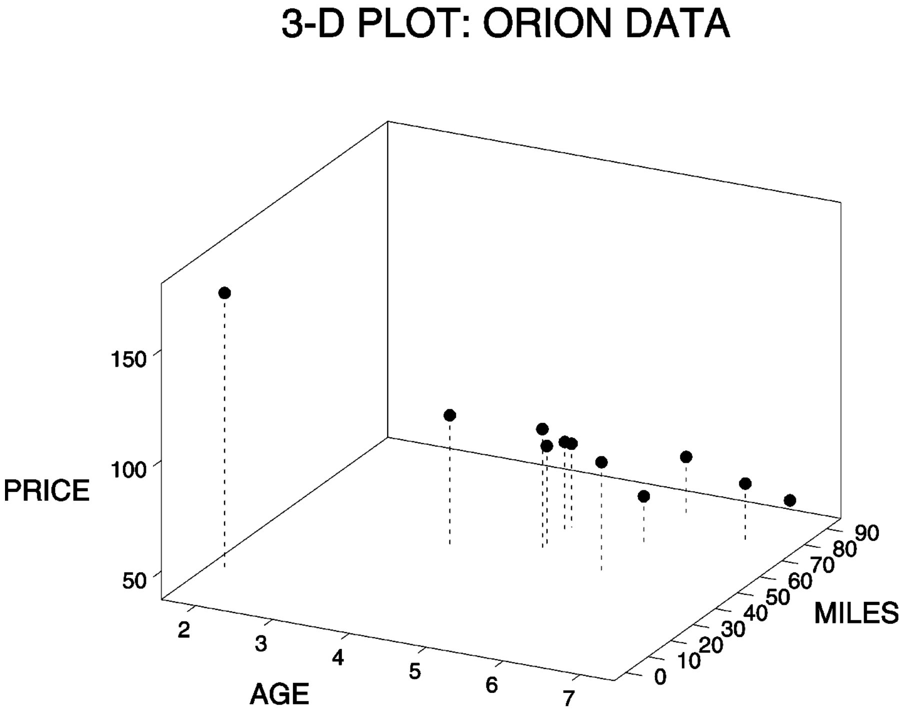


Figure A.3
Three-dimensional plot of the Orion data



The data points are plotted in the same manner as points were plotted in the three-dimensional plot shown in Fig. A.1 on page A-5. For example, the first Orion has an age of 5 years, a mileage of 57,000 and a price of $8500. The point (5, 57) is found in the age-miles plane and then a solid dot is placed above this point at a height of 85.

Figure A.3 shows the data points in three dimensions and is useful in judging whether a plane might be appropriate in relating price to age and mileage.

ESTIMATING REGRESSION PARAMETERS WITH THE METHOD OF LEAST SQUARES

Generally the values of the population regression coefficients ($\beta_0, \beta_1, \ldots, \beta_k$) are unknown and must be estimated from a sample of data points. As in the case of simple linear regression, we will use the **method of least squares** to estimate the regression coefficients from the sample. To find the least squares estimators of $\beta_0, \beta_1, \ldots, \beta_k$, we need to find coefficients $b_0, b_1, \ldots, b_k$ that minimize the sum of squared differences between the observed values of y and the values of y predicted by $b_0 + b_1x_1 + \cdots + b_kx_k$. Thus we minimize

$$\Sigma(y - (b_0 + b_1x_1 + \cdots + b_kx_k))^2.$$

The **sample regression equation** obtained by the method of least squares is written as

$$\hat{y} = b_0 + b_1x_1 + \cdots + b_kx_k,$$

where $\hat{y}$ is the predicted value of y for given values of $x_1, x_2, \ldots, x_k$.

Geometrically, the sample regression (hyper-) plane is the plane that minimizes the sum of squared vertical differences between the observed values of y and the values of y predicted by a regression plane. The computation of the estimated regression coefficients is rather complex, and formulas for their values are cumbersome without using matrix notation. We do not give such formulas here. Calculations of the estimated regression coefficients are generally done by using statistical packages such as Minitab or SAS.

Example A.5 *Obtaining a Sample Regression Equation with Several Predictors: Orion Prices*

In Table A.1 on page A-11 we presented data on age, miles driven, and price for a sample of 11 Orions. We employed Minitab to perform a multiple regression analysis on this data set with age and miles driven as predictor variables for price. In doing so, we instructed Minitab to provide the least detailed output available, as seen in Printout A.1. The procedure for obtaining the Minitab output will be given in Example A.6.

a. Use Printout A.1 to obtain the sample regression equation.

b. Use the regression equation to predict the price of an Orion that is 5 years old and has been driven 52,000 miles.

Printout A.1
Minitab regression output for Orion data

```
The regression equation is
PRICE = 183 - 9.50 AGE - 0.821 MILES

Predictor        Coef     SE Coef          T        P
Constant       183.04       11.35      16.13    0.000
AGE            -9.504       3.874      -2.45    0.040
MILES         -0.8215      0.2552      -3.22    0.012

S = 8.805       R-Sq = 93.6%    R-Sq(adj) = 92.0%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         2      9088.3      4544.2     58.61    0.000
Residual Error     8       620.2        77.5
Total             10      9708.5
```

Solution **a.** The sample regression equation is displayed in the second line of the computer output in Printout A.1: `PRICE = 183 - 9.50 AGE - 0.821 MILES`. Thus the sample regression equation is $\hat{y} = 183 - 9.50x_1 - 0.821x_2$, where x_1 denotes age, in years, x_2 denotes miles driven, in thousands, and $\hat{y}$ denotes predicted price, in hundreds of dollars.

b. For a 5-year-old Orion with 52,000 miles, we have $x_1 = 5$ and $x_2 = 52$. The predicted price for such a car is

$$\hat{y} = 183 - 9.50 \cdot 5 - 0.821 \cdot 52 = 92.81,$$

or $9281. ◆

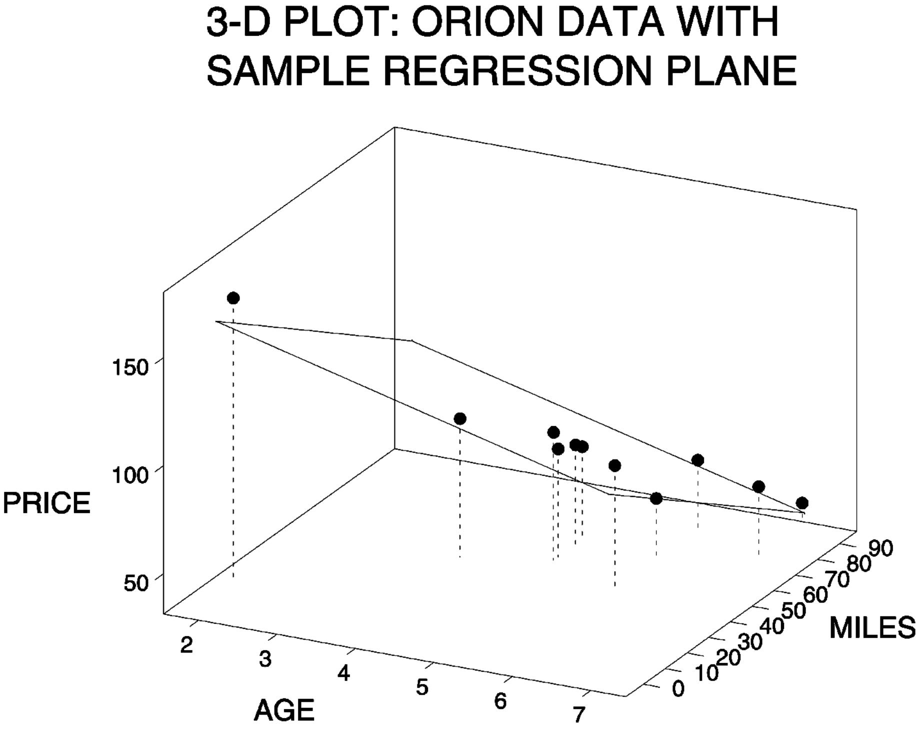
The sample regression equation determined in Example A.5 for the price of an Orion based on the data in Table A.1 is graphed in Fig. A.4 on the next page. This graph shows the data points and the sample regression plane. Visually, we see that the sample regression plane fits the data reasonably well.

A WARNING ON THE USE OF MULTIPLE LINEAR REGRESSION

The multiple linear regression model is based on the assumption that the data points are scattered about a plane in $(k + 1)$-dimensional space, that is, the conditional means of y are represented by a plane in the predictor variables $x_1, x_2, \ldots, x_k$. There are many situations in which the data points are scattered about a curved surface and the multiple linear regression model in the predictor variables $x_1, x_2, \ldots, x_k$ is not appropriate. It is still possible to use the method of least squares to find a best-fitting plane, but this fit may be useless in giving adequate predictions of the response variable.

It is essential that we plot the response variable versus each of the predictor variables and look for indications of curvature in these plots. If there is curva-

Figure A.4 Three-dimensional plot of the Orion data with sample regression plane



ture present in any of these plots, we need to consider using regression models that account for this curvature. Among the standard methods for dealing with curved surfaces are (1) using quadratic or higher-degree polynomial terms in the appropriate predictors in the regression equation, or (2) transforming the response variable in such a way as to "straighten" the relationship between the transformed response and the predictor variables. Sometimes both methods need to be applied in a regression analysis to obtain an adequate regression model.

Detecting curvature in the plots of response versus predictors is not always easy to do, because we are trying to view data in $k + 1$ dimensions by looking at simple two-dimensional plots. It is difficult to understand all the relationships that might exist between the response and predictor variables from these plots.

Consider, for example, the shadow that you cast on the ground or on a wall on a sunny day. You are a three-dimensional object that is being projected by the sunlight to a two-dimensional object, your shadow. How much you might be able to determine about yourself from your shadow depends on the position of the sun and your orientation with respect to the sun. For many possible shadows it would be impossible to determine how many arms or legs you have, or whether you have a nose.

Similar problems arise when higher dimensional data are projected down to two or three dimensions. We cannot recover all the features of the data from these lower dimensional projections.

Later in this module we will see that it is extremely useful to look at diagnostic residual plots to assess whether an appropriate regression model has been determined. In Module B we will discuss the strategy of using transformations of the response and predictor variables to obtain more useful regression models when a plane is not an adequate model.

The Technology Center

Performing a Multiple Linear Regression

Many statistical technologies have programs that will automatically perform a multiple regression analysis, but others do not. For instance, Minitab and the Excel add-in DDXL have such a program, while the TI-83 Plus does not.

For a statistical technology that does not have a dedicated program for performing multiple regression analysis, it is often possible to use the macro capabilities of the statistical technology to write a program. Writing such a macro for multiple linear regression would require knowledge of solving systems of linear equations and matrix algebra, and will not be considered here.

We now present output and (optional) step-by-step instructions to implement such programs.

Example A.6 Using Technology to Perform a Multiple Linear Regression: Orion Prices

Table A.1 on page A-11 displays data on age, miles driven, and price for a sample of 11 Orions. Use Minitab or Excel to perform the multiple regression analysis relating the response variable price to the predictor variables age and miles driven.

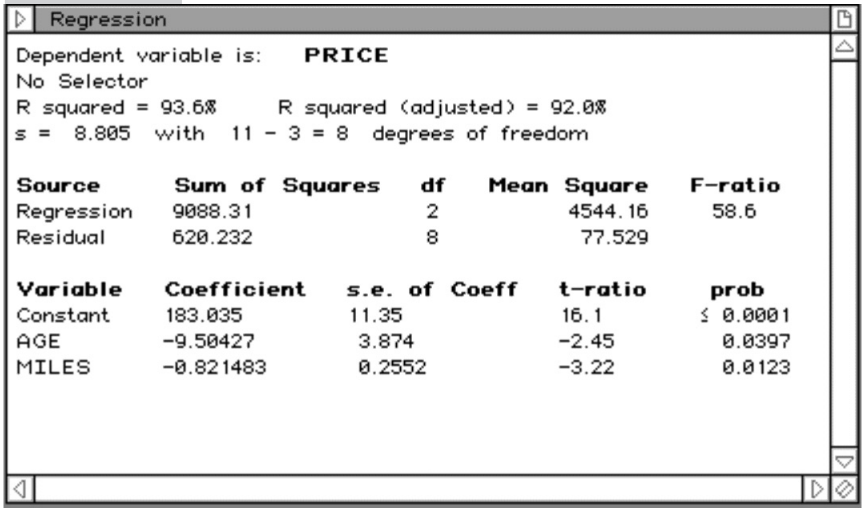
Solution Printout A.1 on page A-14 displays the multiple linear regression output from Minitab. Printout A.2 displays the output from Excel using the DDXL add-in. Both technologies give similar output, although the sample regression equation is not specifically stated in the output from DDXL. In DDXL, the *y*-intercept is given in the column labeled Coefficient and row labeled Constant. The sample regression coefficients for the predictor variables, age and miles driven, are given in the column labeled Coefficient in Printout A.2. ◆

Obtaining the Output (Optional)

Printout A.1 provides output from Minitab and Printout A.2 provides output from the Excel add-in DDXL for a multiple linear regression based on the data for age, miles driven, and price for a sample of 11 Orions in Table A.1. Here are detailed instructions for obtaining that output. First we store the data on age, miles driven, and price in columns or ranges, named AGE, MILES, and PRICE, respectively. Then we proceed as follows.

Printout A.2
Multiple linear regression output for Orion data

EXCEL



```
Dependent variable is:    PRICE
No Selector
R squared = 93.6%     R squared (adjusted) = 92.0%
s =  8.805  with  11 - 3 = 8  degrees of freedom

Source        Sum of Squares    df    Mean Square    F-ratio
Regression    9088.31            2        4544.16     58.6
Residual      620.232            8         77.529

Variable     Coefficient      s.e. of Coeff    t-ratio     prob
Constant     183.035          11.35            16.1        ≤ 0.0001
AGE          -9.50427          3.874           -2.45         0.0397
MILES        -0.821483         0.2552          -3.22         0.0123
```

MINITAB

1. Choose **Stat ➤ Regression ➤ Regression...**
2. Specify PRICE in the **Response** text box
3. Click in the **Predictors** text box and specify AGE and MILES
4. Click the **Results...** button
5. Select the **Regression equation, table of coefficients, s, R-squared, and basic analysis of variance** option button
6. Click **OK**
7. Click **OK**

EXCEL

1. Choose **DDXL ➤ Regression**
2. Select **Multiple Regression** from the **Function type** drop-down box
3. Specify PRICE in the **Response Variable** text box
4. Specify AGE and MILES in the **Explanatory Variables** text box
5. Click **OK**

Obtaining a Scatterplot Matrix

All statistical technologies have programs for obtaining a scatterplot of one variable versus another. However Minitab is the only technology featured here that provides a scatterplot matrix as shown in Figure A.2 on page A-12 for the Orion data set in Table A.1.

Obtaining the Output (Optional)

Figure A.2 on page A-12 provides the scatterplot matrix output from Minitab for the data on age, miles driven, and price for a sample of 11 Orions from Table A.1 on page A-11. Here are detailed instructions for obtaining that output. First we store the data on age, miles driven, and price in columns named AGE, MILES, and PRICE, respectively. Then we proceed as follows.

MINITAB

1 Choose **Graph ➤ Matrix Plot...**
2 In the **Graph variables** text box, specify PRICE, AGE, and MILES.
3 Click **OK**

Obtaining a Three-Dimensional Scatterplot

Some statistical technologies have programs for three-dimensional scatterplots while others do not. Minitab has such a program, but the TI-83 Plus does not. DDXL has the capability of providing a rotating three-dimensional plot of the data. This rotating three-dimensional plot can be very useful in understanding the relationship among three variables. We are able to view the three-dimensional scatterplot from "all sides" when the plot is rotated.

We now present output and (optional) step-by-step instructions to implement such programs.

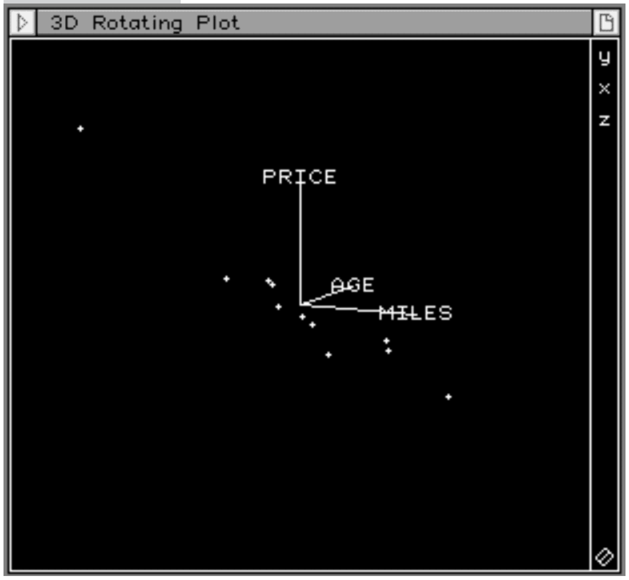
Example A.7 Using Technology to Obtain a Three-Dimensional Scatterplot: Orion Prices

Return to the Orion data in Example A.6. Use Minitab or Excel to obtain a three-dimensional scatterplot of price against age and miles driven.

Solution Figure A.3 on page A-12 displays the Minitab output of the three-dimensional scatterplot of price against age and miles driven. Printout A.3 displays the output from the DDXL add-in to Excel. ◆

Printout A.3
Three-dimensional scatterplot for price, age and miles driven

EXCEL



The scatterplot from Minitab looks quite different from that produced by DDXL. Minitab's scatterplot has detail about the scales on the three axes that the DDXL scatterplot does not have. However, we must remember that the Minitab plot is a "static" plot meant for presentation, while the DDXL plot is a "dynamic" plot that is meant for exploration and finding patterns in the data via rotation. The DDXL plot shown in Printout A.3 is only one of many possible plots that occur when the three-dimensional scatterplot is rotated in DDXL.

Obtaining the Output (Optional)

Here are detailed instructions for obtaining the three-dimensional scatterplot output in Fig. A.3 from Minitab, and Printout A.3 from DDXL. To begin, we store the Orion data as previously described. Then we do the following.

MINITAB

1 Choose **Graph ➤ 3D Plot...**
2 In the **Graph Variables Z** text box specify PRICE
3 In the **Graph Variables Y** text box specify AGE
4 In the **Graph Variables X** text box specify MILES
5 In the **Data display** text boxes in the column labeled **Display,** specify **Project** for **Item2,** and in the column labeled **For each,** specify **Point** for **Item 2.**
6 Click **OK**

EXCEL

1 Choose **DDXL ➤ Charts and Plots**
2 Select **3D Rotating Plot** from the **Function type** drop-down box
3 Specify MILES in the **x-Axis Variable** text box
4 Specify AGE in the **y-Axis Variable** text box
5 Specify PRICE in the **z-Axis Variable** text box
6 Click **OK**

Exercises A.2

Statistical Concepts and Skills

A.36 Regarding a scatterplot matrix:

a. Identify two of its uses.
b. What property should the scatterplots of the response versus each predictor variable have in order to proceed to obtain the regression plane for the data?

A.37 Regarding the criterion used to decide on the plane that best fits a set of data points in $k + 1$ dimensions:

a. What is that criterion called?
b. Specifically, what is the criterion?

A.38 Regarding the plane that best fits a set of data points:

a. What is that plane called?
b. What is the equation of that plane called?

A.39 Regarding the variables in a multiple linear regression analysis:

a. What is the dependent variable called?
b. What are the independent variables called?

A.40 Answer true or false to the following statements and explain your answers.

a. The sample regression coefficients exactly equal the population regression coefficients.
b. It is possible to determine all the features of high dimensional data by projecting the data to two-dimensional plots.
c. The sample multiple linear regression equation cannot be used to predict the value of the response variable at

values of the predictor variables for which we have no observations.

In each of Exercises A.41 and A.42,

a. construct tables giving the values of $x_1, x_2, y, \hat{y}, e = y - \hat{y}$, and e^2 similar to Table 14.4.
b. determine which of the two planes fits the set of data points better according to the least squares criterion.

A.41 Plane A: $y = 2 + 3x_1 + x_2$
Plane B: $y = 3 + 4x_1 + 2x_2$

x_1	1	1	2	2
x_2	1	2	1	2
y	8	10	11	16

A.42 Plane A: $y = 3 - 2x_1 + 3x_2$
Plane B: $y = 4 - x_1 + 2x_2$

x_1	0	0	1	2	2
x_2	1	2	0	1	3
y	7	10	0	1	9

A.43 Advertising and Sales. A household-appliance manufacturer wants to analyze the relationship between total sales and the company's three primary means of advertising. The first three columns of the table below provide the expenditures on advertising, by type, for each of 10 randomly selected sales periods. The fourth column contains the total sales. All data are in millions of dollars. We used Minitab to perform a multiple regression analysis on the data using the variables television, magazine, and radio advertising expenditures as predictor variables for sales. The computer output is shown in Printouts A.4 and A.5 on page A-23.

Television x_1	**Magazines** x_2	**Radio** x_3	**Sales** y
8.3	4.4	6.1	361.1
6.3	4.2	4.9	344.0
9.9	5.9	6.3	377.9
9.4	3.3	6.1	371.5
10.4	2.7	5.2	365.4
9.0	3.5	5.1	364.5
9.2	4.1	6.0	372.9
10.6	4.8	6.4	379.4
9.3	4.2	5.5	362.6
10.5	6.0	5.9	387.5

a. Use the scatterplot matrix in Printout A.4 to assess whether a multiple linear regression model might be appropriate for predicting sales.
b. Use the computer output in Printout A.5 to obtain the sample regression equation for sales in terms of television, magazine, and radio advertising expenditures.
c. Apply the sample regression equation to predict total sales if the amounts spent on television, magazine, and radio advertising are $9.5 million, $4.3 million, and $5.2 million, respectively.

A.44 Corvette Prices. The data on age and price for 10 Corvettes from Exercise 14.43 are repeated in the first and third columns of the table shown below. The second column of the table displays the number of miles, in thousands, that each of the 10 Corvettes has been driven.

Age x_1	**Miles** x_2	**Price** y
6	36	205
6	36	195
6	36	210
2	22	340
2	5	299
5	31	230
4	22	270
5	39	243
1	9	340
4	27	240

We used Minitab to perform a multiple regression analysis on the data, with age and miles as predictor variables for price. The computer output is shown in Printouts A.6 and A.7 on page A-24.

a. Use the scatterplot matrix in Printout A.6 to assess whether a multiple linear regression model might be appropriate for predicting the price of a Corvette.
b. Use the output in Printout A.7 to obtain the sample regression equation for price in terms of age and miles.
c. Apply the sample regression equation to predict the price of a 4-year-old Corvette that has been driven 28,000 miles.

A.45 Graduation Rates. Graduation rates and what influences them have become a concern in U.S. colleges and universities. *U.S. News and World Report*'s "1997 College Guide" provides data on graduation rates for colleges and universities as a function of a number of predictor variables. We consider only the following three predictor variables: student-to-faculty ratio, percentage of freshmen in the top 10% of their high-school class, and percentage of applicants accepted. (Here *graduation rate* refers to the percentage of entering freshmen, attending full time, that graduate within 5 years.) Printouts A.8 and A.9 on page A-25 show the results of a multiple regression analysis for a sample of 48 schools.

Student/ faculty ratio	Percentage in top 10% of HS class	Percent accepted	Graduation rate
8	90	11	97
6	92	12	94
11	85	31	93
7	92	18	93
15	87	16	94
11	94	24	89
12	88	20	93
7	83	30	88
8	88	19	91
3	98	26	82
7	84	21	87
8	82	44	93
11	84	32	91
7	81	33	89
8	76	40	88
7	76	58	81
8	88	24	90
9	66	51	86
13	80	40	93
10	62	60	82
12	79	23	90
12	79	33	91
9	65	47	73
11	63	32	88
17	95	36	80
15	61	68	84
14	74	37	82
9	59	53	81
21	97	39	77
12	69	42	85
11	69	52	77
12	74	48	91
19	95	50	77
11	43	54	85
12	60	44	70
11	56	76	72
8	69	79	73
15	61	41	86
11	43	78	72
12	33	65	70
23	95	73	76
20	90	71	72
13	43	72	67
19	54	49	78
14	51	70	78
18	93	78	70
11	50	82	70
14	44	82	81

a. Use the scatterplot matrix in Printout A.8 to assess whether a multiple linear regression model might be appropriate for predicting graduation rate.

b. Use the computer output in Printout A.9 to determine the sample regression equation.

c. Apply the sample regression equation to predict the graduation rate at a school where the student-to-faculty ratio is 18 to 1, 70% of the freshmen were in the top 10% of their high-school class, and 75% of the applicants are accepted. (The values of the predictor variables are 18, 70, and 75.)

A.46 Custom Home Resales. Hanna Properties specializes in custom-home resales in the Equestrian Estates, an exclusive subdivision in Phoenix, Arizona. Thirty-three properties were randomly selected, and following data were obtained. The variables investigated were square footage,

Sale price	SqFt	Bedrooms	Baths	Days
y	x_1	x_2	x_3	x_4
715	5232	5	5	8
583	4316	4	5	140
484	4238	4	4	229
425	3600	4	4	386
418	4000	4	3	0
418	3730	5	4	260
407	3005	4	3	81
405	3800	4	3	52
385	4127	4	4	108
336	3800	4	3	108
330	3200	4	3	52
330	3319	4	3	66
330	3259	4	3	121
319	3200	4	3	103
314	3400	5	3	60
308	3000	4	3	266
308	3007	4	3	144
297	3041	4	3	74
292	3043	4	3	110
286	3406	4	3	10
285	2539	4	2	44
283	3013	4	3	58
282	3022	4	3	171
272	2792	4	3	274
270	3407	4	3	31
267	3275	4	3	361
266	2826	4	3	88
266	2820	4	3	88
266	2826	4	3	252
264	2610	4	3	48
258	2790	3	3	33
253	2400	4	2	57
249	2780	3	3	223

number of bedrooms, number of bathrooms, number of days on the market, and selling price (in thousands of dollars). Minitab was used to perform a regression analysis for selling price in terms of the other four variables. The resulting computer output is displayed in Printouts A.10 and A.11 on page A-26.

a. Use the scatterplot matrix in Printout A.10 to assess whether a multiple linear regression model might be appropriate for predicting selling price.
b. Use the computer output in Printout A.11 to obtain the sample regression equation.
c. Apply the sample regression equation to find the predicted selling price for a home in the Equestrian Estates that has 3200 sq ft, 4 bedrooms, and 3 bathrooms, and has been on the market for 60 days.

Using Technology

A.47 Advertising and Sales. Refer to Exercise A.43. Use the technology of your choice to do the following.

a. Obtain the scatterplot matrix for the data.
b. Use the scatterplot matrix to assess whether a multiple linear regression model might be appropriate for predicting sales.
c. Determine the regression equation for the data.
d. Apply the regression equation to predict total sales if the amounts spent on television, magazine, and radio advertising are $9.5 million, $4.3 million, and $5.2 million, respectively.

A.48 Corvette Prices. Refer to Exercise A.44. Use the technology of your choice to do the following.

a. Obtain the scatterplot matrix for the data.
b. Use the scatterplot matrix to assess whether a multiple linear regression model might be appropriate for predicting the price of a used Corvette.
c. Determine the regression equation for the data.
d. Apply the regression equation to predict the price of a 4-year-old Corvette that has been driven 28,000 miles.

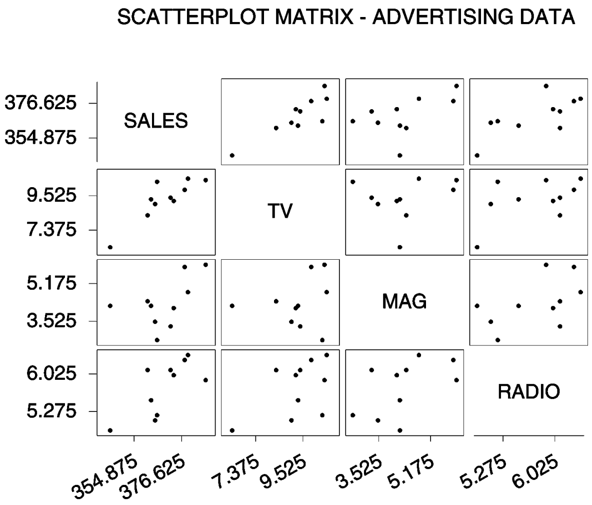
A.49 Graduation Rates. Refer to Exercise A.45. Use the technology of your choice to do the following.

a. Obtain the scatterplot matrix for the data.
b. Use the scatterplot matrix to assess whether a multiple linear regression model might be appropriate for predicting graduation rate.
c. Determine the regression equation for the data.
d. Apply the regression equation to predict the graduation rate at a school where the student-to-faculty ratio is 18 to 1, 70% of the freshmen were in the top 10% of their high-school class, and 75% of the applicants are accepted. (The values of the predictor variables are 18, 70, and 75.)

A.50 Custom Home Resales. Refer to Exercise A.46. Use the technology of your choice to do the following.

a. Obtain the scatterplot matrix for the data.
b. Use the scatterplot matrix to assess whether a multiple linear regression model might be appropriate for predicting the sale price of a home in the Equestrian Estates.
c. Determine the regression equation for the data.
d. Apply the regression equation to find the predicted selling price for a home in the Equestrian Estates that has 3200 sq ft, 4 bedrooms, and 3 bathrooms, and has been on the market for 60 days.

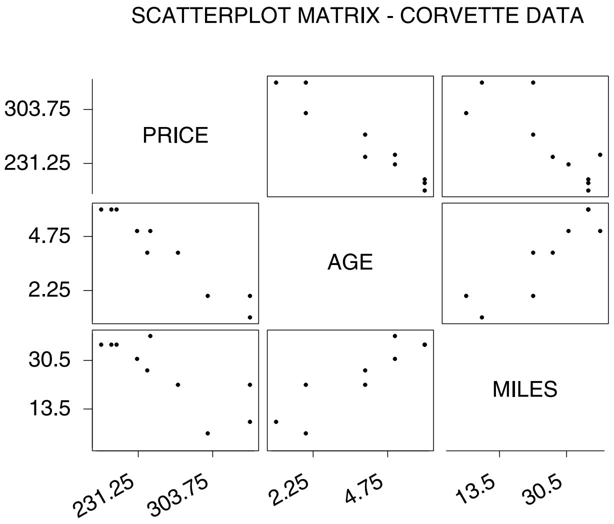
Printout A.4
Minitab output for Exercise A.43



Printout A.5
Minitab output for Exercise A.43 (and Exercises A.64, A.79, and A.91)

```
The regression equation is
SALES = 266 + 6.73 TV + 3.26 MAG + 4.51 RADIO

Predictor        Coef     SE Coef          T        P
Constant       266.23       16.34      16.29    0.000
TV              6.727       1.344       5.01    0.002
MAG             3.257       1.642       1.98    0.095
RADIO           4.507       3.703       1.22    0.269

S = 4.418       R-Sq = 91.1%     R-Sq(adj) = 86.6%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         3     1194.53      398.18     20.40    0.002
Residual Error     6      117.11       19.52
Total              9     1311.64

Predicted Values for New Observations

New Obs     Fit      SE Fit         95.0% CI             95.0% PI
1        367.58       2.59    (  361.24,  373.92)  (  355.05,  380.12)

Values of Predictors for New Observations

New Obs         TV        MAG     RADIO
1             9.50       4.30      5.20
```

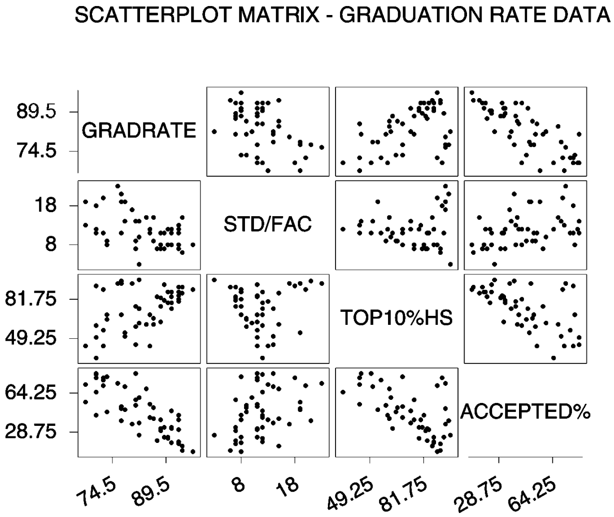
Printout A.6
Minitab output for Exercise A.44



Printout A.7
Minitab output for Exercise A.44 (and Exercises A.65, A.80, and A.92)

```
The regression equation is
PRICE = 367 - 37.4 AGE + 1.64 MILES

Predictor        Coef     SE Coef          T        P
Constant      367.362       9.943      36.95    0.000
AGE           -37.375       5.174      -7.22    0.000
MILES          1.6378      0.8116       2.02    0.083

S = 12.11       R-Sq = 96.0%     R-Sq(adj) = 94.9%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         2       24655       12328     84.07    0.000
Residual Error     7        1026         147
Total              9       25682

Predicted Values for New Observations

New Obs     Fit     SE Fit         95.0% CI             95.0% PI
1        263.72       4.26   (  253.65,  273.80)  (  233.35,  294.10)

Values of Predictors for New Observations

New Obs       AGE     MILES
1            4.00      28.0
```

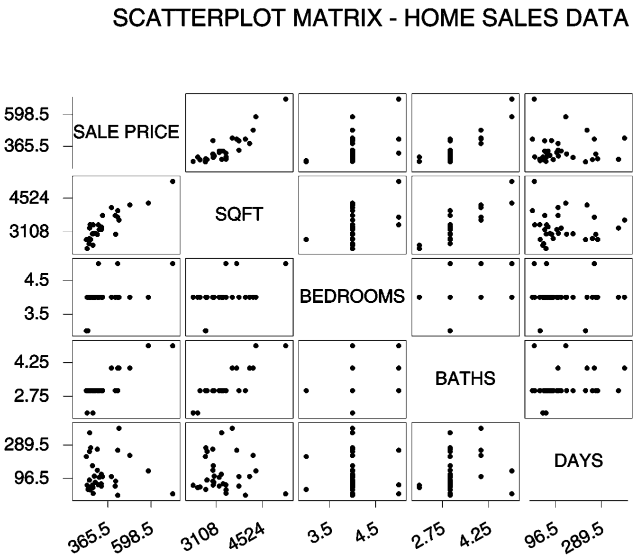
Printout A.8
Minitab output for Exercise A.45



Printout A.9
Minitab output for Exercise A.45 (and Exercises A.66, A.81, and A.93)

```
The regression equation is
GRADRATE = 97.9 - 0.213 STD/FAC + 0.0143 TOP10%HS - 0.293 ACCEPTED%

Predictor         Coef     SE Coef          T        P
Constant        97.873       5.530      17.70    0.000
STD/FAC        -0.2128      0.2033      -1.05    0.301
TOP10%HS       0.01429     0.05727       0.25    0.804
ACCEPTED      -0.29255     0.05199      -5.63    0.000

S = 5.138       R-Sq = 63.8%     R-Sq(adj) = 61.3%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         3     2045.79      681.93     25.83    0.000
Residual Error    44     1161.46       26.40
Total             47     3207.25

Predicted Values for New Observations

New Obs     Fit     SE Fit         95.0% CI            95.0% PI
1        73.102      1.554   (  69.970,  76.235)  (  62.285,  83.920)

Values of Predictors for New Observations

New Obs   STD/FAC  TOP10%HS  ACCEPTED
1            18.0      70.0      75.0
```

Printout A.10
Minitab output for Exercise A.46



Printout A.11
Minitab output for Exercise A.46 (and Exercises A.67, A.82, and A.94)

```
The regression equation is
SALE PRICE = - 233 + 0.0831 SQFT + 22.9 BEDROOMS + 68.2 BATHS - 0.0985 DAYS

Predictor        Coef      SE Coef         T        P
Constant      -232.51        80.24     -2.90    0.007
SQFT          0.08311      0.02575      3.23    0.003
BEDROOMS        22.88        21.96      1.04    0.306
BATHS           68.19        23.75      2.87    0.008
DAYS         -0.09846      0.08374     -1.18    0.250

S = 43.36       R-Sq = 84.1%     R-Sq(adj) = 81.8%

Analysis of Variance

Source           DF          SS          MS         F        P
Regression        4      277643       69411     36.92    0.000
Residual Error   28       52645        1880
Total            32      330288

Predicted Values for New Observations

New Obs     Fit     SE Fit         95.0% CI             95.0% PI
1        323.65       9.09    (  305.02,  342.28)  (  232.90,  414.40)

Values of Predictors for New Observations

New Obs      SQFT  BEDROOMS     BATHS      DAYS
1            3200      4.00      3.00      60.0
```

A.3 INFERENCES CONCERNING THE UTILITY OF THE REGRESSION MODEL

Before discussing how to make statistical inferences about the usefulness of a multiple linear regression, we will consider a popular descriptive measure of the utility of the sample regression equation. This descriptive measure is the **coefficient of multiple determination,** denoted by R^2.

COEFFICIENT OF MULTIPLE DETERMINATION

Recall that in the case of simple linear regression, where there is only one predictor variable for a response variable, y, we use the coefficient of determination, r^2 (Definition 14.5), as a descriptive measure of the utility of the simple linear regression equation. The coefficient of determination is the ratio of the **regression sum of squares (*SSR*)** to the **total sum of squares (*SST*).** Thus

$$r^2 = \frac{SSR}{SST},$$

where

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

and

$$SST = \text{Total sum of squares} = \Sigma(y - \overline{y})^2.$$

The regression sum of squares is the part of the total variation in the observed values of the response variable that is accounted for by the linear regression. The coefficient of determination, r^2, is the proportion of the total variation in the observed values of the response variable that is accounted for by the linear regression. If there is little linear association between y and x, then r^2 will tend to be close to zero. If there is a high degree of linear association between y and x, then r^2 will tend to be close to 1. An exact linear association between y and x always gives $r^2 = 1$.

Also recall that in simple linear regression, we are able to write the total sum of squares (SST) as the sum of the regression sum of squares (SSR) and the **error sum of squares (*SSE*)**. This is the regression identity (Key Fact 14.4):

$$SST = SSR + SSE,$$

where

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

The same definitions of SST, SSR, and SSE can be made in multiple linear regression provided we use $\hat{y} = b_0 + b_1x_1 + \cdots + b_kx_k$. It can be shown that the same regression identity for these three sums of squares holds for the multiple linear regression model with k predictor variables.

KEY FACT A.1 Regression Identity for Multiple Linear Regression

The total sum of squares equals the regression sum of squares plus the error sum of squares. In symbols, $SST = SSR + SSE$. ■

If the sample multiple linear regression equation fits the data well, then the observed and predicted values of the response variable will be "close" together and SSE will be small relative to SST, and SSR will be large relative to SST. As in the case of simple linear regression, we can determine the proportion of the total variation in the observed values of the response variable that is accounted for by the multiple linear regression by considering the ratio SSR/SST.

DEFINITION A.2 Coefficient of Multiple Determination

The ***coefficient of multiple determination, R^2***, is defined by

$$R^2 = \frac{SSR}{SST}.$$

The coefficient of multiple determination is often simply called the ***multiple R^2***. ■

KEY FACT A.2 Interpretation of R^2

The coefficient of multiple determination is the proportion of the total variation in the observed values of the response variable that is explained by the multiple linear regression in the k predictor variables $x_1, x_2, \ldots, x_k$. It always lies between 0 and 1 and is a descriptive measure of the utility of the regression equation for making predictions. Values of R^2 near 0 indicate that the regression equation is not very useful for making predictions, whereas values of R^2 near 1 indicate that the regression equation is very useful for making predictions; R^2 will equal 1 if there is an exact linear relationship between y and $x_1, x_2, \ldots, x_k$. ■

Example A.8 Illustrates the Coefficient of Multiple Determination: Orion Prices

Return to the Orion data set and the multiple linear regression equation that relates price to age and mileage. Determine the coefficient of multiple determination, R^2.

Solution Refer to Printout A.1 on page A-14. In line seven of this printout, the value of R^2 is given by `R-Sq = 93.6%`. Thus $R^2 = 0.936$, and we can say that 93.6% of the total variation in the observed prices is accounted for by the linear regression equation in age and mileage.

We can also compute R^2 by using the formula $R^2 = SSR/SST$. If we look in line nine of Printout A.1, we see that the third column is labeled `SS` (for sum of squares). The first entry in this column is `9088.3` and is the regression sum of squares (SSR) as is indicated by the entry `Regression` in the first column of the table. The third entry in the third column is `9708.5` and is the total sum of squares as is indicated by the entry `Total` in the first column. These two sums of squares allow us to compute R^2: We have $R^2 = SSR/SST = 9088.3/9708.5 = 0.936$. ◆

WARNINGS ON THE USE OF THE COEFFICIENT OF MULTIPLE DETERMINATION

Recall that in the case of simple linear regression, there is a relationship between the coefficient of determination (r^2) and the linear correlation coefficient (r), namely, r^2 is simply the square of the value of r, or r is the square root of r^2 with its sign being the same as the slope of the regression line. No such interpretation of R^2 exists in the case of multiple linear regression. There is no equivalent definition of the square root of R^2.

Also, the same caveat that a high correlation coefficient (or coefficient of determination) in simple linear regression does not imply causation holds true for the multiple R^2. A high value of R^2 does not imply that there is a causal relationship between the predictor variables ($x_1, \ldots, x_k$) and the response variable (y).

Another property of R^2 is that if an additional predictor variable is included in the regression equation, R^2 cannot decrease and generally will increase. This is true no matter what predictor variable is added. The increase will be small if the new predictor variable is not linearly related to the response variable, but there will be an increase. The magnitude of the increase will depend on the added predictor variable and what other predictor variables are already in the regression equation. One should not simply select a model with many predictor variables because it has the highest R^2 value. Methods for deciding which predictor variables to include in a multiple linear regression will be discussed in Module B.

ASSUMPTIONS FOR STATISTICAL INFERENCE

The coefficient of multiple determination, R^2, provides a descriptive measure of the utility of the sample multiple linear regression equation for making predictions. To make inferences about our regression model and its utility, we need to make assumptions like those made in simple linear regression. These are as follows.

KEY FACT A.3 *Assumptions for Multiple Linear Regression Inferences*

1. ***Population regression plane:*** For each set of values, $x_1, x_2, \ldots, x_k$, of the predictor variables, the conditional mean of the response variable y is $\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$. The equation

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

is called the **population regression equation,** and its graph is the **population regression plane.**

2. ***Equal standard deviations:*** The conditional standard deviations of the response variable are the same for all sets of values, $x_1, x_2, \ldots, x_k$, of the predictor variables. We denote this common standard deviation by σ.
3. ***Normal populations:*** For each set of values, $x_1, x_2, \ldots, x_k$, of the predictor variables, the conditional distribution of the response variable is a normal distribution.
4. ***Independent observations:*** The observations of the response variable are independent of one another.

■

Assumptions 1, 2, and 3 require that there are constants $\beta_0, \beta_1, \ldots, \beta_k$ and σ, so that for each set of values, $x_1, x_2, \ldots, x_k$, of the predictor variables, the conditional distribution of the response variable is a normal distribution having mean $\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ and standard deviation σ. Assumption 4 requires that the observations of y be obtained independently.

The multiple linear regression model is often expressed as

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon,$$

where ϵ represents the variations in y about its conditional mean of $\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$. The variable ϵ is commonly referred to as the **error term** in the model. Because of Assumptions 1–4 above, ϵ is a normally distributed variable with mean 0 and standard deviation σ, and the ϵ from one observation is independent of the ϵs from all the other observations. With this representation of the model, we can write

$$\epsilon = y - (\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k).$$

Thus ϵ can be thought of as the difference between the observed value of y and the value of the conditional mean of y at specific values of $x_1, x_2, \ldots, x_k$. The random fluctuations of y from its conditional mean give rise to the variability we see in the observed values of the response variable at specified values of the predictor variables.

Example A.9 *Illustrates Assumptions for Multiple Linear Regression Inferences: Orion Prices*

Consider the variables age, miles driven, and price for Orions, where age is in years, miles driven is in thousands of miles, and price is in hundreds of dollars. Discuss what it would mean for the assumptions for multiple linear regression inferences to be satisfied with age and miles driven as predictor variables for price.

Solution Here we have $k = 2$ since there are two predictor variables, age and miles driven. For the assumptions of multiple linear regression inferences to be satisfied, it would mean that there are constants, $\beta_0, \beta_1, \beta_2$, and σ, such that for each age, x_1, and number of miles driven, x_2, the prices of all Orions of that age that have been driven that number of miles are normally distributed with mean $\beta_0 + \beta_1 x_1 + \beta_2 x_2$ and standard deviation σ.

Consequently, it would mean that the prices of all Orions that are 2 years old ($x_1 = 2$) and have been driven 15,000 miles ($x_2 = 15$) are normally distributed with mean $\beta_0 + \beta_1 \cdot 2 + \beta_2 \cdot 15$ and standard deviation σ; the prices of all Orions that are 3 years old ($x_1 = 3$) and have been driven 32,000 miles ($x_2 = 32$) are normally distributed with mean $\beta_0 + \beta_1 \cdot 3 + \beta_2 \cdot 32$ and standard deviation σ; and so on. Assumption 4 would mean that observations are obtained independently. ◆

THE STANDARD ERROR OF THE ESTIMATE

Assumption 2 for the multiple linear regression model requires that the conditional standard deviations of the response variable be the same regardless of the values of the predictors. The statistic used to estimate σ is called the **standard error of the estimate.**

DEFINITION A.3 *Standard Error of the Estimate*

In a multiple linear regression with k predictors, the ***standard error of the estimate,*** s_e, is defined by

$$s_e = \sqrt{\frac{SSE}{n - (k+1)}},$$

where $SSE = \Sigma(y - \hat{y})^2$. ■

Example A.10 *Illustrates Definition A.3: Orion Prices*

Consider again the Orion data in Table A.1 on page A-11.

a. Use Printout A.1 on page A-14 to obtain the standard error of the estimate.

b. Presuming that the assumptions for multiple linear regression inferences are met with age and miles driven as predictor variables for price, interpret the result in part (a).

Solution

a. The standard error of the estimate is the first entry in the seventh line of Printout A.1: `S = 8.805` (Minitab uses `S` instead of s_e to denote the standard error of the estimate). Thus, for the Orion data in Table A.1, $s_e = 8.805$.

b. Based on the sample data in Table A.1, our best estimate for the (common) conditional standard deviation, σ, of prices for all Orions of any particular age and number of miles driven is $880.5.

◆

In simple linear regression, there is only one predictor variable, so $k = 1$. Then we have $s_e = \sqrt{SSE/(n-2)}$, which is the definition of the standard error of the estimate given in Definition 15.1.

The formula for s_e involves an average value for the squared deviations between the observed and predicted values of the response variable. The average is given by the term $SSE/(n-(k+1))$. The quantity $n-(k+1)$ that divides SSE is the number of degrees of freedom of SSE.

The number of degrees of freedom for SSE equals n, the total sample size, minus $(k+1)$, the total number of regression coefficients $(\beta_0, \beta_1, \ldots, \beta_k)$ estimated by least squares. The average $SSE/(n-(k+1))$ provides an unbiased estimate of σ^2. Thus an estimate of σ is the square root of $SSE/(n-(k+1))$, which is denoted by s_e.

The other sums of squares in the regression identity (SST and SSR) also have degrees of freedom associated with them. SST has $n-1$ degrees of freedom. Recall that $n-1$ is the divisor in calculating the usual standard deviation for the observed values of y. The number of degrees of freedom for SSR is k, the number of predictors in the regression equation. Now there is an identity for degrees of freedom of the sums of squares in multiple linear regression that is analogous to the regression identity for the sums of squares:

KEY FACT A.4 *Regression Identity for Degrees of Freedom*

We have

$$\mathrm{df}(SST) = \mathrm{df}(SSR) + \mathrm{df}(SSE)$$

or

$$n-1 = k + (n-(k+1)),$$

where df(Sum of Squares) represents the degrees of freedom for the sum of squares in question, n is the sample size, and k is the number of predictors. ■

INFERENCES CONCERNING THE UTILITY OF THE MULTIPLE LINEAR REGRESSION EQUATION

Suppose that the variables $x_1, \ldots, x_k$ and y satisfy the Assumptions 1–4 for multiple linear regression inferences. Our first question is whether the regression equation as a whole is useful in making predictions, that is, whether the variables $x_1, \ldots, x_k$ taken together as a group are useful for predicting y.

If $\beta_1, \beta_2, \ldots, \beta_k$ are all 0, then for each set of values $x_1, x_2, \ldots, x_k$ of the predictor variables, the conditional distribution of the response variable is a normal distribution having mean β_0 $(= \beta_0 + 0 \cdot x_1 + 0 \cdot x_2 + \cdots + 0 \cdot x_k)$ and standard deviation σ. So if $\beta_1 = \beta_2 = \cdots = \beta_k = 0$, then the predictor variables provide no information about the conditional distribution of the response variable, and the predictor variables taken together are useless in predicting y. Under that assumption, the conditional mean of y is constant.

Thus to decide on the overall utility of the regression, we test the hypothesis

$$H_0\colon \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

against

$$H_a\colon \text{At least one of the } \beta_i\text{s is not zero.}$$

In determining a test statistic for this null hypothesis, we return to Key Fact A.1, the regression identity for multiple linear regression, $SST = SSR + SSE$. If the k predictor variables are useless in predicting y (H_0 is true), then SSR will tend to be small and SSE will tend to be large relative to SST. If the k predictor variables are actually useful in predicting y (H_0 is false), then SSR will tend to be large and SSE will tend to be small relative to SST.

If we consider the ratio SSR/SSE, then we would expect a small value for this ratio if H_0 is true and a large value if H_0 is false. The test statistic for H_0 uses a slightly modified version of this ratio. The actual test statistic is

$$F = \frac{SSR/k}{SSE/(n-(k+1))}.$$

Thus each sum of squares is divided by its respective degrees of freedom, and then the ratio is formed to determine the test statistic F.

DEFINITION A.4 *Mean Squares and F-Statistic in Regression*

The ratio SSR/k is called the ***mean square for regression*** and denoted ***MSR.*** The ratio $SSE/(n-(k+1))$ is called the ***mean square for error*** and denoted ***MSE.*** The ***F-statistic*** is

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

■

If the null hypothesis ($H_0\colon \beta_1 = \beta_2 = \cdots = \beta_k = 0$) is true, the F-statistic has an F-distribution with degrees of freedom k and $n-(k+1)$. The F-distribution was previously discussed in Section 16.1.

ANALYSIS OF VARIANCE TABLE

The information provided by the regression identities for the sums of squares and the degrees of freedom and the F-statistic are usually summarized in a table called an **analysis of variance table.** An analysis of variance table has five columns: (1) a column for the sources of variation in the data (regression, error, and total), (2) a column for the degrees of freedom (regression, error, and total), (3) a column for the sums of squares (regression, error, and total), (4) a column for the mean squares (regression and error), and (5) a column for the F-statistic.

The general form of an analysis of variance table for a multiple linear regression analysis is shown in Table A.2. Procedure A.1 uses analysis of variance tables in carrying out the F-test for the utility of a multiple linear regression.

Table A.2 Analysis of variance table for regression

Source	df	*SS*	*MS* = *SS*/df	*F*-statistic
Regression	k	SSR	$MSR = \dfrac{SSR}{k}$	$F = \dfrac{MSR}{MSE}$
Error	$n-(k+1)$	SSE	$MSE = \dfrac{SSE}{n-(k+1)}$	
Total	$n-1$	SST		

Procedure A.1 The *F*-Test for the Utility of a Multiple Regression

Assumptions

The four assumptions for multiple linear regression inferences (Key Fact A.3)

Critical-Value Approach

Step 1 The null and alternative hypotheses are:

H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$

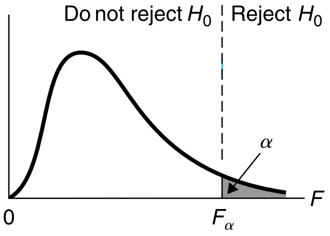
H_a: At least one of the β_is is not zero.

Step 2 Decide on the significance level, α.

Step 3 Obtain the three sum of squares, SST, SSR and SSE.

Step 4 Construct the analysis of variance table, including the value of the F-statistic, $F = MSR/MSE$.

Step 5 The critical value is F_α with df $= (k, n-(k+1))$, where n is the total number of observations, and k is the number of predictor variables. Use Table VIII to find the critical value.



Step 6 If the value of the F-statistic, $F = MSR/MSE$, falls in the rejection region, reject H_0; otherwise, do not reject H_0.

Step 7 Interpret the results of the hypothesis test.

P-Value Approach

Step 1 The null and alternative hypotheses are:

H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$

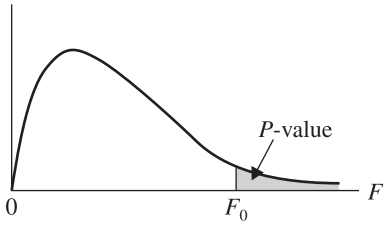
H_a: At least one of the β_is is not zero.

Step 2 Decide on the significance level, α.

Step 3 Obtain the three sum of squares, SST, SSR and SSE.

Step 4 Construct the analysis of variance table, including the value of the F-statistic, $F = MSR/MSE$, and denote that value by F_0.

Step 5 The F-statistic has df $= (k, n-(k+1))$, where n is the total number of observations, and k is the number of predictor variables. Use Table VIII to estimate the P-value, or obtain it exactly using technology.



Step 6 If $P \leq \alpha$, reject H_0; otherwise, do not reject H_0.

Step 7 Interpret the results of the hypothesis test. ■

Example A.11 *Illustrates Procedure A.1: Orion Prices*

Consider again the data on age, miles driven, and price for a sample of 11 Orions, displayed in Table A.1 on page A-11. At the 5% significance level, do the data provide sufficient evidence to conclude that, taken together, age and miles driven are useful for predicting price?

Solution **Step 1** *State the null and alternative hypotheses.*

Let β_1 and β_2 be the population regression coefficients multiplying x_1 (age) and x_2 (miles driven) respectively. Then the null and alternative hypotheses are

H_0: $\beta_1 = \beta_2 = 0$ (Taken together, age and miles driven are not useful for predicting price.)

H_a: At least one of β_1 and β_2 is not zero. (Taken together, age and miles driven are useful for predicting price.)

Step 2 *Decide on the significance level, α.*

We are to perform the test at the 5% significance level; so $\alpha = 0.05$.

Step 3 *Obtain the three sums of squares, SST, SSR, and SSE.*

Printout A.1 on page A-14 shows the three sums of squares under the column labeled `SS` in the analysis of variance table: $SSR = 9088.3$, $SSE = 620.2$, $SST = 9708.5$.

Note that SSE is in the line labeled `Residual Error` in Printout A.1.

Step 4 *Construct the analysis of variance table, including the value of the F-statistic, $F = MSR/MSE$.*

The analysis of variance table is given in Printout A.1 and is as follows.

Source	df	SS	$MS = SS/\text{df}$	F-statistic
Regression	2	9088.3	4544.2	58.61
Error	8	620.2	77.5	
Total	10	9708.5		

The value of the F-statistic is

$$F = MSR/MSE = 4544.2/77.5 = 58.61.$$

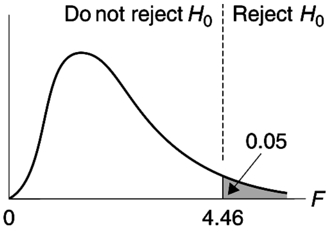
This F-statistic value is also given as the fifth entry on the tenth line of Printout A.1, the line whose first entry is `Regression`. It is also shown in the analysis of variance table above.

Critical-Value Approach

Step 5 *The critical value is F_α with $df = (k, n - (k+1))$, where n is the total number of observations, and k is the number of predictor variables. Use Table VIII to find the critical value.*

We have $\alpha = 0.05$, and we know that the number of predictor variables is $2 (k = 2)$ and the total number of observations is 11 $(n = 11)$. Therefore, $\text{df} = (k, n - (k+1)) = (2, 11 - (2+1)) = (2, 8)$. Consulting Table VIII, we find that the critical value is $F_\alpha = F_{0.05} = 4.46$, as seen in Figure A.5A.

Figure A.5A



Step 6 *If the value of the F-statistic, $F = MSR/MSE$, falls in the rejection region, then reject H_0; otherwise, do not reject H_0.*

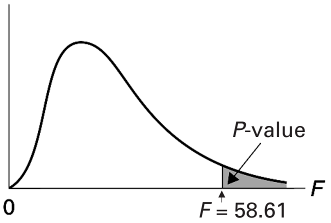
From Step 4, the value of the F-statistic is 58.61. Now $F = 58.61$ falls in the rejection region since $F = 58.61 > 4.46 = F_{0.05}$. Thus we reject H_0.

P-Value Approach

Step 5 *The F-statistic has $df = (k, n - (k+1))$, where n is the total number of observations, and k is the number of predictor variables. Use Table VIII to estimate the P-value, or obtain it exactly using technology.*

From Step 4 we know that the value of the test statistic is $F_0 = 58.61$. Also we know that the number of predictor variables is $2(k = 2)$ and the total number of observations is $11(n = 11)$. Therefore, $\text{df} = (k, n - (k+1)) = (2, 11 - (2+1)) = (2, 8)$. Consulting Table VIII, we find that the area to the right of 58.61 for the F-distribution with $\text{df} = (2, 8)$ is less than 0.001. Thus, $P < 0.001$. The P-value for our test of H_0: $\beta_1 = \beta_2 = 0$ is also given in Printout A.1. In line 10, the sixth entry (the one under the column labeled P) gives the P-value of the hypothesis test as $P = 0.000$ (to three decimal places).

Figure A.5B



Step 6 *If $P \leq \alpha$, reject H_0; otherwise, do not reject H_0.*

Since the P-value is less than 0.001 and thus less than the specified significance level of 0.05, we reject H_0. Furthermore, by referring to Table 9.12, we can see that the data provide very strong evidence against the null hypothesis, and hence in favor of the alternative hypothesis.

Step 7 *Interpret the results of the hypothesis test.*

What Does it Mean ?

The test results are statistically significant at the 5% level; that is at the 5% significance level the data provide sufficient evidence to conclude that at least one of the population regression coefficients (β_1, β_2) is not zero. Thus we conclude that, taken together, age and miles driven are useful in predicting the price of an Orion.

◆

RELATIONSHIP BETWEEN THE COEFFICIENT OF MULTIPLE DETERMINATION AND THE *F*-STATISTIC

The coefficient of multiple determination is given by the formula $R^2 = SSR/SST$. The F-statistic is given by the formula $F = MSR/MSE$. The regression identity for the sum of squares tells us that $SST = SSR + SSE$. Using these facts and a bit of algebra, it can be shown that the F-statistic can be computed from the value of R^2 by using the following formula:

$$F = \frac{R^2/k}{(1 - R^2)/(n - (k + 1))}.$$

Note that if R^2 is close to 0, the value of F will be small. This case corresponds to the situation in which little variation in the observed values of the response variable can be accounted for by the multiple linear regression equation in the predictor variables. If the value of R^2 is close to 1, the denominator term involving $1 - R^2$ will be close to 0 and the value of F will be large. The case that R^2 is close to 1 corresponds to a large proportion of the variation in the observed values of the response variable being accounted for by the multiple linear regression equation.

Exercises A.3

Statistical Concepts and Skills

A.51 Fill in the blanks.

a. A measure of total variation in the observed values of the response variable in a multiple linear regression analysis is the ______. This measure of total variation is denoted by ______.

b. A measure of the amount of variation in the observed values of the response variable that is explained by the multiple linear regression is the ______. This measure is denoted by ______.

c. A measure of the amount of variation in the observed values of the response variable that is not explained by the multiple linear regression is the ______. This measure is denoted by ______.

A.52 In this section we introduced a descriptive measure of the utility of the multiple linear regression equation for making predictions.

a. Define and interpret this descriptive measure.

b. Identify the symbol used for this descriptive measure.

A.53 Suppose x_1, x_2, and x_3 are predictor variables and y is the response variable of a population. Consider the population consisting of all members of the original population having specified values of the three predictor variables. The distribution, mean, and standard deviation of the response variable on that population are called, respectively, the ______, ______, and ______ of the response variable corresponding to the specified values of the predictor variables.

A.54 State the four conditions required for making inferences in multiple linear regression analysis.

In each of Exercises A.55–A.59, assume the predictor variables x_1, x_2, and x_3 and the response variable y satisfy the assumptions for multiple linear regression inferences.

A.55 Fill in the blanks.

a. The plane in four dimensions, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$, is called the ______.

b. The common conditional standard deviation of the response variable is denoted by the symbol ______.

c. For $x_1 = 5$, $x_2 = 8$, and $x_3 = 13$, the conditional distribution of the response variable y is a ______ distribution having mean ______ and standard deviation ______.

A.56 How do we denote the statistics used to estimate

a. the y-intercept of the population regression plane?

b. the partial slopes of the population regression plane?

c. the common conditional standard deviation, σ, of the response variable?

A.57 Regarding the estimate of the conditional standard deviation, σ, which of the three sums of squares figures in its computation?

A.58 Regarding the test of significance for the usefulness of the predictor variables x_1, x_2, and x_3 in predicting y, which of the three sums of squares figures in its calculation?

A.59 State the regression identity for degrees of freedom in this case of three predictor variables.

A.60 Fill in the blanks.

a. When a sum of squares is divided by its degrees of freedom, the ratio is called a ______.

b. The F-statistic for testing the utility of a multiple linear regression equation is defined to be ______.

c. The degrees of freedom for the F-statistic for testing the utility of a multiple linear regression equation involving k predictor variables are ______ and ______.

A.61 Answer true or false to the following statements and explain your answers.

a. If the F-test for the utility of a multiple linear regression rejects the null hypothesis, then all the population regression coefficients are different from 0.

b. If the F-test for the utility of a multiple linear regression does not reject the null hypothesis, we can conclude that there is no relationship whatsoever between the response variable and the predictor variables.

c. If the value of the F-statistic for testing the utility of the multiple linear regression is close to 0, the value of the coefficient of multiple determination will be close to 0.

A.62 For a particular multiple linear regression analysis, there are five predictor variables. A sample of size 37 is obtained on the five predictor variables and the response variable. It is found that $SST = 1256$ and $SSE = 372$.

a. Construct the analysis of variance table for the analysis.

b. Find R^2 and interpret its value.

c. Find the standard error of the estimate, s_e.

d. At the 5% significance level, do the data provide sufficient evidence to conclude that the five predictor variables taken together are useful for predicting the response variable?

e. State how useful you feel the regression equation is for making predictions about the response variable.

A.63 For a particular multiple linear regression analysis there are seven predictor variables. A sample of 68 observations is obtained on the seven predictor variables and the response variable. It is found that $SSR = 987$ and $SSE = 1826$.

a. Construct the analysis of variance table for this analysis.

b. Find R^2 and interpret its value.

c. Find the standard error of the estimate, s_e.

d. At the 1% significance level, do the data provide sufficient evidence to conclude that the seven predictor variables taken together are useful for predicting the response variable?

e. State how useful you feel the regression equation is for making predictions about the response variable.

A.64 Advertising and Sales. Refer to Exercise A.43 on page A-20 regarding the household-appliance manufacturer that wants to analyze the relationship between total sales and the company's expenditures on three primary means of advertising (television, magazines, and radio). Use Printout A.5 on page A-23 to help answer the following questions.

a. Explain what it would mean for the assumptions for multiple linear regression inferences to be satisfied with television, magazine, and radio advertising expenditures as predictor variables for sales.

b. Use the computer output to obtain the coefficient of determination, R^2. Interpret your result.

c. Determine and interpret the standard error of the estimate, s_e.

d. At the 5% significance level, do the data provide sufficient evidence to conclude that, taken together, television, magazine, and radio advertising expenditures are useful for predicting sales?

A.65 Corvette Prices. Refer to Exercise A.44 on page A-20 regarding the relationship between the price of a Corvette and its age and number of miles driven. Use Printout A.7 on page A-24 to help answer the following questions.

a. Explain what it would mean for the assumptions for multiple linear regression inferences to be satisfied with age and miles driven as predictor variables for price.

b. Use the computer output to obtain the coefficient of determination, R^2. Interpret your result.

c. In Exercise 14.65, we found that the coefficient of determination for the age and price data is $r^2 = 0.937$. This means that 93.7% of the total variation in the price data is explained by age. Which regression equation better explains the variation in the price data: the multiple regression equation, using both age and miles as predictor variables; or the simple linear regression equation, using only age as a predictor variable?

d. Determine and interpret the standard error of the estimate, s_e.

e. At the 5% significance level, do the data provide sufficient evidence to conclude that, taken together, age and miles driven are useful for predicting price?

A.66 Graduation Rates. Refer to Exercise A.45 on page A-20 regarding the relationship between college graduation rate and the predictor variables of student-to-faculty ratio, the percentage of freshmen in the top 10% of their high-school class, and the percentage of applicants accepted. Use Printout A.9 on page A-25 to help answer the following questions.

a. Explain what it would mean for the assumptions for multiple linear regression inferences to be satisfied with student-to-faculty ratio, percentage of freshmen in the top 10% of their high school class, and percentage of applicants accepted as predictor variables for graduation rate.
b. Use the computer output to obtain the coefficient of determination, R^2. Interpret your result.
c. How useful do the variables student-to-faculty ratio, percentage of freshmen in the top 10% of their high-school class, and percentage of applicants accepted appear to be for predicting graduation rates at colleges and universities?
d. Determine and interpret the standard error of the estimate, s_e.
e. At the 5% significance level, do the data provide sufficient evidence to conclude that, taken together, student-to-faculty ratio, percentage of freshmen in the top 10% of their high-school class, and percentage of applicants accepted are useful for predicting graduation rate?

A.67 Custom Home Resales. Refer to Exercise A.46 on page A-21 regarding predicting the selling price of a home in the Equestrian Estates using the predictor variables square footage, number of bedrooms, number of bathrooms, and number of days on the market. Use Printout A.11 on page A-26 to help answer the following questions.

a. Explain what it would mean for the assumptions for multiple linear regression inferences to be satisfied with square footage, number of bedrooms, number of bathrooms, and number of days on the market as predictor variables for selling price.
b. Use the computer output to obtain the coefficient of determination, R^2. Interpret your result.
c. Determine and interpret the standard error of the estimate, s_e.
d. Do the data provide sufficient evidence to conclude that, taken together, square footage, number of bedrooms, number of bathrooms, and number of days on the market are useful for predicting selling price? Perform the required hypothesis test at the 1% significance level.

Extending the Concepts and Skills

A.68 Suppose that $R^2 = 1$ for a data set. What can you say about

a. *SSE*?
b. *SSR*?
c. the utility of the sample multiple linear regression equation for making predictions?

A.69 Suppose that $R^2 = 0$ for a data set. What can you say about

a. *SSE*?
b. *SSR*?
c. the utility of the sample multiple linear regression equation for making predictions?

A.70 Use the regression identity for multiple linear regression to show that

$$R^2 = 1 - \frac{SSE}{SST}.$$

a. Explain why this formula shows that the coefficient of multiple determination can also be interpreted as the percentage reduction in the total squared error obtained by using the regression equation instead of the mean, $\bar{y}$, to predict the observed values of the response variable.
b. Referring to Exercise A.65, what percentage reduction in the total squared error is obtained by using the regression equation instead of the mean of the observed prices to predict the observed prices?
c. Referring to Exercise A.66, what percentage reduction in the total squared error is obtained by using the regression equation instead of the mean of the observed graduation rates to predict the observed graduation rates?

Using Technology

A.71 Advertising and Sales. Refer to Exercise A.43 on page A-20. Use the technology of your choice to do the following.

a. Obtain the coefficient of determination. Interpret your result.
b. Determine and interpret the standard error of the estimate, s_e.
c. Test at the 5% level of significance whether the data provide enough evidence to conclude that, taken together, television, magazine, and radio advertising expenditures are useful for predicting sales.

A.72 Corvette Prices. Refer to Exercise A.44 on page A-20. Use the technology of your choice to do the following.

a. Obtain the coefficient of determination. Interpret your result.
b. Determine and interpret the standard error of the estimate, s_e.
c. At the 5% significance level, do the data provide sufficient evidence to conclude that, taken together, age and miles driven are useful for predicting price?

A.73 Graduation Rates. Refer to Exercise A.45 on page A-20 Use the technology of your choice to do the following.

a. Obtain the coefficient of determination. Interpret your result.

b. Determine and interpret the standard error of the estimate, s_e.

c. At the 5% significance level, do the data provide sufficient evidence to conclude that, taken together, student-to-faculty ratio, percentage of freshmen in the top 10% of their high-school class, and percentage of applicants accepted are useful for predicting graduation rate?

A.74 Custom Home Resales. Refer to Exercise A.46 on page A-21. Use the technology of your choice to do the following.

a. Obtain the coefficient of determination. Interpret your result.

b. Determine and interpret the standard error of the estimate, s_e.

c. Do the data provide sufficient evidence to conclude that, taken together, square footage, number of bedrooms, number of bathrooms, and number of days on the market are useful for predicting selling price? Use $\alpha = 0.01$.

A.4 INFERENCES CONCERNING THE UTILITY OF PARTICULAR PREDICTOR VARIABLES

To decide whether a particular predictor variable, say, x_i, is useful for predicting y, we proceed as we did in simple linear regression. Namely, we perform the hypothesis test

$$H_0\colon \beta_i = 0 \quad (x_i \text{ is not useful for predicting } y)$$
$$H_a\colon \beta_i \neq 0 \quad (x_i \text{ is useful for predicting } y).$$

Rejection of the null hypothesis indicates that x_i is useful as a predictor of y. Non-rejection of the null hypothesis suggests that x_i may not be useful as a predictor of y and that it may be worthwhile to do a regression analysis with the variable x_i omitted.

We use the sample regression coefficient b_i to estimate β_i, and b_i will be the basis for our test statistic of the null hypothesis H_0: $\beta_i = 0$. Recall that in the case of simple linear regression, the slope, b_1, of the sample regression equation has a sampling distribution that is normal provided Assumptions 1–4 for regression inference hold (Key Fact 15.1). A similar result is true for the sampling distribution of each b_i when Assumptions 1–4 for multiple linear regression inferences hold (Key Fact A.3 on page A-29).

KEY FACT A.5 The Sampling Distribution of a Sample Regression Coefficient

Suppose that the variables y and $x_1, \ldots, x_k$ satisfy Assumptions 1–4 for multiple linear regression inferences. Then, for samples of size n, each with the same values of the predictor variables, the following properties hold for the sample regression coefficient b_i, for each $i = 1, \ldots, k$.

- The mean of b_i equals the population regression parameter β_i: $\mu_{b_i} = \beta_i$.
- The standard deviation of b_i is $\sigma_{b_i} = \sigma \cdot c_i$. The value of c_i depends on the values of the predictor variables $x_1, \ldots, x_k$ and cannot be expressed easily without matrix notation.
- The variable b_i is normally distributed.

Thus the distribution of all possible sample estimates of β_i is a normal distribution with mean β_i and standard deviation $\sigma \cdot c_i$. ■

As a consequence of Key Fact A.5, the standardized variable

$$z = \frac{b_i - \beta_i}{\sigma \cdot c_i}$$

has the standard normal distribution. However, since σ is unknown, we cannot use z as a basis for a test statistic of H_0: $\beta_i = 0$. But we can use s_e to estimate σ. Then if the estimated standard deviation of b_i is denoted by s_{b_i}, we have

$$s_{b_i} = s_e \cdot c_i.$$

The standard deviation of b_i, s_{b_i}, is also called the standard error of b_i. The value of s_{b_i} is commonly provided by statistical programs such as Minitab. Direct knowledge of c_i is not necessary.

We note that the form of the estimated standard deviation of the sample regression coefficient, b_i, is the same as that of the estimated standard deviation of b_1 in simple linear regression. In each case the estimated standard deviation of a sample regression coefficient is s_e times a constant that depends on the values of the predictor variables.

Now replacing $\sigma \cdot c_i$ by s_{b_i} in the equation for z gives a statistic that has a t-distribution, as might be expected from the similar result in Key Fact 15.4. We have the following.

KEY FACT A.6 *t*-Distribution for Inferences about β_i

Suppose the variables y and $x_1, x_2, \ldots, x_k$ satisfy Assumptions 1–4 for multiple linear regression inferences. Then, for samples of size n, each with the same values of the predictor variables, the variable

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

has the t-distribution with $\text{df} = n - (k + 1)$. Note that the number of degrees of freedom for t is the same as the error degrees of freedom. ■

By using Key Fact A.6, we obtain the test statistic for the null hypothesis H_0: $\beta_i = 0$ to be the sample regression coefficient divided by its estimated standard deviation:

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

The critical values and P-values for this test statistic are available from the t-table, Table IV.

Procedure A.2 The *t*-Test for the Utility of x_i in the Multiple Linear Regression Equation

Assumptions

The four assumptions for multiple linear regression inferences (Key Fact A.3)

Critical-Value Approach

Step 1 The null and alternative hypotheses are:

H_0: $\beta_i = 0$ (predictor variable x_i is not useful in making predictions)

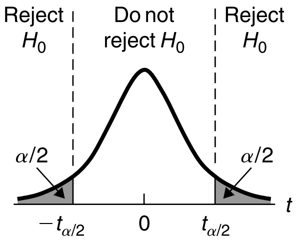
H_a: $\beta_i \neq 0$ (predictor variable x_i is useful for making predictions).

Step 2 Decide on the significance level, α.

Step 3 Compute the value of the test statistic

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

Step 4 The critical values are $\pm t_{\alpha/2}$ with df $= n - (k + 1)$, where n is the total number of observations, and k is the number of predictor variables. Use Table IV to find the critical values.



Step 5 If the value of the test statistic falls in the rejection region, reject H_0; otherwise, do not reject H_0.

Step 6 Interpret the results of the hypothesis test.

P-Value Approach

Step 1 The null and alternative hypotheses are:

H_0: $\beta_i = 0$ (predictor variable x_i is not useful in making predictions)

H_a: $\beta_i \neq 0$ (predictor variable x_i is useful for making predictions).

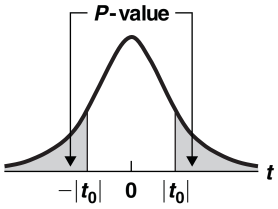
Step 2 Decide on the significance level, α.

Step 3 Compute the value of the test statistic

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

and denote that value by t_0.

Step 4 The *t*-statistic has df $= n - (k + 1)$, where n is the total number of observations, and k is the number of predictor variables. Use Table IV to estimate the *P*-value, or obtain it exactly using technology.



Step 5 If $P \leq \alpha$, reject H_0; otherwise, do not reject H_0.

Step 6 Interpret the results of the hypothesis test. ■

We must be careful when interpreting the conclusion of a hypothesis test for a β-parameter. Although x_i may not be useful for predicting y in conjunction with the other predictor variables under consideration, it may be useful when employed as the only predictor variable or with some other collection of predictor variables. Thus, in this section, when we say that x_i is not useful for predicting y, we really mean that in the regression with $x_1, x_2, \ldots, x_k$ as the predictor variables, x_i is not useful for predicting y. This issue will be discussed again in Module B when we consider the problem of multicollinearity (correlations among the predictor variables).

Likewise, although x_i may be useful for predicting y in conjunction with the other predictor variables under consideration, it may not be useful when employed as the only predictor variable or with some other collection of predictor variables. Thus, in this section, when we say that x_i is useful for predicting y, we really mean that in the regression with $x_1, x_2, \ldots, x_k$ as the predictor variables, x_i is useful for predicting y.

Example A.12 *Illustrates Procedure A.2: Orion Prices*

Consider again the data on age, miles driven, and price for a sample of 11 Orions, presented in Table A.1 on page A-11. At the 5% significance level, do the data provide sufficient evidence to conclude that, in conjunction with age, number of miles driven is useful for predicting price?

Solution We will assume that the assumptions for multiple linear regression inference are satisfied for the variables of age, mileage, and price. We will apply Procedure A.2 to perform the required hypothesis test.

Step 1 *State the null and alternative hypotheses.*

Let β_2 denote the regression coefficient for mileage in our multiple linear regression equation. Then the null and alternative hypotheses are

$$H_0\colon \beta_2 = 0 \text{ (miles driven is not useful for predicting price)}$$

$$H_a\colon \beta_2 \neq 0 \text{ (miles driven is useful for predicting price).}$$

Step 2 *Decide on the significance level, α.*

We are to perform the hypothesis test at the 5% significance level; so $\alpha = 0.05$.

Step 3 *Compute the value of the test statistic.*

$$t = \frac{b_2}{s_{b_2}}.$$

Look at the sixth line of the computer output in Printout A.1 on page A-14, the line whose first entry is `MILES`. The second entry in that line in the column labeled `Coef` gives the coefficient b_2 of the sample regression plane; so $b_2 = -0.8215$. The third entry in that line in the column labeled `SE Coef` shows the estimated standard deviation of b_2, that is s_{b_2}; so $s_{b_2} = 0.2552$. The fourth entry in that line in the column labeled `T` displays the value of the test statistic

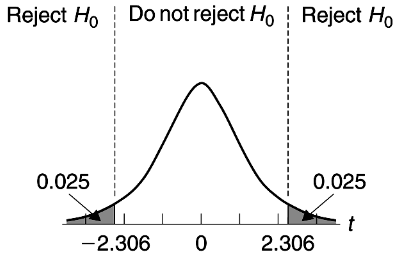
$$t = \frac{b_2}{s_{b_2}} = \frac{-0.8215}{0.2552} = -3.22.$$

Critical-Value Approach

Step 4 *The critical values are* $\pm t_{\alpha/2}$ *with* $df = n - (k + 1)$. *Use Table IV to find the critical values.*

From Step 2, $\alpha = 0.05$. We also have $n = 11$ and $k = 2$, so $df = n - (k + 1) = 11 - (2 + 1) = 8$. Using Table IV, we find that the critical values are $\pm t_{\alpha/2} = \pm t_{0.05/2} = \pm t_{0.025} = \pm 2.306$. These values and the rejection region are shown in Figure A.6A.

Figure A.6A



Step 5 *If the value of the test statistic falls in the rejection region, reject* H_0*; otherwise, do not reject* H_0.

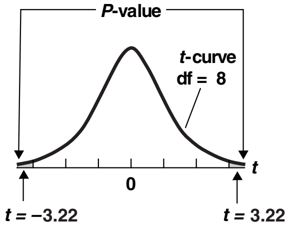
From Step 3, the value of the test statistic is $t = -3.22$. Since this falls in the rejection region, we reject H_0.

P-Value Approach

Step 4 *The* t*-statistic has* $df = n - (k + 1)$, *where* n *is the total number of observations, and* k *is the number of predictor variables. Use Table IV to estimate the P-value, or obtain it exactly using technology.*

From Step 3 we see that the value of the test statistic is $t_0 = -3.22$. We also have $n = 11$ and $k = 2$, so $\text{df} = n - (k + 1) = 11 - (2 + 1) = 8$. We find from Table IV that the area to the right of $|-3.22|$ is between 0.005 and 0.01. Multiplying by two, gives $0.01 < P < 0.02$. The P-value for this test statistic can also be found in the sixth line of the computer output in Printout A.1 on page A-14, the line whose first entry is `MILES`. The final entry in this line gives the P-value for the hypothesis test to be $P = 0.012$.

Figure A.6B



Step 5 *If* $P \leq \alpha$, *reject* H_0*; otherwise, do not reject* H_0.

Since $0.01 < P < 0.02$, the P-value is less than the specified significance level of 0.05, and we reject H_0. Furthermore, by referring to Table 9.12, we see that the data provide strong evidence against the null hypothesis.

Step 6 *Interpret the results of the hypothesis test.*

What Does it Mean?

The test results are statistically significant at the 5% level; that is, at the 5% significance level, the data provide sufficient evidence to conclude that the regression coefficient β_2 of the population regression plane is not 0. Hence, in conjunction with age, mileage is a useful predictor of the price of an Orion.

◆

CONFIDENCE INTERVAL FOR β_i

Recall that in simple linear regression, it is useful to obtain a confidence interval for the slope, β_1, of the population regression line. Likewise, in multiple linear regression it is useful to construct a confidence interval for each of the regression coefficients, $\beta_1, \beta_2, \ldots, \beta_k$, of the population regression plane. We know that a point estimate of β_i is given by b_i. To determine a confidence interval estimate for β_i, we apply Key Fact A.6 on page A-41 to obtain the following procedure.

Procedure A.3 The *t*-Interval Procedure for the Regression Coefficients of a Population Regression Plane

Assumptions

The four assumptions for multiple linear regression inferences

Step 1 For a confidence level of $1 - \alpha$, use Table IV to find $t_{\alpha/2}$ with $\text{df} = n - (k + 1)$.

Step 2 The endpoints of the confidence interval for β_i are

$$b_i \pm t_{\alpha/2} \cdot s_{b_i}.$$

Step 3 Interpret the confidence interval. ■

Example A.13 *Illustrates Procedure A.3: Orion Prices*

Use the data in Table A.1 on page A-11 to obtain individual 95% confidence intervals for the regression coefficients β_1 and β_2 of the population regression plane that relates price to age and mileage for Orions.

Solution We apply Procedure A.3.

Step 1 *For a confidence level of* $1 - \alpha$, *use Table IV to find* $t_{\alpha/2}$ *with* $df = n - (k + 1)$.

For a 95% confidence interval, $\alpha = 0.05$. Since $n = 11$ and $k = 2$, $\text{df} = 11 - (2 + 1) = 8$. Using Table IV, we obtain $t_{\alpha/2} = t_{0.05/2} = t_{0.025} = 2.306$.

Step 2 *The endpoints for the confidence interval for* β_1 *are*

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

By using the information on the fifth line of Printout A.1 on page A-14 we have $b_1 = -9.504$, and $s_{b_1} = 3.874$. So the endpoints of the 95% confidence interval for β_1 are

$$-9.504 \pm 2.306 \cdot 3.874$$

or -18.44 to -0.57. In a similar fashion, the 95% confidence interval for β_2 can be found to be $-0.8215 \pm 2.306 \cdot 0.2552$, or -1.41 to -0.23.

Step 3 *Interpret the confidence interval.*

What Does it Mean?

We are 95% confident that the regression coefficient, β_1, is somewhere between -18.44 and -0.57. In other words, we are 95% confident that, for a fixed mileage, the yearly decrease in the mean price for Orions is somewhere between \$57 and \$1844. Also, we are 95% confident that the regression coefficient, β_2, is somewhere between -1.41 to -0.23. In other words, we are 95% confident that, for a fixed age, the mean price for Orions decreases somewhere between \$23 and \$141 for each increase in mileage of 1000 miles.

Note that the confidence interval for the regression coefficient of age in our two-predictor regression equation is considerably different than that found for the slope of the simple linear regression line relating price to the single predictor, age, in Example 15.6. The 95% confidence interval for the slope for regressing price on age in Example 15.7 is -26.59 to -13.93.

This difference occurs because of the additional predictor, mileage, in our two-predictor regression equation. Remember that in multiple linear regression, the inference we make about a regression coefficient depends on what other predictor variables are in the regression equation.

Exercises A.4

Statistical Concepts and Skills

A.75 Explain why the predictor variables are useless as predictors of the response variable if the partial slopes of the population regression plane are all zero.

A.76 For variables x_1, x_2, x_3, and y satisfying the assumptions for multiple linear regression inferences, the population regression equation is $y = 27 - 4.7x_1 + 2.3x_2 + 5.8x_3$. For samples of size 20 and given values of the predictor variables, the distribution of the estimates of β_1 for all possible sample regression planes is a ______ distribution with mean ______ and standard deviation ______.

A.77 What test statistic is used for a hypothesis test that the population regression coefficient for the predictor variable x_1 is zero? What is the distribution of this test statistic?

A.78 Answer true or false to the following statements and explain your answers.

a. In a multiple linear regression analysis, the test of H_0: $\beta_1 = 0$ can be affected by what other predictor variables are in the multiple linear regression equation.

b. If the F-test for the utility of the multiple linear regression equation rejects the null hypothesis of H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$, then each individual t-test of H_0: $\beta_i = 0, i = 1, \ldots, k$, will result in rejection of the null hypothesis.

c. The number of degrees of freedom for the $t_{\alpha/2}$-value used in constructing a confidence interval for a population regression coefficient is the same as the number of degrees of freedom for the mean square error (MSE).

A.79 Advertising and Sales. Refer to Exercise A.43 on page A-20 regarding the household-appliance manufacturer that wants to analyze the relationship between total sales and the company's expenditures on three primary means of advertising (television, magazines, and radio). Use Printout A.5 on page A-23 to help answer the following questions.

a. At the 5% significance level, do the data provide sufficient evidence to conclude that television advertising expenditure is useful for predicting sales? Be precise in your conclusion.

b. Repeat part (a) for radio advertising expenditure.

c. Find a 95% confidence interval for the coefficient β_1 of the predictor variable x_1 (television advertising expenditure).

d. Interpret your result from part (c) in words.

e. Repeat part (c) for the coefficient β_3 of the predictor variable x_3 (radio advertising expenditure).

A.80 Corvette Prices. Refer to Exercise A.44 on page A-20 regarding the relationship between the price of a Corvette and its age and number of miles driven. Use Printout A.7 on page A-24 to help answer the following questions.

a. At the 5% significance level, do the data provide sufficient evidence to conclude that age is useful for predicting price? Be precise in your conclusion.

b. Repeat part (a) for miles driven. What if the hypothesis test is performed at the 10% significance level?

c. Find a 95% confidence interval for the coefficient β_1 of the predictor variable x_1 (age).

d. Interpret your result from part (c) in words.

e. Repeat part (c) for the coefficient β_2 of the predictor variable x_2 (miles driven).

A.81 Graduation Rates. Refer to Exercise A.45 on page A-20 regarding the relationship between college graduation rate and the predictor variables student-to-faculty ratio, percentage of freshmen in the top 10% of their high-school class, and percentage of applicants accepted. Use Printout A.9 on page A-25 to help answer the following questions.

a. At the 5% significance level, do the data provide sufficient evidence to conclude that percentage of freshmen in the top 10% of their high-school class is useful for predicting graduation rate? Be precise in your conclusion.

b. Repeat part (a) for each of the variables student-to-faculty ratio and percentage accepted.

c. Do you think another regression analysis is called for? If so, which predictor variable(s) might you include? Explain your answers.

d. Find a 95% confidence interval for the coefficient β_2 of the predictor variable x_2 (percentage of freshmen in the top 10% of their high-school class).

e. Interpret your result from part (d) in words.

f. Repeat part (d) for the coefficient β_1 of the predictor variable x_1 (student-to-faculty ratio).

A.82 Custom Home Resales. Refer to Exercise A.46 on page A-20 regarding predicting the selling price of a home in the Equestrian Estates using the predictor variables square footage, number of bedrooms, number of bathrooms, and number of days on the market. Use Printout A.11 on page A-26 to help answer the following questions.

a. At the 1% significance level, do the data provide sufficient evidence to conclude that square footage is useful for predicting selling price? Be precise in your conclusion.

b. Repeat part (a) for the variable number of bedrooms. Explain why number of bedrooms might not be useful for predicting selling price of homes in the Equestrian Estates.

c. Suppose you decide to run another regression analysis on the data, this time with fewer predictor variables. Which predictor variables would you include? Explain your answer.

d. Find a 95% confidence interval for the coefficient β_1 of the predictor variable x_1 (square footage).

e. Interpret your result from part (d) in words.

f. Repeat part (d) for the coefficient β_2 of the predictor variable x_2 (number of bedrooms).

Using Technology

A.83 Advertising and Sales. Referring to Exercise A.79, use the technology of your choice to perform the tests of hypotheses in parts (a) and (b).

A.84 Corvette Prices. Referring to Exercise A.80, use the technology of your choice to perform the tests of hypotheses in parts (a) and (b).

A.85 Graduation Rates. Referring to Exercise A.81, use the technology of your choice to perform the tests of hypotheses in parts (a) and (b).

A.86 Custom Home Resales. Referring to Exercise A.82, use the technology of your choice to perform the tests of hypotheses in parts (a) and (b).

A.5 CONFIDENCE INTERVALS FOR MEAN RESPONSE; PREDICTION INTERVALS FOR RESPONSE

In this section we will learn how a sample multiple linear regression equation can be used to make two important inferences:

- Estimating the conditional mean of the response variable corresponding to a particular set of values of the k predictor variables.
- Predicting the value of the response variable for a particular set of values of the k predictor variables.

We will use the Orion example to illustrate the relevant ideas. In doing so, we will presume that the assumptions for multiple linear regression inferences (Key Fact A.3 on page A-29) are satisfied by the variables age, miles driven, and price for Orions.

To determine a point estimate for the conditional mean of the response variable, y, corresponding to particular values, $x_{1p}, x_{2p}, \ldots, x_{kp}$, of the k predictor variables, we proceed as in simple linear regression. A point estimate for the conditional mean of y is obtained by substituting the particular predictor variable values into the sample multiple linear regression equation:

$$\hat{y}_p = b_0 + b_1 x_{1p} + b_2 x_{2p} + \cdots + b_k x_{kp}.$$

Example A.14 *Estimating Conditional Means in Multiple Linear Regression: Orion Prices*

In Table A.1 on page A-11, we presented data on age, miles driven, and price for a sample of 11 Orions. Determine a point estimate for the mean price of all Orions that are 5 years old and have been driven 52,000 miles.

Solution By Assumption 1 of the assumptions for multiple linear regression inferences, the population regression equation gives the mean prices for Orions of various ages and miles driven. In particular, the mean price of all Orions that are 5 years old and have been driven 52,000 miles is $\beta_0 + \beta_1 \cdot 5 + \beta_2 \cdot 52$. Since β_0, β_1, and β_2 are unknown, we estimate this mean price by the corresponding value, $b_0 + b_1 \cdot 5 + b_2 \cdot 52$, on the sample regression plane.

We employed Minitab to perform a multiple regression analysis on the data in Table A.1 with age and miles driven as predictor variables for price. In doing so, we instructed Minitab to provide more detailed output than in Printout A.1 (page A-14) to obtain information on the estimate of the conditional mean price and on certain confidence and prediction intervals. The procedure for obtaining the resulting output, shown in Printout A.12, will be given in The Technology Center at the end of this section.

Noting that the sample regression equation obtained for the regression of price on age and miles driven, given in Printout A.12, is $\hat{y} = 183 - 9.50x_1 - 0.821x_2$, our estimate for the mean price of all Orions that are 5 years old and have been driven 52,000 miles is

$$\hat{y} = 183 - 9.50 \cdot 5 - 0.821 \cdot 52 = 92.80$$

or $9280.[1] This estimate can also be obtained directly from Printout A.12; it is the second entry in the fourth line from the bottom under the column labeled `Fit`.

[1] In calculating this estimate, we did not use the rounded values, 183, −9.50, and −0.821, of the sample regression coefficients; rather, we kept full computer accuracy. Using the rounded values yields $9281.

Printout A.12
Minitab regression output for Orion data with confidence and prediction intervals

```
The regression equation is
PRICE = 183 - 9.50 AGE - 0.821 MILES

Predictor        Coef     SE Coef          T        P
Constant       183.04       11.35      16.13    0.000
AGE            -9.504       3.874      -2.45    0.040
MILES         -0.8215      0.2552      -3.22    0.012

S = 8.805       R-Sq = 93.6%     R-Sq(adj) = 92.0%

Analysis of Variance

Source            DF          SS          MS          F        P
Regression         2      9088.3      4544.2      58.61    0.000
Residual Error     8       620.2        77.5
Total             10      9708.5

Predicted Values for New Observations

New Obs     Fit     SE Fit        95.0% CI              95.0% PI
1         92.80       2.74   (   86.48,    99.12)  (   71.53,  114.06)

Values of Predictors for New Observations

New Obs       AGE     MILES
1            5.00      52.0
```

Note that the estimate for the mean price of all Orions that are 5 years old and have been driven 52,000 miles is the same as the predicted price for an Orion of this age and mileage. Both are obtained by substituting $x_1 = 5$ and $x_2 = 52$ into the sample multiple linear regression equation. ◆

The estimate of $9280 for the mean price of all Orions that are 5 years old and have been driven 52,000 miles is a point estimate. As in the case of simple linear regression, we would like some idea of how accurate this point estimate is. Thus it would be useful to provide a confidence-interval estimate for the mean price of all Orions that are 5 years old and have been driven 52,000 miles. We will now learn how to obtain such confidence-interval estimates.

CONFIDENCE INTERVALS FOR CONDITIONAL MEANS IN MULTIPLE LINEAR REGRESSION

We proceed to develop a confidence interval for conditional means in multiple linear regression in the same way we developed such a confidence interval in simple linear regression. First we must identify the distribution of the predicted value of the response variable for a particular set of values of the k predictor variables.

For motivation, let us return to the Orion example and consider the particular values of 5 years and 52 thousand miles for the two predictor variables age and mileage. Recall that there are 11 Orions in our data set and that the predicted price of an Orion that is 5 years old and has been driven 52,000 miles is \$9280.

Consider now all possible samples of 11 Orions whose combinations of age and miles driven are the same as those given in the age and mileage columns of Table A.1 on page A-11. For all these samples, the predicted price of an Orion that is 5 years old and has been driven 52,000 miles varies from one sample to another and is therefore a variable. Using the assumptions for multiple linear regression inferences, it can be shown that its distribution is a normal distribution with mean equal to the mean price of all Orions that are 5 years old and have been driven 52,000 miles. More generally, we have the following fact.

KEY FACT A.7 *Distribution of the Predicted Value of a Response Variable in Multiple Linear Regression*

Suppose the variables $x_1, x_2, \dots, x_k$ and y satisfy Assumptions 1–4 for multiple linear regression inferences. Let $x_{1p}, x_{2p}, \dots, x_{kp}$ denote particular values of the k predictor variables and $\hat{y}_p$ the corresponding value predicted for the response variable by the sample multiple linear regression equation, that is, $\hat{y}_p = b_0 + b_1x_{1p} + b_2x_{2p} + \cdots + b_kx_{kp}$. Then, for samples of size n, each with the same values of the predictor variables, the following properties hold for $\hat{y}_p$.

- The variable $\hat{y}_p$ is normally distributed.
- The mean of $\hat{y}_p$ equals the conditional mean of the response variable corresponding to the values $x_{1p}, x_{2p}, \dots, x_{kp}$ of the k predictor variables:

$$\mu_{\hat{y}_p} = \beta_0 + \beta_1x_{1p} + \cdots + \beta_kx_{kp}.$$

- The standard deviation of $\hat{y}_p$ is

$$\sigma_{\hat{y}_p} = \sigma \cdot d_p,$$

where d_p depends on the predictor variable values and on the values of $x_{1p}, x_{2p}, \dots, x_{kp}$.

■

In particular, the distribution of all possible predicted values of the response variable corresponding to $x_{1p}, x_{2p}, \dots, x_{kp}$ is a normal distribution with mean $\beta_0 + \beta_1x_{1p} + \cdots + \beta_kx_{kp}$. The standard deviation of this distribution equals $\sigma \cdot d_p$.

The formula for d_p is rather complicated and can only be expressed easily using matrix notation. As in simple linear regression (see the equation for $\sigma_{\hat{y}_p}$ in Key Fact 15.5), it involves the predictor variable values as well as the values of the predictors at which we want to predict the response variable. We will not need the formula for d_p.

In view of Key Fact A.7, if we standardize the variable $\hat{y}_p$, the resulting variable has the standard normal distribution. However, because the standardized variable contains the unknown parameter σ, it cannot be used as a basis for a confidence interval for the conditional mean of y. However, we can replace σ by its estimate s_e, the standard error of the estimate. The resulting variable has a t-distribution.

KEY FACT A.8 *t-Distribution for Confidence Intervals for Conditional Means in Multiple Linear Regression*

Suppose the variables $x_1, x_2, \ldots, x_k$ and y satisfy Assumptions 1–4 for multiple linear regression inferences. Then, for samples of size n, each with the same values of the predictor variables, the variable

$$t = \frac{\hat{y}_p - \left(\beta_0 + \beta_1 x_{1p} + \cdots + \beta_k x_{kp}\right)}{s_{\hat{y}_p}}$$

has the t-distribution with $\text{df} = n - (k+1)$, where $s_{\hat{y}_p} = s_e \cdot d_p$. ■

Recalling that $\beta_0 + \beta_1 x_{1p} + \cdots + \beta_k x_{kp}$ is the conditional mean of the response variable corresponding to the predictor variable values $x_{1p}, x_{2p}, \ldots, x_{kp}$, we can use Key Fact A.8 to derive the following confidence-interval procedure for means in multiple linear regression.

Procedure A.4 The *t*-Interval Procedure for a Conditional Mean of the Response Variable

Assumptions

The four assumptions for multiple linear regression inferences

Step 1 For a confidence level of $1 - \alpha$, use Table IV to find $t_{\alpha/2}$ with $\text{df} = n - (k+1)$.

Step 2 Compute the point estimate, $\hat{y}_p = b_0 + b_1 x_{1p} + b_2 x_{2p} + \cdots + b_k x_{kp}$, for the conditional mean of the response variable corresponding to the particular values $x_{1p}, x_{2p}, \ldots, x_{kp}$ of the k predictor variables.

Step 3 The endpoints of the confidence interval for the conditional mean of the response variable are

$$\hat{y}_p \pm t_{\alpha/2} \cdot s_{\hat{y}_p},$$

where $s_{\hat{y}_p} = s_e \cdot d_p$.

Step 4 Interpret the confidence interval. ■

Recall that we have not given a formula for d_p. Most statistical programs, such as Minitab, will provide either the value of the estimated standard deviation of $\hat{y}_p$, that is, $s_{\hat{y}_p}$, or the endpoints of the confidence interval for the conditional mean for a particular set of values of the predictor variables. Direct knowledge of the value of d_p is not necessary. Note that $s_{\hat{y}_p}$ is also called the standard error of $\hat{y}_p$.

Example A.15 *Illustrates Procedure A.4: Orion Prices*

Refer to the data on age, miles driven, and price for a sample of 11 Orions shown in Table A.1 on page A-11. Obtain a 95% confidence interval for the mean price of all Orions that are 5 years old and have been driven 52,000 miles.

Solution We apply Procedure A.4.

Step 1 *For a confidence level of* $1-\alpha$*, use Table IV to find* $t_{\alpha/2}$ *with* $df = n-(k+1)$.

We want a 95% confidence interval. Thus we have $\alpha = 0.05$. Since $n = 11$ and $k = 2$, we have $\text{df} = n-(k+1) = 11-(2+1) = 8$. Consulting Table IV, we find that $t_{\alpha/2} = t_{0.05/2} = t_{0.025} = 2.306$.

Step 2 *Compute the point estimate,* $\hat{y}_p = b_0 + b_1x_{1p} + b_2x_{2p} + \cdots + b_kx_{kp}$*, for the conditional mean of the response variable corresponding to the particular values* $x_{1p}, x_{2p}, \ldots, x_{kp}$ *of the k predictor variables.*

From Example A.5, the sample regression equation for the data in Table A.1 is $\hat{y} = 183 - 9.50x_1 - 0.821x_2$. Here we want $x_1 = x_{1p} = 5$ (years) and $x_2 = x_{2p} = 52$ (thousand miles). So the point estimate for the mean price of all Orions that are 5 years old and have been driven 52,000 miles is

$$\hat{y}_p = 183 - 9.50 \cdot 5 - 0.821 \cdot 52 = 92.80.$$

This point estimate of the conditional mean is also given as the second item in the fourth line from the bottom of Printout A.12 (page A-49) under the heading `Fit`.

Step 3 *The endpoints of the confidence interval for the conditional mean of the response variable are*

$$\hat{y}_p \pm t_{\alpha/2} \cdot s_{\hat{y}_p}.$$

The required value of $s_{\hat{y}_p}$ is given by the third item in the fourth line from the bottom of Printout A.12, the item headed `SE Fit`. Thus $s_{\hat{y}_p} = 2.74$. Consequently, the endpoints of the confidence interval for the conditional mean are

$$92.80 \pm 2.306 \cdot 2.74,$$

or 86.48 to 99.12. This confidence interval is also given as the fourth item in the fourth line from the bottom of Printout A.12, the item headed `95% CI`.

Step 4 *Interpret the confidence interval.*

What Does it Mean ? We can be 95% confident that the mean price of all Orions that are 5 years old and have been driven 52,000 miles is somewhere between \$8648 and \$9912.

◆

PREDICTION INTERVALS

In simple linear regression, a primary use of a sample regression equation is for making predictions. The same is true for multiple linear regression. For example, the sample regression equation for the Orion data is $\hat{y} = 183 - 9.50x_1 - 0.821x_2$. So the predicted price of an Orion that is 5 years old and has been driven 52,000 miles is $\hat{y} = 183 - 9.50 \cdot 5 - 0.821 \cdot 52 = 92.80$, or \$9280. Since the prices of such cars vary, it makes more sense to find a **prediction interval** for the price of an Orion that is 5 years old and has been driven 52,000 miles than to give a single predicted value.

Recall from Section 15.3 that the term *prediction* is used for interval estimates of a response variable, such as the price of an Orion of a particular age and mileage. The term *confidence* is used for interval estimates of parameters, such as the mean price of all Orions of a particular age and mileage.

To develop a prediction-interval procedure, we first identify the distribution of the difference between the observed and predicted values of the response variable for a particular set of values of the predictor variables. To illustrate the ideas, we return to the Orion example and consider the particular value of age to be 5 years and of mileage to be 52,000 miles.

As we have seen, the predicted price, in hundreds of dollars, for an Orion that is 5 years old and has been driven 52,000 miles is 92.80. Suppose we observe the price of a 5-year-old Orion with mileage of 52,000 miles and find it to be 94.95. Then the difference between the observed price and the predicted price is 94.95 − 92.80, or 2.15.

Consider now all possible samples of 11 Orions whose combinations of age and miles driven are the same as those given in Table A.1 on page A-11. For such samples, the predicted price of an Orion that is 5 years old and has been driven 52,000 miles varies from one sample to another and is therefore a variable. The observed price of an Orion that is 5 years old and has been driven 52,000 miles is a variable as well. Thus the difference between the observed price and predicted price is also a variable. Using the assumptions for multiple linear regression inferences, it can be shown that the distribution of this difference is a normal distribution with mean 0. More generally, we have the following fact.

KEY FACT A.9 *Distribution of the Difference between the Observed and Predicted Values of the Response Variable in Multiple Linear Regression*

Suppose the variables $x_1, x_2, \ldots, x_k$ and y satisfy Assumptions 1–4 for multiple linear regression inferences. Let $x_{1p}, x_{2p}, \ldots, x_{kp}$ denote particular values of the k predictor variables and let $\hat{y}_p$ be the corresponding value predicted for the response variable by the sample multiple linear regression equation. Furthermore, let y_p be an independently observed value of the response variable corresponding to the values $x_{1p}, x_{2p}, \ldots, x_{kp}$ of the k predictor variables. Then for samples of size n, each with the same values of the predictor variables, the following properties hold for $y_p - \hat{y}_p$, the difference between the observed and predicted values.

- The variable $y_p - \hat{y}_p$ is normally distributed.
- The variable $y_p - \hat{y}_p$ has mean equal to 0: $\mu_{y_p-\hat{y}_p} = 0$.
- The standard deviation of $y_p - \hat{y}_p$ is

$$\sigma_{y_p-\hat{y}_p} = \sigma \cdot \sqrt{1 + d_p^2}.$$

In particular, the distribution of all possible differences between the observed and predicted values of the response variable corresponding to $x_{1p}, x_{2p}, \ldots, x_{kp}$ is a normal distribution with mean 0. ■

In view of Key Fact A.9, if we standardize the variable $y_p - \hat{y}_p$, the resulting variable has the standard normal distribution. However, because the standardized variable contains the unknown parameter σ, it cannot be used as a basis for a prediction-interval formula. So we replace σ by its estimate s_e, the standard error of the estimate. The resulting variable has a t-distribution.

KEY FACT A.10 *t*-Distribution for Prediction Intervals in Multiple Linear Regression

Suppose the variables $x_1, x_2, \ldots, x_k$ and y satisfy Assumptions 1–4 for multiple linear regression inferences. Then for samples of size n, each with the same values of the predictor variables, the variable

$$t = \frac{y_p - \hat{y}_p}{s_{y_p-\hat{y}_p}}$$

has the t-distribution with $\text{df} = n - (k + 1)$, where $s_{y_p-\hat{y}_p} = s_e \cdot \sqrt{1 + d_p^2}$. ■

Using Key Fact A.10, we can derive the following procedure for obtaining a prediction interval.

Procedure A.5 The t-Interval Procedure for a Prediction of the Response Variable in Multiple Linear Regression

Assumptions

The four assumptions for multiple linear regression inferences

Step 1 For a prediction level of $1 - \alpha$, use Table IV to find $t_{\alpha/2}$ with $\text{df} = n - (k + 1)$.

Step 2 Compute the predicted value, $\hat{y}_p = b_0 + b_1x_{1p} + b_2x_{2p} + \cdots + b_kx_{kp}$, of the response variable corresponding to the particular values $x_{1p}, x_{2p}, \ldots, x_{kp}$ of the k predictor variables.

Step 3 The endpoints of the prediction interval for the observed value of the response variable are

$$\hat{y}_p \pm t_{\alpha/2} \cdot s_{y_p - \hat{y}_p},$$

where $s_{y_p-\hat{y}_p} = s_e \cdot \sqrt{1 + d_p^2}$.

Step 4 Interpret the prediction interval.

It should be noted that if the estimated standard deviation of $y_p - \hat{y}_p$ is squared, we obtain

$$s^2_{y_p-\hat{y}_p} = s_e^2 \cdot (1 + d_p^2) = s_e^2 + s_e^2 \cdot d_p^2.$$

However, from Key Fact A.8 on page A-51, we can see that $s_e^2 \cdot d_p^2$ is the square of the estimated standard deviation of $\hat{y}_p$, that is, $s_e^2 \cdot d_p^2 = s^2_{\hat{y}_p}$. Thus

$$s^2_{y_p-\hat{y}_p} = s_e^2 + s_e^2 \cdot d_p^2 = s_e^2 + s^2_{\hat{y}_p},$$

or, by taking square roots of both sides, we obtain

$$s_{y_p-\hat{y}_p} = \sqrt{s_e^2 + s^2_{\hat{y}_p}}.$$

So if we know s_e and $s_{\hat{y}_p}$, we can determine $s_{y_p-\hat{y}_p}$ from the previous equation.

■

Example A.16 *Illustrates Procedure A.5: Orion Prices*

Refer to the data on age, miles driven, and price for a sample of 11 Orions in Table A.1 on page A-11. Obtain a 95% prediction interval for the price of an Orion that is 5 years old and has been driven 52,000 miles.

Solution We apply Procedure A.5.

Step 1 *For a prediction level of* $1 - \alpha$, *use Table IV to find* $t_{\alpha/2}$ *with* $df = n - (k + 1)$.

We want a 95% prediction interval, so $\alpha = 0.05$. Also since $n = 11$ and $k = 2$, we have $\text{df} = n - (k + 1) = 11 - (2 + 1) = 8$. Consulting Table IV we find that $t_{\alpha/2} = t_{0.05/2} = t_{0.025} = 2.306$.

Step 2 *Compute the predicted value,* $\hat{y}_p = b_0 + b_1x_{1p} + b_2x_{2p} + \cdots + b_kx_{kp}$, *of the response variable corresponding to the particular values* $x_{1p}, x_{2p}, \ldots, x_{kp}$ *of the k predictor variables.*

The sample multiple linear regression equation is given in Example A.5 on page A-13 as $\hat{y} = 183 - 9.50x_1 - 0.821x_2$, where x_1 is age (in years) and x_2 is mileage (in thousands of miles). The predicted price at $x_1 = 5$ years and $x_2 = 52$ thousand miles is therefore

$$\hat{y}_p = 183 - 9.50 \cdot 5 - 0.821 \cdot 52 = 92.80.$$

Step 3 *The endpoints of the prediction interval for the observed value of the response variable are*

$$\hat{y}_p \pm t_{\alpha/2} \cdot s_{y_p - \hat{y}_p}.$$

We can compute $s_{y_p - \hat{y}_p}$ by using the formula

$$s_{y_p - \hat{y}_p} = \sqrt{s_e^2 + s_{\hat{y}_p}^2},$$

which is given directly after Procedure A.5. From Printout A.12 on page A-49, we have that `S = 8.805` is the standard error of the estimate, which is given as the first entry in the seventh line of the printout. Thus $s_e^2 = (8.805)^2 = 77.5$. Note that since $s_e^2 = MSE$, we can also obtain s_e^2 from the analysis of variance table in Printout A.12.

Also, we know that $s_{\hat{y}_p}$ is given in Printout A.12 in the fourth line from the bottom under the heading `SE Fit`. Thus $s_{\hat{y}_p} = 2.74$, and $s_{\hat{y}_p}^2 = (2.74)^2 = 7.51$. So the estimated standard deviation of $y_p - \hat{y}_p$ is

$$s_{y_p - \hat{y}_p} = \sqrt{s_e^2 + s_{\hat{y}_p}^2} = \sqrt{77.5 + 7.51} = \sqrt{85.01} = 9.22.$$

Thus the endpoints of the 95% prediction interval are

$$92.80 \pm 2.306 \cdot 9.22,$$

or 71.54 to 114.06.

This required prediction interval is also given as the final item in the fourth line from the bottom of Printout A.12, the item under the heading `95.0% PI`. The slight difference of 0.01 in the lower endpoint of this interval compared to the one we obtained in the previous paragraph is due to rounding error in the calculation of $s_{y_p - \hat{y}_p}$ given above.

Step 4 *Interpret the prediction interval.*

What Does it Mean ?

We can be 95% certain that the observed price of an Orion that is 5 years old and has been driven 52,000 miles will be somewhere between $7,154 and $11,406.

◆

EXTRAPOLATION

We discussed the problem of **extrapolation** in simple linear regression in Section 14.2. In simple linear regression, extrapolation is the act of using a regression equation to make predictions of the response variable outside of the range of the observed values of the single predictor variable. Highly inaccurate predictions can result from extrapolation, as is illustrated in Section 14.2.

In simple linear regression, it is rather easy to detect whether extrapolation is taking place. For instance, in our example of predicting the price of an Orion based upon its age, the observed range of ages is from 2 to 7 years. If we predict the price of an 11-year-old Orion, we are clearly extrapolating because 11 years is outside the range from 2 to 7 years.

In multiple linear regression, we define extrapolation as the act of using a regression equation to make predictions outside the region of the observed values of the predictor variables. Here we need to look at the region obtained by considering simultaneously the observed values of all the predictor variables, not just the range of the observed values of each predictor variable individually.

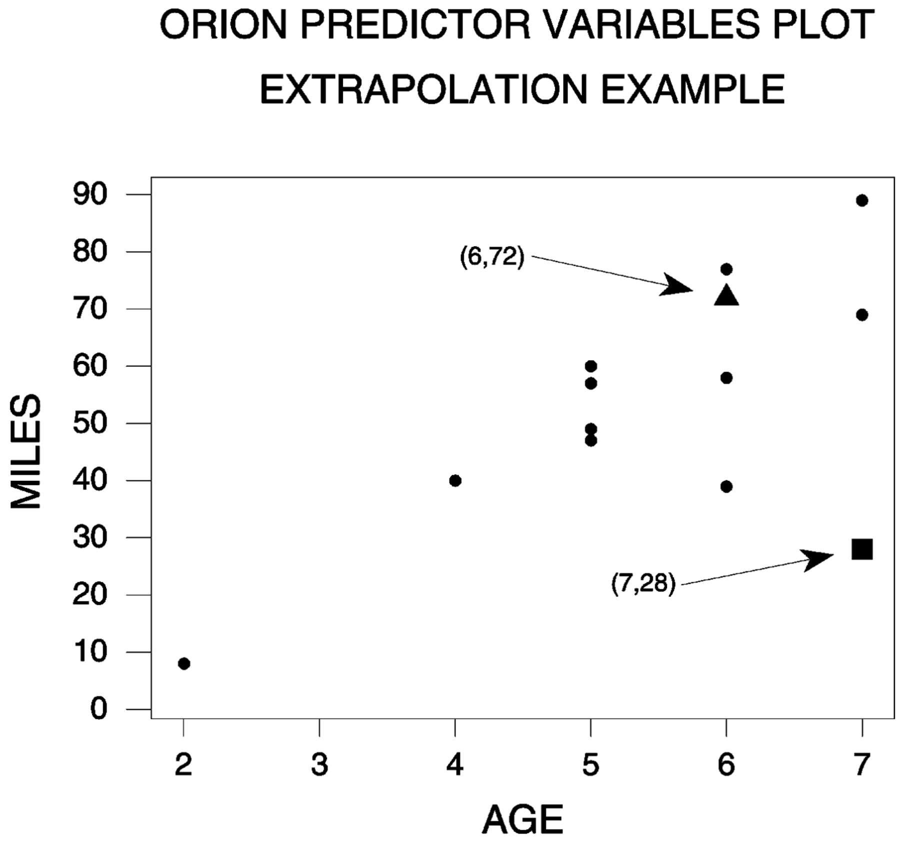
To illustrate, let's return to our example of predicting the price of an Orion based on its age and mileage. The region over which it is reasonable to make predictions about the price of an Orion is the collection of values for age and mileage contained within the swarm of observed age-mileage points shown in the scatterplot of mileage versus age in Fig. A.7. For instance, it is appropriate to predict the price of an Orion that is 6 years old and has been driven 72,000 miles (plotted as the triangle in Fig. A.7) because the point (6, 72) falls within the region of the observed age-mileage points.

Consider now an Orion that is 7 years old and has been driven 28,000 miles. Is it appropriate to predict the price of this Orion? The age of the Orion (7 years) falls within the range of observed ages (2 to 7 years), and the mileage falls within the range of observed miles driven (8 to 89 thousand miles). However, it is not enough to have each predictor variable value at which prediction is to be done fall within the range of observed values of that predictor. The point (7, 28) is plotted as a square in Fig. A.7. Clearly this point has an age and mileage combination that is outside the region of observed predictor variable points. Thus to predict the price of such an Orion by using our regression equation would be extrapolation.

With more predictors it becomes more difficult to discover whether extrapolation is occurring. We can determine whether the value of each predictor variable falls in the range of values observed for each respective predictor. We can also check whether the point determined by each pair of predictor variables' values falls within the region in two dimensions determined by the scatterplot of observed points for each respective pair of predictor variables. However, it is not enough to simply check each predictor variable and each pair of predictor variables. We must consider the k-dimensional scatterplot of observed predictor variable points. This is not possible if there are more than three predictor variables.

Extrapolation can be difficult to detect when there are many predictors. The problem is made worse when there are moderate to high correlations among

Figure A.7
Prediction and extrapolation in the Orion data set



the predictor variables. This is the case with the Orion data set since age and mileage are correlated—as age increases, mileage tends to increase. Irregularly shaped regions of the set of observed predictor variable points also make it difficult to detect extrapolation.

An indication that extrapolation might be taking place is a wide confidence interval for the conditional mean of the response variable at the specified predictor variable values relative to the widths of the confidence intervals for the conditional means of the response variable at the observed values of the predictors.

The Technology Center

Some statistical technologies have programs that will provide predicted values, confidence intervals for the conditional mean of the response variable, and prediction intervals. Minitab is the only technology featured here that provides such output.

Recall the data in Table A.1 on page A-11 that gives data on age, miles driven, and price of a sample of 11 Orions. The sample regression equation that relates price to age and miles driven is given in Printout A.1 on page A-14 and Printout A.12 on page A-49. Printout A.12 also provides information about the predicted value, confidence interval for the conditional mean, and prediction interval for an Orion that is five years old and has been driven 52,000 miles. The predicted values, confidence interval for the conditional mean, and the prediction interval are discussed in Example A.14 (page A-48), Example A.15 (page A-52), and Example A.16 (page A-55), respectively.

Obtaining the Output (Optional)

Here are detailed instructions for obtaining the predicted value, confidence interval, and prediction interval given in the Minitab output in Printout A.12. First we store the data on age, miles driven, and price in columns named AGE, MILES, and PRICE, respectively. Then we proceed as follows.

MINITAB

1 Choose **Stat ➤ Regression ➤ Regression...**
2 Specify Price in the **Response** text box
3 Click in the **Predictors** text box and specify AGE and MILES
4 Click the **Options...** button
5 Click in the **Prediction intervals for new observations** text box and type 5 52
6 Click in the **Confidence level** text box and type 95
7 Click **OK**
8 Click the **Results...** button
9 Select the **Regression equation, table of coefficients, s, R-squared, and basic analysis of variance** option button
10 Click **OK**
11 Click **OK**

Exercises A.5

Statistical Concepts and Skills

A.87 What two regression inferences did we discuss in this section? What assumptions are required for such inferences?

A.88 State whether each of the following is true or false and explain your answer.

a. The estimate of the mean of the response variable at a particular set of predictor variable values is the same as the predicted value of the response variable based on the sample multiple linear regression equation.

b. The prediction interval for the value of the response variable at a particular set of values of the predictor variables is wider than the confidence interval for the conditional mean of the response variable at the same set of values of the predictor variables.

A.89 A sample multiple linear regression equation predicts the response variable y to be 87.62 at values of the predictor variables $x_1 = 5$, $x_2 = 12$, and $x_3 = 27$. What is the estimated value of the conditional mean of y at $x_1 = 5$, $x_2 = 12$, and $x_3 = 27$?

A.90 Answer true or false to the following statements and explain your answers.

a. In multiple linear regression, we can determine whether we are extrapolating in predicting the value of the response variable for a given set of predictor variable values by determining whether each predictor variable value falls in the range of observed values of that predictor.

b. Irregularly shaped regions of the values of predictor variables are easy to detect with two-dimensional scatterplots of pairs of predictor variables, and thus it is easy to determine whether we are extrapolating when predicting the response variable.

A.91 Advertising and Sales. Refer to Exercise A.43 on page A-20 regarding the household-appliance manufacturer that wants to analyze the relationship between total sales and the company's expenditures on three primary means of advertising (television, magazines, and radio). Printout A.5 on page A-23 shows output that results by applying Minitab's regression procedure to these data. The confidence and prediction intervals are for television advertising of $9.5 million, magazine advertising of $4.3 million, and radio advertising of $5.2 million.

a. Obtain a point estimate for mean sales when the amounts spent on television, magazine, and radio advertising are $9.5 million, $4.3 million, and $5.2 million, respectively.

b. Find a 95% confidence interval for mean sales when the amounts spent on television, magazine, and radio advertising are $9.5 million, $4.3 million, and $5.2 million, respectively.

c. Determine the predicted sales if the amounts spent on television, magazine, and radio advertising are $9.5 million, $4.3 million, and $5.2 million, respectively.

d. Determine a 95% prediction interval for sales if the amounts spent on television, magazine, and radio advertising are $9.5 million, $4.3 million, and $5.2 million, respectively.

e. In part (c) we predicted total sales for expenditures of $9.5 million, $4.3 million, and $5.2 million on television, magazine, and radio advertising, respectively. Is this prediction appropriate, or are we extrapolating? (*Hint:* Find the point representing these expenditures on the pairwise plots of the predictor variables.)

f. Is it appropriate to predict total sales for expenditures of $7.1 million, $3.5 million, and $6.1 million on television, magazine, and radio advertising, respectively?

A.92 Corvette Prices. Refer to Exercise A.44 on page A-20 regarding the relationship between the price of a Corvette and its age and number of miles driven. Printout A.7 on page A-24 shows output that results by applying Minitab's regression procedure to these data. The confidence and prediction intervals are for Corvettes that are 4 years old and have been driven 28,000 miles.

a. Obtain a point estimate for the mean price of all Corvettes that are 4 years old and have been driven 28,000 miles.

b. Obtain a 95% confidence interval for the mean price of all Corvettes that are 4 years old and have been driven 28,000 miles.

c. Determine the predicted price of a Corvette that is 4 years old and has been driven 28,000 miles.

d. Find a 95% prediction interval for the price of a randomly selected Corvette that is 4 years old and has been driven 28,000 miles.

e. In part (a) we estimated the mean price of all Corvettes that are 4 years old and have been driven 28,000 miles. Is this estimation appropriate, or are we extrapolating? (*Hint:* Find the point representing the age and miles driven on the pairwise plot of the predictor variables.)

f. Is it appropriate to predict the price of a Corvette that is 10 years old and has been driven 55,000 miles?

A.93 Graduation Rates. Refer to Exercise A.45 on page A-20 regarding the relationship between college graduation rate and the predictor variables student-to-faculty ratio, percentage of freshmen in the top 10% of their high-school class, and percentage of applicants accepted. Printout A.9 on page A-25 shows output that results by applying Minitab's regression procedure to these data. The confidence and prediction intervals are for the case where the student-to-faculty ratio is 18 to 1, 70% of the freshmen were in the top 10% of their high-school class, and 75% of the applicants are accepted.

a. Obtain a point estimate for the mean graduation rate of all schools where the student-to-faculty ratio is 18 to 1, 70% of the freshmen were in the top 10% of their high-school class, and 75% of the applicants are accepted.

b. Determine a 95% confidence interval for the mean graduation rate of all schools where the student-to-faculty ratio is 18 to 1, 70% of the freshmen were in the top 10% of their high-school class, and 75% of the applicants are accepted.

c. Determine the predicted graduation rate at a randomly selected school where the student-to-faculty ratio is 18 to 1, 70% of the freshmen were in the top 10% of their high-school class, and 75% of the applicants are accepted.

d. Determine a 95% prediction interval for the graduation rate at a randomly selected school where the student-to-faculty ratio is 18 to 1, 70% of the freshmen were in the top 10% of their high-school class, and 75% of the applicants are accepted.

e. In solving parts (a) and (b), what assumptions are you making?

A.94 Custom Home Resales. Refer to Exercise A.46 on page A-21 regarding predicting the selling price of a home in the Equestrian Estates using the predictor variables square footage, number of bedrooms, number of bathrooms, and number of days on the market. Printout A.11 on page A-26 shows output that results by applying Minitab's regression procedure to these data. The confidence and prediction intervals are for homes that have 3200 sq ft, 4 bedrooms, 3 bathrooms, and that remain on the market for 60 days.

a. Obtain a point estimate for the mean selling price of all homes in the Equestrian Estates that have 3200 sq ft, 4 bedrooms, 3 bathrooms, and that remain on the market for 60 days.

b. Obtain a 95% confidence interval for the mean selling price of all homes in the Equestrian Estates that have 3200 sq ft, 4 bedrooms, 3 bathrooms, and that remain on the market for 60 days.

c. Determine the predicted selling price of a randomly selected home in the Equestrian Estates that has 3200 sq ft, 4 bedrooms, 3 bathrooms, and that remains on the market for 60 days.

d. Determine a 95% prediction interval for the selling price of a randomly selected home in the Equestrian Estates that has 3200 sq ft, 4 bedrooms, 3 bathrooms, and that remains on the market for 60 days.

e. In solving parts (a) and (b), what assumptions are you making?

Using Technology

A.95 Advertising and Sales. Referring to Exercise A.91, use the technology of your choice to obtain the requested point estimate and confidence interval for the mean total sales, and the predicted value and prediction interval for total sales for television advertising of $9.5 million, magazine advertising of $4.3 million, and radio advertising of $5.2 million.

A.96 Corvette Sales. Referring to Exercise A.92, use the technology of your choice to obtain the requested point estimate and confidence interval for the mean price, and the predicted value and prediction interval for the price of a Corvette that is 4 years old and has been driven 28,000 miles.

A.97 Graduation Rates. Referring to Exercise A.93, use the technology of your choice to obtain the requested point estimate and confidence interval for the mean graduation rate, and the predicted value and prediction interval for the graduation rate of a school where the student-to-faculty ratio is 18 to 1, 70% of the freshmen were in the top 10% of their high-school class, and 75% of the applicants are accepted.

A.98 Custom Home Resales. Referring to Exercise A.94, use the technology of your choice to obtain the requested point estimate and confidence interval for the mean selling price, and the predicted value and prediction interval for the selling price of a home in the Equestrian Estates that has 3200 sq ft, 4 bedrooms, 3 bathrooms, and that remains on the market for 60 days.

A.6 CHECKING MODEL ASSUMPTIONS AND RESIDUAL ANALYSIS

In simple linear regression, we saw that in order to assess the assumptions of the regression model, we analyze the residuals, which are the differences between the observed values of the response variable and the values predicted by the sample regression equation:

$$\textbf{Residual} = e = y - \hat{y}.$$

Likewise, the residuals from a multiple linear regression can be used to examine the assumptions made for performing inferences in multiple linear regression. We use residuals to assess:

1. Whether the multiple linear regression equation is a suitable description of the relationship between the response variable, y, and the predictor variables, $x_1, \dots, x_k$.
2. Whether the conditional standard deviation of the response variable is constant over all possible predictor values.

3. Whether for each set of values of the predictor variables, the distribution of the response variable is normal.
4. Whether the observations are independent.

We will concentrate here on the evaluation of Assumptions 1–3. Assessment of the independence assumption is very difficult in most cases and generally requires that we have additional information about the data, such as the time sequence in which the data were collected. We will not consider methods for checking the independence assumption.

Additionally, residuals may be used to detect **outliers** or atypical observations in the data. Outliers can have a great effect on the sample regression coefficients and on statistical inferences made about the model parameters.

Recall from simple linear regression that the residuals from a least squares regression analysis always sum to 0, that is, $\Sigma e = 0$, and thus $\bar{e} = 0$. This also holds true for the residuals of a least squares multiple linear regression analysis. It can also be shown that the residuals are uncorrelated with each of the predictor variables and with the predicted values of the response variable. These facts along with the assumptions for the regression model lead to an understanding of the expected behavior of plots of the residuals if the regression model assumptions are true. **Residual plots** can be used as diagnostic tools for assessing the validity of the regression model assumptions.

In order to check for the suitability of the regression equation, we plot the residuals versus the predicted values of the response variable (e versus $\hat{y}$), and the residuals versus each of the predictor variables (e versus x_i, for $i = 1, 2, \ldots, k$). If the regression equation is suitable, each of these plots should show the residuals roughly centered and symmetric about the horizontal axis. Deviations from this expected pattern indicate that the multiple linear regression equation may not be suitable. Methods for finding a more suitable regression equation will be discussed in Module B.

The constant standard deviation assumption can be checked using the same residual plots as those used to check for the suitability of the regression equation. If the standard deviation is constant across all predictor variable values, the plot of residuals versus predicted values and the plots of residuals versus predictor variables are expected to exhibit roughly constant variation as the predicted values or the predictor variable values change. If the variation in the residuals changes as the predicted values change or as the values of one (or more) of the predictor variables change, the assumption of constant standard deviation comes into question.

A normal probability plot of the residuals may be used to assess the normality assumption. Such a plot should be roughly linear. Departures from linearity indicate possible nonnormal populations.

To check for possible outlier data points, we look for residuals that are large in magnitude. The determination of whether a residual is large is made by obtaining the number of standard deviations a residual is from its mean of 0. The standard deviation of each residual is approximated by the standard error of the estimate, s_e. As a rough rule, we consider any data point with a residual whose absolute value is larger than $2s_e$ as a potential outlier that should be investigated further.

It should be noted that s_e is only an estimate of the standard deviation of a residual. The actual standard deviation of a residual depends on the individual observation and changes from residual to residual. Some statistical programs give the value of the standard deviation of each residual, or give the value of the residual divided by its standard deviation. A residual divided by its standard deviation is often called the *standardized residual.*

KEY FACT A.11 Residual Analysis for Multiple Linear Regression

If the assumptions for multiple linear regression inferences are met, then the following conditions for the residual plots should hold.

- A plot of the residuals against each predictor variable's values should fall roughly in a horizontal band centered and symmetric about the horizontal axis.
- A plot of the residuals against the predicted values of the response variable should fall roughly in a horizontal band centered and symmetric about the horizontal axis.
- A normal probability plot of the residuals should be roughly linear.

Failure of any of these three conditions casts doubt on the validity of one or more of the assumptions for regression inferences for the variables under consideration. ■

If an inappropriate regression equation has been used to fit the data, the residuals contain information about the true population regression equation as well as information about the appropriateness of the constant conditional standard deviation and the normality assumptions. It is important to note that checking the residuals for constant standard deviation and normality should not be done until an appropriate regression equation is obtained.

We refer to the residual plots in Fig. 15.6 to illustrate: (a) a plot revealing no violations of the linearity and constant standard deviation assumptions; (b) a plot indicating that an inappropriate regression equation has been used to fit the data; and (c) a plot indicating that the conditional standard deviation is not constant.

Example A.17 Analysis of Residuals: Orion Prices

Perform a residual analysis on the Orion data in Table A.1 on page A-11, with age and miles driven as predictor variables for price.

Solution We apply Key Fact A.11. The predicted values of the response variable (price) and residuals for the Orion data can be found, respectively, in the fourth and sixth columns of the table near the bottom of Printout A.13 on the following page. (Minitab uses the term Fit instead of *predicted value.*)

From the residual and age data, we obtain the residual plot in Fig. A.8(a) on page A-65; from the residual and miles data, we obtain the residual plot in Fig. A.8(b); and from the residual and predicted-price data, we obtain the residual plot in Fig. A.8(c). Keeping in mind the small sample size, we can say

Printout A.13
Minitab regression output for Orion data with predicted values and residuals

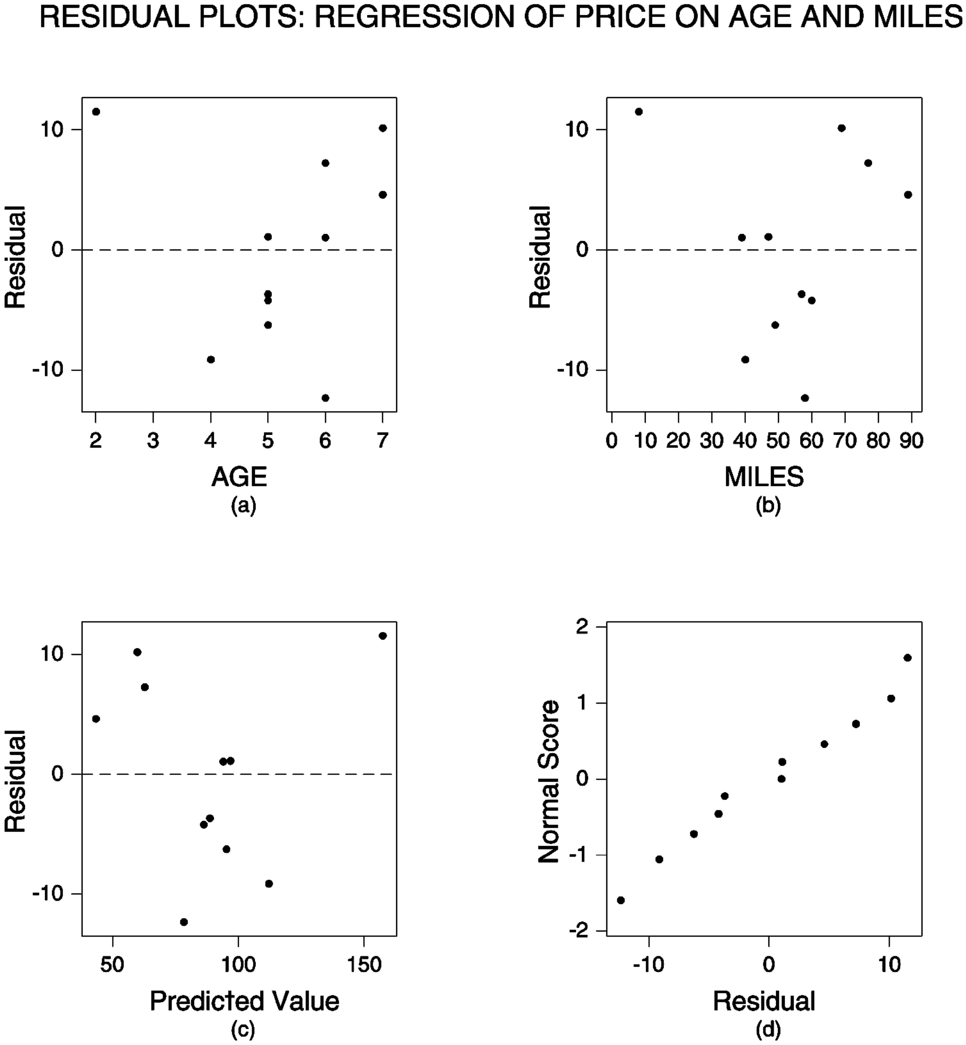
```
The regression equation is
PRICE = 183 - 9.50 AGE - 0.821 MILES

Predictor        Coef     SE Coef          T        P
Constant       183.04       11.35      16.13    0.000
AGE            -9.504       3.874      -2.45    0.040
MILES         -0.8215      0.2552      -3.22    0.012

S = 8.805       R-Sq = 93.6%     R-Sq(adj) = 92.0%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         2      9088.3      4544.2     58.61    0.000
Residual Error     8       620.2        77.5
Total             10      9708.5

Source       DF      Seq SS
AGE           1      8285.0
MILES         1       803.3

Obs       AGE      PRICE         Fit      SE Fit    Residual    St Resid
  1      5.00      85.00       88.69        3.20       -3.69       -0.45
  2      4.00     103.00      112.16        3.71       -9.16       -1.15
  3      6.00      70.00       62.76        4.59        7.24        0.96
  4      5.00      82.00       86.22        3.66       -4.22       -0.53
  5      5.00      89.00       95.26        2.73       -6.26       -0.75
  6      5.00      98.00       96.90        2.84        1.10        0.13
  7      6.00      66.00       78.36        3.32      -12.36       -1.52
  8      6.00      95.00       93.97        6.93        1.03        0.19
  9      2.00     169.00      157.45        6.99       11.55        2.15R
 10      7.00      70.00       59.82        4.71       10.18        1.37
 11      7.00      48.00       43.39        5.35        4.61        0.66

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

that all three plots fall roughly in a horizontal band centered and symmetric about the horizontal axis. Finally, Fig. A.8(d) shows a normal probability plot of the residuals, which is quite linear.

These plots also do not indicate any residuals that are far away from their mean of 0 relative to the standard error of the estimate, s_e. The standard error of the estimate is $s_e = 8.805$, and the residual with the largest magnitude is -12.36 (observation seven). Thus all the residuals are within $2s_e$ of their mean of 0. By using this criterion, we do not have any potential outliers.

Printout A.13 also provides the standardized residuals in the column labeled `St Resid`. These values are the residuals divided by their standard deviations and are different from those obtained by dividing the residual by s_e. If the

Figure A.8 Residual plots for the Orion data set



standardized residuals are used, the residual for observation nine has a standardized value of 2.15. This is slightly above our cutoff value of 2, and might indicate a possible outlier.

This data point (observation nine) is discussed in Chapter 14 when the price of an Orion is regressed only on its age. There it is not considered an outlier, but is seen to have a strong influence on the simple linear regression line. It is an **influential observation**—when this point is deleted from the data set, the new estimated regression coefficients for the regression of price on age show a considerable change from those when this data point is retained.

A similar effect can be seen in the multiple linear regression of price on age and miles if observation nine is omitted. The regression equation becomes `PRICE = 155 - 5.50 AGE - 0.742 MILES`. The *y*-intercept and the coefficient for age change considerably from those obtained with all the data.

What should be done about this data point? We should check that no errors have been made in recording the values of the response and predictor variables or in entering the data into our computer. In this case the data point is not in error. In our analysis we note that it has a large standardized residual and exerts an influence on our regression equation. Observation nine is a rather new car (2 years old) and has been driven only 8000 miles. Since there is so little data in this example, if we want to try to predict the prices of cars that are fairly new and have low mileage, we cannot remove the point from the analysis. A better fitting model might be obtained if more data were collected in the low age and low mileage region.

In the final analysis we conclude that the regression analysis seems to be reasonable. ◆

The Technology Center

Obtaining Predicted Values and Residuals

Most statistical technologies that perform a multiple regression will provide the predicted values and the residuals for each observation. Minitab provides a list of predicted values and residuals, while DDXL provides only a list of the predicted values. We now present output and (optional) step-by-step instructions to obtain the output.

Example A.18 Using Technology to Obtain Predicted Values and Residuals: Orion Prices

Table A.1 on page A-11 gives data on age, miles driven, and price for a sample of 11 Orions. Use Minitab or Excel to perform a multiple regression relating the response variable price to the predictor variables age and miles driven, and obtain the predicted values and residuals.

Solution Printout A.13 on page A-64 gives the usual Minitab regression output plus a table of predicted values (fits), residuals, and related quantities. Printout A.14 shows the predicted values provided by DDXL in addition to the standard multiple regression output. DDXL does not provide the residuals. ◆

Obtaining the Output (Optional)

For a multiple regression based on the data for age, miles driven, and price of 11 Orions in Table A.1, Printout A.12 provides the Minitab output showing the predicted values and residuals, and Printout A.14 provides the DDXL output of the predicted values. Here are detailed instructions for obtaining that output. First we store age, miles driven, and price in columns or ranges, named AGE, MILES, and PRICE, respectively. For Excel, we assume that the data are stored in a worksheet named TbA-1.xls. Then we proceed as follows.

MINITAB

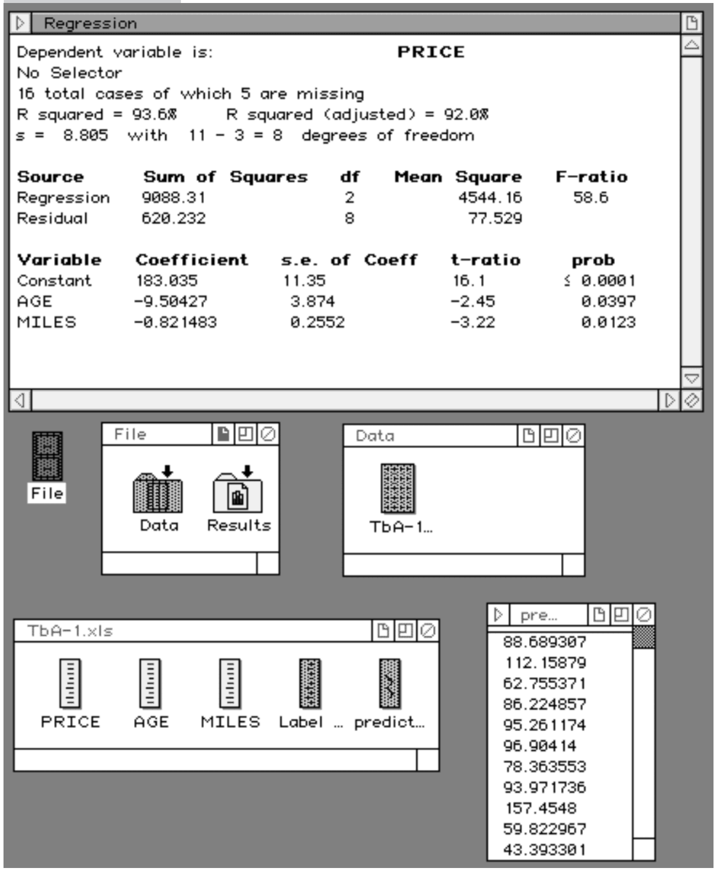
1 Choose **Stat ➤ Regression ➤ Regression...**
2 Specify PRICE in the **Response** text box
3 Click in the **Predictors** text box and specify AGE and MILES
4 Click the **Results...** button
5 Select the **In addition, the full table of fits and residuals** option button
6 Click **OK**
7 Click **OK**

EXCEL

1 Choose **DDXL ➤ Regression**
2 Select **Multiple Regression** from the **Function type** drop-down box
3 Specify PRICE in the **Response Variable** text box
4 Specify AGE and MILES in the **Explanatory Variables** text box
5 Click **OK**
6 Double click the **File** icon
7 Double click the **Data** icon in the **File** folder
8 Double click the **TbA-1...** icon in the **Data** folder
9 Double click the **predict...** icon in the **TbA-1.xls** folder

Printout A.14
Output of predicted values for multiple regression of Orion data

EXCEL



```
Dependent variable is:                    PRICE
No Selector
16 total cases of which 5 are missing
R squared = 93.6%     R squared (adjusted) = 92.0%
s =  8.805  with  11 - 3 = 8  degrees of freedom

Source       Sum of Squares   df   Mean Square   F-ratio
Regression   9088.31           2       4544.16     58.6
Residual     620.232           8        77.529

Variable    Coefficient   s.e. of Coeff   t-ratio    prob
Constant    183.035       11.35           16.1     ≤ 0.0001
AGE         -9.50427       3.874          -2.45      0.0397
MILES       -0.821483      0.2552         -3.22      0.0123
```

```
 88.689307
 112.15879
 62.755371
 86.224857
 95.261174
 96.90414
 78.363553
 93.971736
 157.4548
 59.822967
 43.393301
```

Note that if the sample size is large, using the Minitab option in the **Results...** menu that gives the full table of fits and residuals can generate a great deal of output. For large data sets, only the unusual observations can be obtained by using the option **In addition, sequential sum of squares, and the unusual observations in the table of fits and residuals.** In the Minitab regression procedure, if the absolute value of their standardized residual is 2 or more, an R is placed next to the standardized residual. An X indicates an influential observation.

Obtaining Residual Plots

Most statistical technologies provide a plot of residuals against the predicted values, plots of residuals against the predictor variable values, and a normal probability plot of residuals. These plots are very useful in assessing whether the assumptions for multiple linear regression are appropriate. Minitab provides all of the plots mentioned above, whereas DDXL provides only a plot of residuals against predicted values and a normal probability plot of residuals. We now present output and (optional) step-by-step instructions to obtain the output.

Example A.19 Using Technology to Obtain Residual Plots: Orion Prices

Return to the data for 11 Orions in Example A.18. Use Minitab or Excel to perform a multiple regression relating the response variable price to the predictor variables age and miles driven, and to obtain the plots of residuals against the predicted values, against the values of age, and against the values of miles driven, and also to obtain a normal probability plot of residuals.

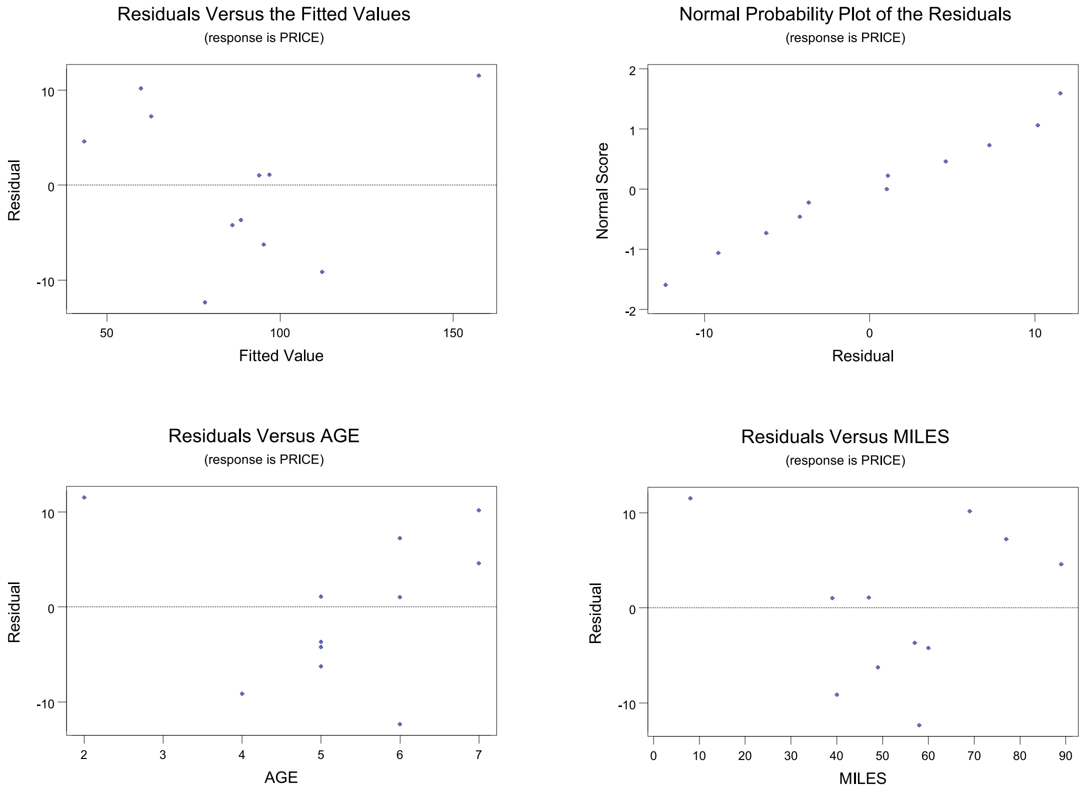
Solution Printout A.15 displays the required four residual plots from Minitab, and two of the four required residual plots from DDXL. Discussion of these residual plots is found in Example A.17 on page A-63. ◆

Obtaining the Output (Optional)

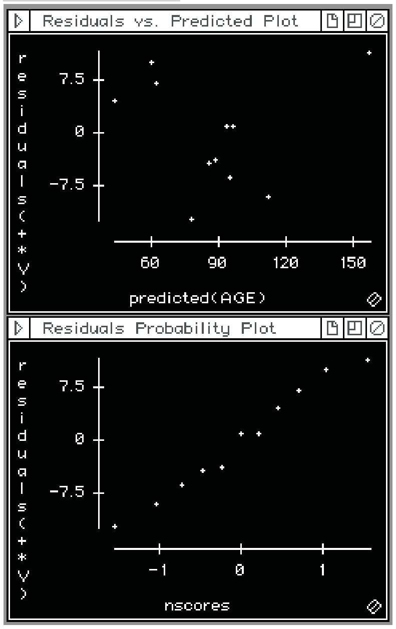
Printout A.15 provides the residual plots that are produced by Minitab and DDXL for the multiple regression of price on age and miles driven for a sample of 11 Orions in Table A.1. Here are detailed instructions for obtaining that output. First we store the data on age, miles driven, and price in columns

Printout A.15
Residual plots for multiple regression of Orion Data

MINITAB



EXCEL



or ranges, named AGE, MILES, and PRICE, respectively. Then we proceed as follows.

MINITAB

1 Choose **Stat ➤ Regression ➤ Regression ...**
2 Specify PRICE in the **Response** text box
3 Click in the **Predictors** text box and specify AGE and MILES
4 Click the **Graphs...** button
5 Select the **Normal plot of residuals** checkbox
6 Select the **Residuals versus fits** check box
7 Click in the **Residuals versus the variables** text box and specify AGE and MILES
8 Click **OK**
9 Click **OK**

EXCEL

1 Choose **DDXL ➤ Regression**
2 Select **Multiple Regression** from the **Function type** drop-down box
3 Specify PRICE in the **Response Variable** text box
4 Specify AGE and MILES in the **Explanatory Variables** text box
5 Click **OK**
6 Click the **Check Residuals** button

Exercises A.6

Statistical Concepts and Skills

A.99 Fill in the blanks.

a. In multiple linear regression analysis, a residual is the difference between an ______ and ______.

b. A plot of the residuals versus the values of a predictor variable should fall roughly in a ______ centered and symmetric about the ______, if the assumptions for multiple linear regression inferences are met.

c. The sum of the residuals from the least squares multiple linear regression is ______.

A.100 Describe the difference between a residual and a standardized residual. Which of these two residuals is more useful in detecting outliers?

A.101 Fill in the blanks.

a. In multiple linear regression, a(n) ______ is a data point that lies far from the multiple linear regression plane, relative to the other data points.

b. In multiple linear regression, a(n) ______ is a data point whose removal causes the multiple linear regression equation to change considerably.

A.102 Answer true or false to the following statements and explain your answers.

a. In a multiple linear regression analysis, a data point with a large standardized residual should be removed from the data set, and the regression equation should be recalculated.

b. In a multiple linear regression analysis, if a plot of the residuals against a predictor variable shows curvature, the normality assumption for regression inferences is violated.

A.103 Advertising and Sales. Refer to Exercise A.43 on page A-20 regarding the household-appliance manufacturer that wants to analyze the relationship between total sales and the company's expenditures on three primary means of advertising (television, magazines, and radio).

a. Printout A.16 on page A-72 shows output that results by applying Minitab's regression procedure to these data. Printouts A.17(a), (b), (c), and (d) on page A-73 display, respectively, plots of residuals against television advertising expenditures, residuals against magazine advertising expenditures, residuals against radio advertising expenditures, and residuals against predicted total sales.

Printout A.17(e) shows a normal probability plot of the residuals. Perform a residual analysis to assess the assumptions of linearity of the regression equation, constancy of the conditional standard deviation, and normality of the conditional distributions. Check for outliers and influential observations.

b. Does your analysis in part (a) reveal any violations of the assumptions for multiple regression inferences? Explain your answer.

A.104 Corvette Prices. Refer to Exercise A.44 on page A-20 regarding the relationship between the price of a Corvette and its age and number of miles driven.

a. Printout A.18 on page A-74 shows output that results by applying Minitab's regression procedure to these data. Printouts A.19(a), (b), and (c) on page A-75 display, respectively, plots of residuals against age, residuals against miles driven, and residuals against predicted price. Printout A.19(d) shows a normal probability plot of the residuals. Perform a residual analysis to assess the assumptions of linearity of the regression equation, constancy of the conditional standard deviation, and normality of the conditional distributions. Check for outliers and influential observations.

b. Does your analysis in part (a) reveal any violations of the assumptions for multiple regression inferences? Explain your answer.

Using Technology

A.105 Advertising and Sales. Refer to Exercise A.43 on page A-20 regarding the household-appliance manufacturer that wants to analyze the relationship between total sales and the company's expenditures on three primary means of advertising (television, magazines, and radio).

a. Obtain output similar to that in Printout A.13 on page A-64 and Fig. A.8 on page A-65.

b. Perform a residual analysis to assess the assumptions of linearity of the regression equation, constancy of the conditional standard deviation, and normality of the conditional distributions. Check for outliers and influential observations.

c. Does your analysis in part (b) reveal any violations of the assumptions for multiple regression inferences? Explain your answer.

A.106 Corvette Prices. Refer to Exercise A.44 on page A-20 regarding the relationship between the price of a Corvette and its age and number of miles driven.

a. Obtain output similar to that in Printout A.13 on page A-64 and Fig. A.8 on page A-65.

b. Perform a residual analysis to assess the assumptions of linearity of the regression equation, constancy of the conditional standard deviation, and normality of the conditional distributions. Check for outliers and influential observations.

c. Does your analysis in part (b) reveal any violations of the assumptions for multiple regression inferences? Explain your answer.

A.107 Graduation Rates. Refer to Exercise A.45 on page A-20 regarding the relationship between college graduation rate and the predictor variables student-to-faculty ratio, percentage of freshmen in the top 10% of their high-school class, and percentage of applicants accepted.

a. Obtain output similar to that in Printout A.13 on page A-64 and Fig. A.8 on page A-65.

b. Perform a residual analysis to assess the assumptions of linearity of the regression equation, constancy of the conditional standard deviation, and normality of the conditional distributions. Check for outliers and influential observations.

c. Does your analysis in part (b) reveal any violations of the assumptions for multiple regression inferences? Explain your answer.

A.108 Custom Homes Resales. Refer to Exercise A.46 on page A-21 regarding predicting the selling price of a home in the Equestrian Estates using the predictor variables square footage, number of bedrooms, number of bathrooms, and number of days on the market.

a. Obtain output similar to that in Printout A.13 on page A-64 and Fig. A.8 on page A-65.

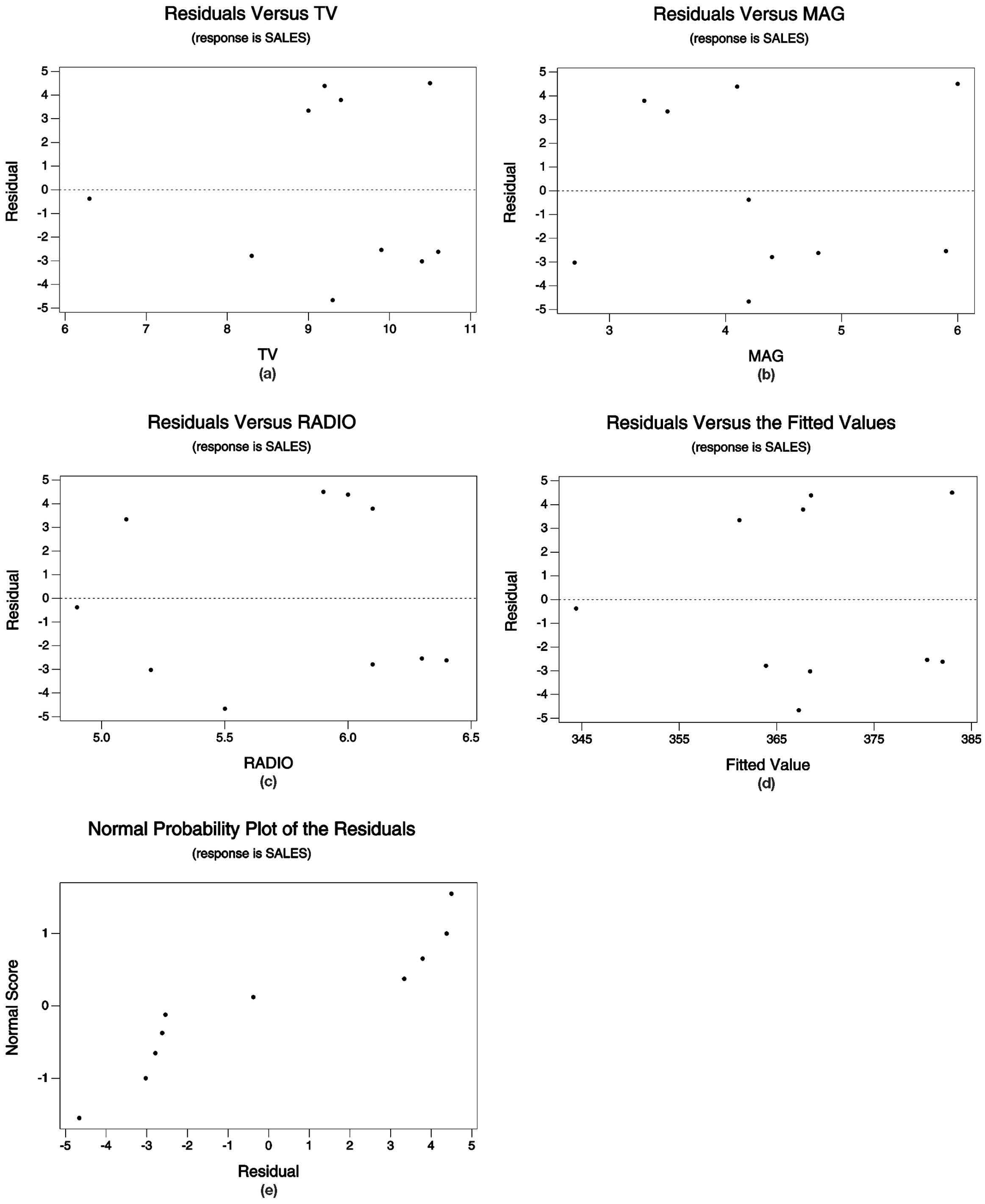
b. Perform a residual analysis to assess the assumptions of linearity of the regression equation, constancy of the conditional standard deviation, and normality of the conditional distributions. Check for outliers and influential observations.

c. Does your analysis in part (b) reveal any violations of the assumptions for multiple regression inferences? Explain your answer.

Printout A.16
Minitab output for Exercise A.103

```
The regression equation is
SALES = 266 + 6.73 TV + 3.26 MAG + 4.51 RADIO

Predictor        Coef     SE Coef          T        P
Constant       266.23       16.34      16.29    0.000
TV              6.727       1.344       5.01    0.002
MAG             3.257       1.642       1.98    0.095
RADIO           4.507       3.703       1.22    0.269

S = 4.418       R-Sq = 91.1%     R-Sq(adj) = 86.6%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         3     1194.53      398.18     20.40    0.002
Residual Error     6      117.11       19.52
Total              9     1311.64

Source       DF      Seq SS
TV            1      991.57
MAG           1      174.04
RADIO         1       28.92

Obs         TV      SALES         Fit      SE Fit    Residual    St Resid
  1        8.3     361.10      363.89        2.63       -2.79       -0.79
  2        6.3     344.00      344.38        3.89       -0.38       -0.18
  3        9.9     377.90      380.44        2.66       -2.54       -0.72
  4        9.4     371.50      367.71        2.88        3.79        1.13
  5       10.4     365.40      368.43        3.48       -3.03       -1.11
  6        9.0     364.50      361.16        2.38        3.34        0.90
  7        9.2     372.90      368.52        1.84        4.38        1.09
  8       10.6     379.40      382.02        2.38       -2.62       -0.70
  9        9.3     362.60      367.26        1.64       -4.66       -1.14
 10       10.5     387.50      383.00        3.34        4.50        1.55
```

Printout A.17
Residual plots for Exercise A.103



(a)

(b)

(c)

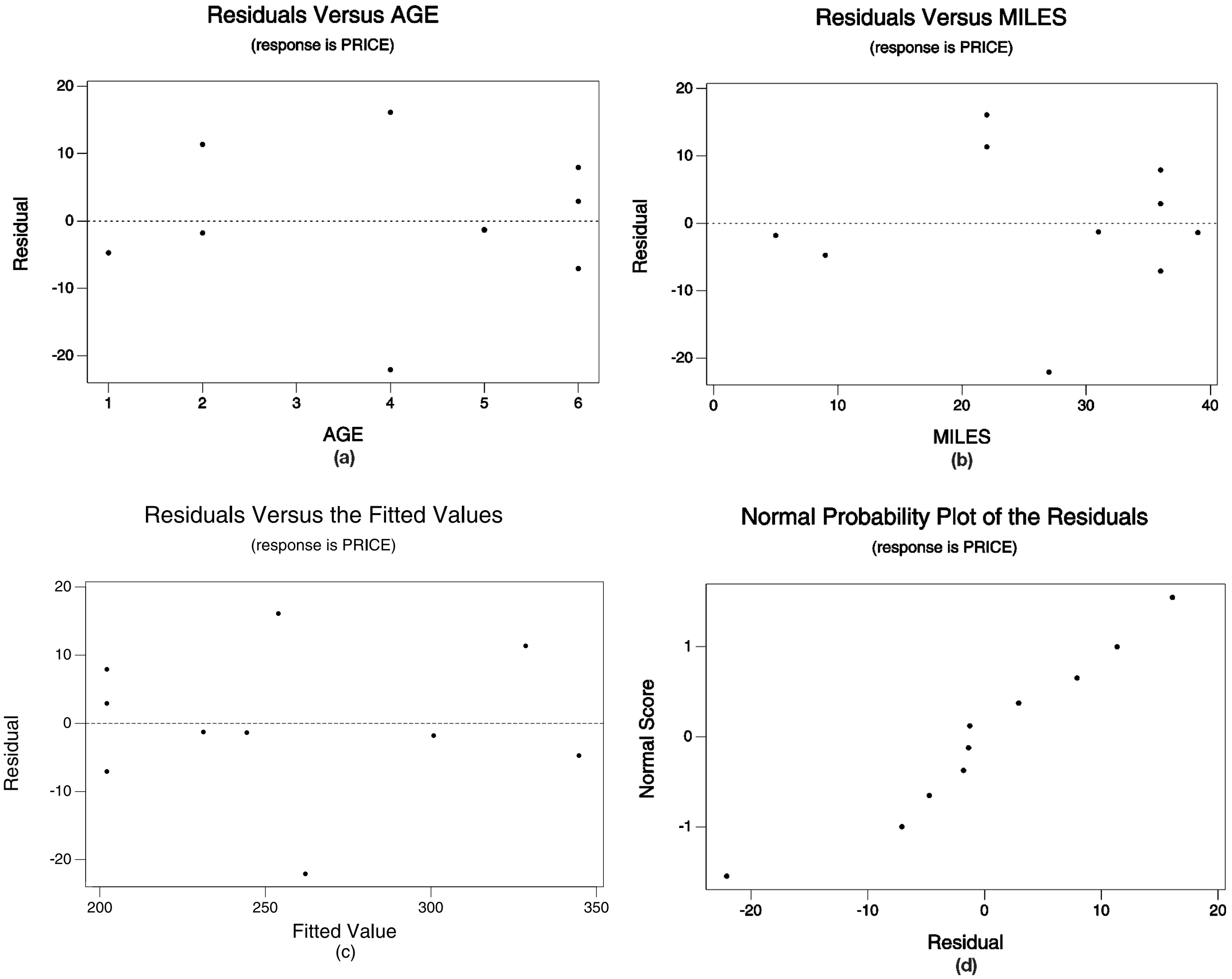
(d)

(e)

Printout A.18
Minitab output for Exercise A.104

```
The regression equation is
PRICE = 367 - 37.4 AGE + 1.64 MILES

Predictor        Coef     SE Coef          T        P
Constant      367.362       9.943      36.95    0.000
AGE           -37.375       5.174      -7.22    0.000
MILES          1.6378      0.8116       2.02    0.083

S = 12.11       R-Sq = 96.0%     R-Sq(adj) = 94.9%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         2       24655       12328     84.07    0.000
Residual Error     7        1026         147
Total              9       25682

Source       DF      Seq SS
AGE           1       24058
MILES         1         597

Obs        AGE       PRICE         Fit      SE Fit    Residual    St Resid
  1       6.00      205.00      202.07        5.73        2.93        0.27
  2       6.00      195.00      202.07        5.73       -7.07       -0.66
  3       6.00      210.00      202.07        5.73        7.93        0.74
  4       2.00      340.00      328.64        8.72       11.36        1.35
  5       2.00      299.00      300.80        9.53       -1.80       -0.24
  6       5.00      230.00      231.26        4.32       -1.26       -0.11
  7       4.00      270.00      253.89        4.88       16.11        1.45
  8       5.00      243.00      244.36        7.45       -1.36       -0.14
  9       1.00      340.00      344.73        7.78       -4.73       -0.51
 10       4.00      240.00      262.08        3.97      -22.08       -1.93
```

Printout A.19
Residual plots for Exercise A.104



Residuals Versus AGE
(response is PRICE)

(a)

Residuals Versus MILES
(response is PRICE)

(b)

Residuals Versus the Fitted Values
(response is PRICE)

(c)

Normal Probability Plot of the Residuals
(response is PRICE)

(d)

Module Review

You Should Be Able To

1. use and understand the formulas presented in this module.

2. define and apply the concepts related to linear equations with more than one independent variable.

3. explain the least-squares criterion.

4. define and use the terminology *predictor variable* and *response variable*.

5. understand when it is appropriate to obtain a multiple linear regression equation for a set of data points.

6. interpret the regression parameter estimates and use the sample multiple linear regression equation to make predictions.

7. understand the difference between the population multiple linear regression equation and a sample multiple linear regression equation.

8. interpret the three sum of squares, *SST*, *SSE*, and *SSR*.

9. interpret the coefficient of multiple determination, R^2.

10. state the assumptions for multiple linear regression inferences.

11. interpret the information in an analysis of variance table.

12. determine the standard error of the estimate from *MSE*.

13. perform a hypothesis test to decide whether a specified collection of predictor variables taken together is useful for predicting the response variable.

14. perform a hypothesis test to decide whether a population regression parameter, β_i, of a population multiple linear regression equation is not zero, and hence whether the predictor x_i is useful for predicting the response variable.

15. obtain a confidence interval for β_i.

16. obtain a point estimate and a confidence interval for the conditional mean of the response variable corresponding to particular values of the predictor variables.

17. obtain a predicted value and a prediction interval for the response variable corresponding to particular values of the predictor variables.

18. understand the concept of extrapolation.

19. perform a residual analysis to check the assumptions for multiple linear regression inferences.

20. identify outliers and influential observations.

★ **21.** use a statistical technology to perform the analyses covered in this module.

★ **22.** interpret the output obtained from a statistical technology for the analyses discussed in this module.

Key Terms

analysis of variance table, *A-33*
coefficient of multiple determination (R^2), *A-28*
conditional distribution, *A-30*
conditional mean, *A-7*
conditional standard deviation, *A-30*
error sum of squares (*SSE*), *A-27*
error term (ϵ), *A-30*
extrapolation, *A-57*
influential observation, *A-65*
linear equation in several independent variables, *A-6*
mean square for error (*MSE*), *A-33*
mean square for regression (*MSR*), *A-33*
method of least squares, *A-13*
multiple linear regression, *A-4*
multiple linear regression equation, *A-7*
multiple linear regression model, *A-7*
outlier, *A-62*
partial slope, *A-6*
plane, *A-4*
population regression equation, *A-7*, *A-30*
population regression plane, *A-30*
prediction interval, *A-53*
predictor variable, *A-4*
regression identity for degrees of freedom, *A-32*
regression identity for multiple linear regression, *A-28*
regression sum of squares (*SSR*), *A-27*
residual, *A-61*
residual plot, *A-62*
response variable, *A-4*
sample regression equation, *A-13*
sampling distribution of a sample regression coefficient, *A-41*
scatterplot matrix, *A-10*
simple linear regression, *A-2*
standard error of the estimate (s_e), *A-31*
total sum of squares (*SST*), *A-27*
y-intercept, *A-6*

Review Test

Statistical Concepts and Skills

1. For a linear equation $y = b_0 + b_1x_1 + b_2x_2 + b_3x_3$, identify the

a. independent variables.
b. dependent variable.
c. y-intercept.
d. partial slopes.

2. Consider the linear equation $y = 5 + 4x_1 - 3x_2$.

a. At what y-value does it intersect the y-axis?
b. At what x_1- and x_2-values does it intersect the y-axis?
c. What are the partial slopes?
d. By how much will the y-value on the plane change when x_1 increases by 1 unit and x_2 remains unchanged?
e. By how much will the y-value on the plane change when x_2 decreases by 2 units and x_1 remains unchanged?

3. Answer true or false to each of the following statements and explain your answers.

a. The y-intercept of a plane has no effect on the rate at which the plane increases or decreases as one of the independent variables changes.
b. If a plane has a negative partial slope for the independent variable x_1, then y-values on the plane increase as x_1 increases provided the other independent variables remain unchanged.
c. A linear equation that involves independent variables x_1 and x_2 and that has both partial slopes equal to zero has a graph that is a plane that does not increase or decrease as x_1 or x_2 change value.

4. What kind of plot is useful for deciding whether it is reasonable to find a regression plane for a set of data points involving several predictor variables?

5. Identify one use of a multiple linear regression equation.

6. Fill in the blanks.

a. In multiple linear regression analysis, the independent variables are called ______.
b. In multiple linear regression analysis, the dependent variable is called ______.

7. Regarding multiple linear regression analysis:

a. What sum of squares is made as small as possible when finding the best fitting plane based on the method of least squares?
b. What do we call the best fitting plane that is obtained by the method of least squares?
c. What do we call the act of using a multiple linear regression equation to make predictions for values of the predictor variables that are outside the region of the observed values of the predictor variables?

8. Explain how the coefficient of multiple determination, R^2, is used as a descriptive measure in multiple linear regression.

9. For each of the following sums of squares in multiple linear regression, identify its name and what it measures.

a. SST
b. SSR
c. SSE

10. Answer true or false to the following statements and explain your answers.

a. A value of R^2 close to 1 indicates a causal relationship between the response variable and the predictor variables.
b. If a response variable is not linearly related to the predictor variables, the value of R^2 will be close to 0.
c. If the value of R^2 is close to 1, then the linear correlation coefficient of the response variable with each predictor variable will be close to 1 or -1.

11. Electricity Costs. The monthly service charge for electricity for a residential customer who is using the local electric utility's Two Period Plan is \$4.10. There is a charge of \$0.148 per kilowatt hour (kWh) of electricity used during "on peak" hours in a month (x_1), and a charge of \$0.046 per kWh of electricity used during "off peak" hours in a month (x_2). Let y denote the cost of electricity during a month.

a. Obtain the equation that expresses y in terms of x_1 and x_2.
b. Find the y-intercept, b_0, and partial slopes, b_1 and b_2, of the linear equation in (a).
c. Find the cost for a customer who uses 200 kWh of electricity during "on peak" hours and 800 kWh of electricity during "off peak" hours during a month.

12. Suppose x_1 and x_2 are predictor variables for a response variable y.

a. The distribution of all possible values of the response variable corresponding to particular values of the two predictor variables is called a ______ distribution of the response variable.
b. State the four assumptions for multiple linear regression inferences.

13. Fill in the blanks.

a. The F-statistic for a test of the utility of a multiple linear regression equation is obtained by dividing ______ by ______.

b. The degrees of freedom for the F-statistic for a test of the utility of a multiple linear regression equation are ______ and ______.

c. The test statistic for a t-test of the utility of a particular predictor variable in a multiple linear regression equation is obtained by dividing ______ by ______.

d. The number of degrees of freedom for the test statistic of the utility of a particular predictor variable in a multiple linear regression equation is ______.

14. Answer true or false to each of the following statements and explain your answers.

a. If the F-test for the utility of a multiple linear regression equation rejects the null hypothesis, then each of the predictor variables in the regression equation is useful in predicting the response variable.

b. The t-test for the utility of a particular predictor variable in a multiple linear regression equation can be affected by the other predictor variables in the equation.

15. Which interval is wider: (a) the 95% confidence interval for the conditional mean of the response variable at a particular set of values of the predictor variables or (b) the 95% prediction interval for the response variable at the same set of values of the predictor variables? Explain your answer.

16. What plots did we use in this module to decide whether it is reasonable to presume that the assumptions for multiple linear regression inferences are met by the predictor variables and response variable? What properties should these plots have?

17. Regarding analysis of residuals, decide in each case which assumption for regression inferences may be violated.

a. The plot of the residuals against the predicted values shows curvature.

b. A plot of the residuals against a predictor variable becomes narrower as the predictor variable values increase.

c. A normal probability plot of the residuals shows curvature.

18. Annual Income. The Census Bureau collects data on income by educational attainment, sex, and age. Results are published in *Current Population Reports.* From a random sample of 75 males between the ages of 25 and 50, all of whom have at least a ninth-grade education, data were collected on age, number of years of school completed, and annual income. Then Minitab was used to perform a multiple regression analysis for annual income (in thousands of dollars) with the variables age and number of years of school completed as predictor variables. The resulting output is shown in Printout A.20.

a. Use the computer output to obtain the regression equation for annual income in terms of age and number of years of school completed.

b. Apply the regression equation to predict the annual income of a male who is 32 years old and has completed exactly 4 years of college (i.e., 16 years of school).

c. Find and interpret the coefficient of determination, R^2.

d. Find and interpret the standard error of the estimate.

e. At the 5% significance level, do the data provide sufficient evidence to conclude that, taken together, age and number of years of school completed are useful for predicting annual income for males (the type of males under consideration)?

f. At the 5% significance level, do the data provide sufficient evidence to conclude that age is useful as a predictor of annual income for males? Be precise in your conclusion.

g. Repeat part (f) for the predictor variable, number of years of school completed.

19. Annual Income. Refer to Problem 18 and the computer output in Printout A.20. In the last five lines of the output you will find information on annual income of males who are 32 years old and have completed exactly 4 years of college (i.e., 16 years of school). Presuming that the assumptions for regression inferences are met, use Printout A.20 to solve the following problems.

a. Find a point estimate for the mean annual income of all males who are 32 years old and have completed exactly 4 years of college (i.e., 16 years of school).

b. Obtain a 95% confidence interval for the mean annual income of all males who are 32 years old and have completed exactly 4 years of college.

c. Determine the predicted annual income of a randomly selected male who is 32 years old and has completed exactly 4 years of college.

d. Find a 95% prediction interval for the annual income of a randomly selected male who is 32 years old and has completed exactly 4 years of college.

20. Annual Income. Refer to Problem 18. Printouts A.21(a), (b), and (c) on page A-80 display, respectively, plots of residuals against education, residuals against age, and residuals against predicted income; Printout A.21(d) shows a normal probability plot of the residuals. Do these graphs suggest any violations of the assumptions for multiple linear regression inferences for the variables under consideration? Explain your answer.

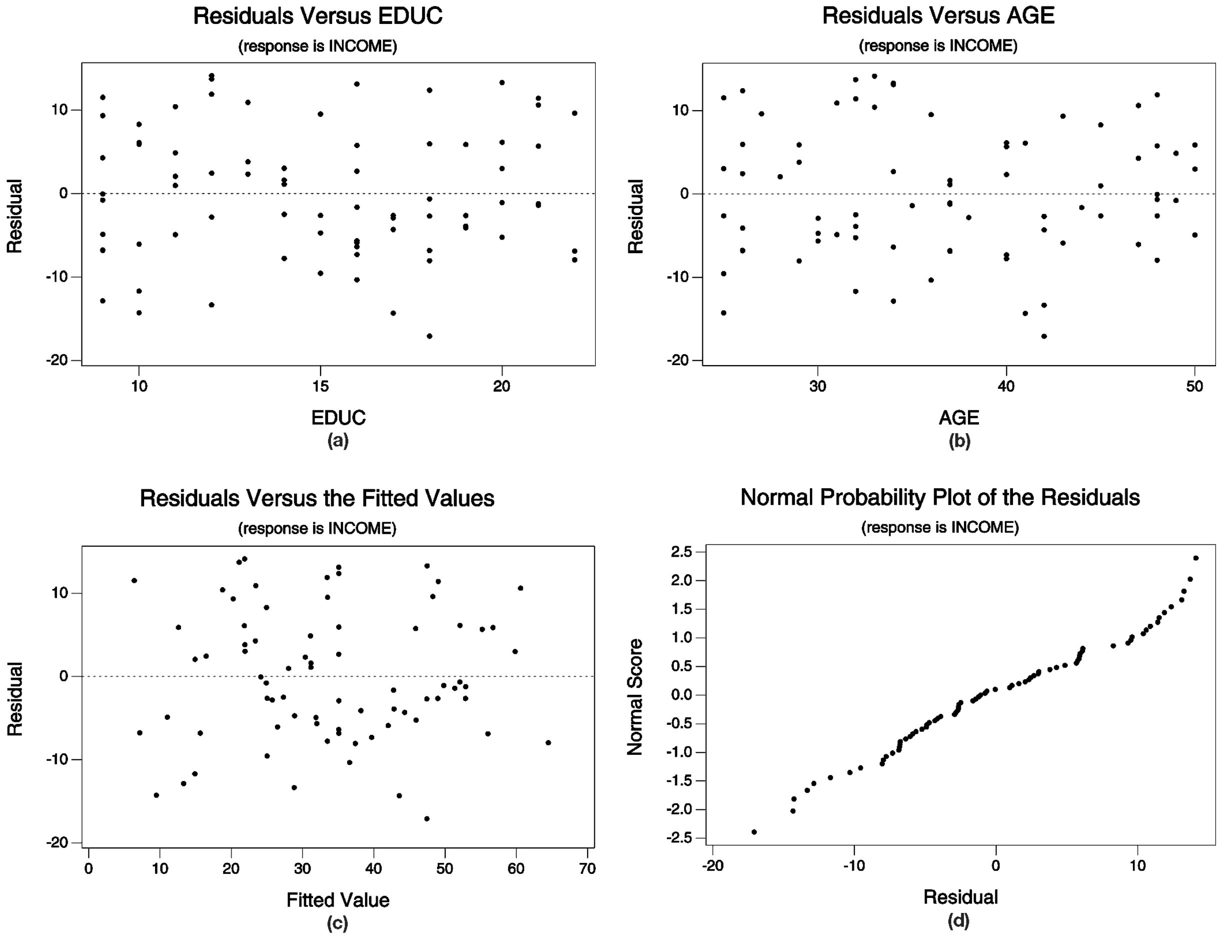
Using Technology

21. Annual Income. Refer to Problem 18. Using your statistical technology, explain how you would obtain output similar to that shown in Printout A.20, and in Printouts A.21(a)–(d) on page A-80.

Printout A.20
Minitab output for Problems 18 and 19

```
The regression equation is
INCOME = - 40.9 + 0.772 AGE + 3.11 EDUC

Predictor        Coef     SE Coef          T        P
Constant      -40.855       5.511      -7.41    0.000
AGE            0.7719      0.1165       6.62    0.000
EDUC           3.1052      0.2250      13.80    0.000

S = 7.886       R-Sq = 76.6%     R-Sq(adj) = 75.9%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         2     14657.1      7328.6    117.84    0.000
Residual Error    72      4477.6        62.2
Total             74     19134.7

Source       DF      Seq SS
AGE           1      2808.6
EDUC          1     11848.5

Unusual Observations
Obs        AGE     INCOME         Fit      SE Fit    Residual    St Resid
 20       42.0     30.374      47.457       1.286     -17.083       -2.20R

R denotes an observation with a large standardized residual

Predicted Values for New Observations

New Obs     Fit     SE Fit         95.0% CI              95.0% PI
1        33.528      1.096    (  31.342,  35.714)  (  17.656,  49.400)

Values of Predictors for New Observations

New Obs        AGE      EDUC
1             32.0      16.0
```

Printout A.21
Residual plots for Problem 20



Internet Project

Homicides in Detroit

Several studies have shown that homicide rates are linked to economic conditions. Poverty levels or the percent of unemployed people are sometimes used to predict changes in homicide rates. Also, drug-use patterns (which also may be related to economic factors), or the prevalence of handguns has been directly linked to the murder rate. Others have claimed that longer prison sentences, changes in the youth population, or the amount of money available to police departments are responsible for nationwide changes in violent crimes.

Obviously, homicide is a complex problem and these studies and articles may all be partially correct. However, there seems to be no single factor that can fully explain or predict the fluctuations in the homicide rate. In this project, you will examine data containing several of the variables that have been linked to changes in the homicide rate. Because no single factor can explain the complexity in the data, you will use multiple regression to attempt to find an answer.

URL for access to Internet Projects Page: www.aw.com/weiss

Focusing on Data Analysis

HS GPA, GPA, and SAT Scores

Recall from Chapter 1 that the Focus database contains information on 500 randomly selected Arizona State University sophomores. Among the variables considered are high-school GPA, SAT math score, SAT verbal score, and cumulative GPA. For these database exercises, you should eliminate all cases (students) in which one or more of the four variables, high-school GPA, SAT math score, SAT verbal score, and cumulative GPA, equal 0. Use the technology of your choice to solve the following problems regarding the relationship between the response variable, cumulative GPA, and the three predictor variables, high-school GPA, SAT math score, and SAT verbal score.

a. Obtain a scatterplot matrix of the data for these four variables. What do these scatterplots tell you about the relationships among the variables?

b. Does a multiple linear regression equation relating cumulative GPA to high-school GPA, SAT math score, and SAT verbal score seem appropriate for these data? Explain your answer.

c. Determine the multiple linear regression equation relating the response variable of cumulative GPA to the three predictor variables.

d. Interpret the estimated regression coefficients.

e. Determine the proportion of variation in the observed cumulative GPAs that can be explained by the multiple linear regression.

f. Construct residual plots and assess the appropriateness of the multiple linear regression equation.

g. Construct residual plots and a normal probability plot to assess the appropriateness of the assumptions of constant conditional standard deviation and normality for multiple linear regression inferences.

h. Identify any potential outliers.

Presuming now that the assumptions for multiple linear regression inferences hold for the four variables cumulative GPA, high-school GPA, SAT math score, and SAT verbal score, solve the following problems.

i. Obtain the analysis of variance table for the regression of cumulative GPA on the three predictor variables.

j. At the 5% level of significance, do the data provide sufficient evidence to conclude that the three predictor variables taken together are useful for predicting the cumulative GPA of sophomores at Arizona State University?

k. At the 5% level of significance, do the data provide sufficient evidence to conclude that high school GPA is useful for predicting cumulative GPA when SAT math score and SAT verbal score are also in the regression equation?

l. Repeat part (k) to determine the usefulness of SAT math score in predicting cumulative GPA when the other two predictor variables are in the regression equation.

m. Repeat part (k) to determine the usefulness of SAT verbal score in predicting cumulative GPA when the other two predictor variables are in the regression equation.

n. Should all three predictor variables remain in the multiple linear regression equation?

o. Do you think the three predictor variables do a good job of predicting cumulative GPA?

p. Determine a point estimate for the mean cumulative GPA of all sophomores at Arizona State University who have a high-school GPA of 3.0, an SAT math score of 500, and an SAT verbal score of 450.

q. Find a 95% confidence interval for the mean cumulative GPA of all sophomores at Arizona State University who have a high-school GPA of 3.0, an SAT math score of 500, and an SAT verbal score of 450.

r. Determine the predicted cumulative GPA of a sophomore at Arizona State University who has a high-school GPA of 3.0, an SAT math score of 500, and an SAT verbal score of 450.

s. Obtain a 95% prediction interval for the cumulative GPA of a sophomore at Arizona State University who has a high-school GPA of 3.0, an SAT math score of 500, and an SAT verbal score of 450.

t. Suggest other possible variables that could be considered as predictors of cumulative GPA.

case study discussion

Automobile Insurance Rates

At the beginning of this module on page A-3, we discussed studying the effect various factors might have on the average automobile insurance rate in a state. Data were obtained from the 117th edition (1997) of the *Statistical Abstract of the United States* for each of the 50 states on the response variable, average automobile insurance rate, and five predictor variables: population density, automobile theft rate, automobile death rate per 100 million miles driven, average drive time to work, and average cost of a day's stay in a hospital. The data are shown in Table A.3. Using the technology of your choice, answer the following questions.

a. Draw a scatterplot matrix of the data for the six variables. What do these scatterplots tell you about the relationship among the variables?

b. Does a multiple linear regression equation relating insurance rate to the five predictor variables seem appropriate for these data? Explain your answer.

c. Find the multiple linear regression equation relating the response variable of insurance rate to the five predictor variables.

d. Interpret the sample regression coefficients.

e. Determine the proportion of variation in the observed insurance rates that can be accounted for by the multiple linear regression equation in the five predictor variables.

f. Should all of the predictors remain in the regression equation?

g. Construct residual plots and assess the appropriateness of the multiple linear regression equation.

h. Construct residual plots to assess the assumptions of constant conditional standard deviation and normality.

i. Identify potential outliers and influential observations.

j. Do you think these predictor variables do a good job of predicting the response?

k. Suggest other possible variables that should be considered as predictors in this regression model.

Table A.3
Automobile insurance rates and other related data for the 50 states

State	Ave. ins. rate	Pop. density	Auto theft rate	Deaths/100M miles	Ave. drive time	Hospital cost/day
AK	730	1.1	522	2.0	16.7	1341
AL	549	84.2	347	2.2	21.2	819
AR	500	48.2	325	2.5	19.0	704
AZ	727	39.0	1158	2.6	21.6	1191
CA	831	204.4	888	1.5	24.6	1315
CO	722	36.9	388	1.9	20.7	1069
CT	881	675.8	540	1.1	21.1	1264
DE	784	370.8	414	1.7	20.0	1058
FL	739	266.7	786	2.3	21.8	1004
GA	596	127.0	608	1.8	22.7	836
HI	963	184.3	691	1.6	23.8	956
IA	429	51.0	223	2.0	16.2	702
ID	447	14.4	242	2.2	17.3	719
IL	612	213.1	523	1.7	25.1	1050
IN	542	162.8	466	1.5	20.4	963
KS	474	31.4	324	1.7	17.2	732
KY	555	97.7	259	2.1	20.7	795
LA	787	99.9	598	2.3	22.3	902
MA	898	777.3	605	0.9	22.7	1157
MD	732	518.8	718	1.5	27.0	1064
ME	472	40.3	135	1.5	19.0	916
MI	645	168.9	646	1.8	21.2	994
MN	630	58.5	342	1.4	19.1	736
MO	572	77.8	473	1.9	21.6	967
MS	579	57.9	361	3.0	20.6	584
MT	468	6.0	308	2.3	14.8	493
NC	501	150.3	311	2.0	19.8	832
ND	381	9.3	179	1.1	13.0	521
NE	452	21.5	351	1.6	15.8	661
NH	609	129.6	145	1.1	21.9	915
NJ	1013	1076.7	632	1.3	25.3	962
NM	639	14.1	513	2.3	19.1	1073
NV	759	14.6	745	2.4	19.8	1072
NY	906	385.1	566	1.4	28.6	909
OH	532	272.8	415	1.4	20.7	1061
OK	526	48.1	496	1.8	19.3	861
OR	565	33.4	702	1.9	19.6	1141
PA	667	269.0	413	1.6	21.6	963
RI	870	947.6	441	1.0	19.2	1092
SC	582	122.8	385	2.3	20.5	923
SD	429	9.7	121	2.0	13.8	476
TN	519	129.1	649	2.2	21.5	871
TX	711	73.0	560	1.7	22.2	1063
UT	547	24.3	389	1.7	18.9	1213
VA	553	168.6	293	1.3	24.0	901
VT	512	63.6	136	1.7	18.0	714
WA	650	83.1	554	1.4	22.0	1318
WI	506	95.0	364	1.4	18.3	794
WV	646	75.8	166	2.2	21.0	763
WY	433	5.0	168	2.5	15.4	545

Answers to Selected Exercises

Exercises A.1

A.1

a. $y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k$
b. $b_0, b_1, \ldots, b_k$ are constants; $y, x_1, \ldots, x_k$ are variables
c. $x_1, \ldots, x_k$ represent the independent variables; y represents the dependent variable

A.3

a. b_0 is the y-intercept of the plane given by the linear equation. b_0 is the value of y at which the plane intersects the y-axis.
b. b_1 and b_2 are the partial slopes of the plane. b_1 (b_2) tells us how much y will change when x_1 (x_2) increases by 1 unit, while x_2 (x_1) is held fixed.

A.5

a. $y = 20 + 39.95x_1 + 0.20x_2$
b. $b_0 = 20, b_1 = 39.95, b_2 = 0.20$
c. \$39.95
d. \$319.75

A.7

a. $y = 75 + 40x_1 + 16x_2 + 5x_3$
b. $b_0 = 75, b_1 = 40, b_2 = 16, b_3 = 5$
c. \$16
d. \$1535

A.9

a. y-intercept is $b_0 = 20$; partial slopes are $b_1 = 39.95$ and $b_2 = 0.20$.
b. The y-intercept, $b_0 = 20$, gives the value of y at which the plane, $y = 20 + 39.95x_1 + 0.20x_2$, intersects the y-axis. The partial slope, $b_1 = 39.95$, indicates that the value of y increases by 39.95 units for every increase in x_1 of 1 unit, while x_2 is held fixed. The partial slope, $b_2 = 0.20$, indicates that the value of y increases by 0.20 for every increase in x_2 of 1 unit, while x_1 is held fixed.
c. The y-intercept, $b_0 = 20$, represents the fact that the cost of renting a car is \$20 for 0 days and 0 miles driven; it is the one time surcharge because the car is rented at the airport. The partial slope, $b_1 = 39.95$, represents the fact that the cost per day is \$39.95; it is the amount the total cost increases for each additional day of rental. The partial slope, $b_2 = 0.20$, represents the fact that the cost per mile is \$0.20; it is the amount the total cost increases for each additional mile driven.

A.11

a. y-intercept is $b_0 = 75$; partial slopes are $b_1 = 40$, $b_2 = 16$, and $b_3 = 5$.
b. The y-intercept, $b_0 = 75$, gives the value of y at which the plane, $y = 75 + 40x_1 + 16x_2 + 5x_3$, intersects the y-axis. The partial slope, $b_1 = 40$, indicates that the value of y increases by 40 units for every increase in x_1 of 1 unit, while x_2 and x_3 are held fixed. The partial slope, $b_2 = 16$, indicates that the value of y increases by 16 units for every increase in x_2 of 1 unit, while x_1 and x_3 are held fixed. The partial slope, $b_3 = 5$, indicates that the value of y increases by 5 units for every increase in x_3 of 1 unit, while x_1 and x_2 are held fixed.
c. The y-intercept, $b_0 = 75$, represents the fact that the cost of renting the banquet room is \$75 for 0 hours, with 0 buffet dinners, and 0 drinks being ordered; it is the one time clean up charge. The partial slope, $b_1 = 40$, represents the fact that the cost per hour is \$40; it is the amount the total cost increases for each additional hour of usage. The partial slope, $b_2 = 16$, represents the fact that the cost per dinner is \$16; it is the amount the total cost increases for each additional dinner ordered. The partial slope, $b_3 = 5$, represents the fact that the cost per drink is \$5; it is the amount the total cost increases for each additional drink ordered.

A.13

a. y-intercept, $b_0 = 3$; partial slopes, $b_1 = 4$ and $b_2 = 7$
b. slopes upward

A.15

a. y-intercept, $b_0 = 6$; partial slopes, $b_1 = -7$ and $b_2 = 10$
b. slopes downward

A.17

a. y-intercept, $b_0 = -2$; partial slopes, $b_1 = 3$ and $b_2 = 0$
b. slopes upward

A.19

a. y-intercept, $b_0 = 7$; partial slopes, $b_1 = 0$ and $b_2 = 3$
b. horizontal

A.21

a. y-intercept, $b_0 = 9$; partial slopes, $b_1 = 0$ and $b_2 = 0$
b. horizontal

A.23

a. slopes upward
b. $y = 5 + 2x_1 + 7x_2$

A.25

a. slopes downward
b. $y = -2 + 3x_1 - 7x_2$

A.27

a. slopes upward
b. $y = -0.05x_1 + 1.5x_2$

A.29

a. slopes upward
b. $y = 3x_2$

A.31 More than one predictor variable may be useful in predicting the response variable and provide greater precision in prediction.

A.33

a. Engine size, weight, length, width, and luggage capacity of an automobile
b. $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$

Exercises A.2

A.37

a. the least squares criterion
b. Find the plane for which the sum of squared deviations between the observed and predicted values of the response variable is the smallest.

A.39

a. response variable
b. predictor variables

A.41

a. Plane A: $y = 2 + 3x_1 + x_2$

x_1	x_2	y	$\hat{y}$	e	e^2
1	1	8	6	2	4
1	2	10	7	3	9
2	1	11	9	2	4
2	2	16	10	6	36
					53

Plane B: $y = 3 + 4x_1 + 2x_2$

x_1	x_2	y	$\hat{y}$	e	e^2
1	1	8	9	−1	1
1	2	10	11	−1	1
2	1	11	13	−2	4
2	2	16	15	1	1
					7

b. Plane B

A.43

a. Sales tend to increase in a roughly linear manner as television, magazine, and radio advertising expenditures increase.
b. $\hat{y} = 266 + 6.73x_1 + 3.26x_2 + 4.51x_3$
c. \$367.4 million

A.44

a. Price tends to decrease in a roughly linear manner as age and miles driven increase.
b. $\hat{y} = 367 - 37.4x_1 + 1.64x_2$
c. \$26,372

Exercises A.3

A.51

a. total sum of squares; *SST*
b. regression sum of squares; *SSR*
c. error sum of squares; *SSE*

A.53 conditional distribution; conditional mean; conditional standard deviation

A.55

a. population regression plane
b. σ
c. normal; $\beta_0 + 5\beta_1 + 8\beta_2 + 13\beta_3$; σ

A.57 *SSE*

A.59 $n - 1 = 3 + (n - 4)$

A.61

a. False. Rejection of the null hypothesis only implies that at least one of the population regression coefficients is not zero.
b. False. If the null hypothesis is not rejected, we can only conclude that the multiple linear regression equation is not useful in predicting the response variable. The response variable might be related to the predictor variables by some other regression equation.
c. True. See the displayed equation on page A-37.

A.63

a.

Source	df	*SS*	*MS*	*F*
Regression	7	987	141.00	4.63
Error	60	1826	30.43	
Total	67	2813		

b. $R^2 = 987/2813 = 0.351$
c. $s_e = 5.52$
d. H_0: $\beta_1 = \beta_2 = \ldots = \beta_7 = 0$, H_a: At least one of the β_is is not zero; $\alpha = 0.01$; critical value $= 2.95$. Since $F = 4.63 >$

2.95, reject H_0. We conclude that, taken together, the seven predictors are useful for predicting the response variable.

e. Although we rejected the null hypothesis that all the regression coefficients are zero, the value of R^2 is only 0.351. There is a great deal of variation in the response variable that is not accounted for by the seven predictors, and we might not be able to make predictions as accurately as we would like.

A.64

a. It would mean that there are constants, β_0, β_1, β_2, β_3, and σ, such that for each amount spent for television advertising, x_1, for magazine advertising, x_2, and for radio advertising, x_3, the total sales are normally distributed with mean $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ and standard deviation σ. Assumption 4 would mean that the observations are obtained independently.

b. $R^2 = 0.911$. 91.1% of the total variation in the observed values of total sales is accounted for by the linear regression equation in television, magazine, and radio advertising expenditures.

c. $s_e = 4.418$. Our best estimate of the (common) conditional standard deviation, σ, of the total sales for any particular amounts expended for television, magazine, and radio advertising is $4.418 million.

d. H_0: $\beta_1 = \beta_2 = \beta_3 = 0$, H_a: At least one of the β_is is not zero; $\alpha = 0.05$; critical value $= 4.76$. Since $F = 20.40 > 4.76$, reject H_0; at the 5% significance level, the data provide sufficient evidence to conclude that at least one of the population regression coefficients (β_1, β_2, β_3) is not zero. Taken together, the predictor variables (television, magazine, and radio advertising expenditures) are useful in predicting total sales. For the P-value approach, note that $F = 20.40$ has $P = 0.002 < 0.05$.

A.65

a. It would mean that there are constants β_0, β_1, β_2, and σ, such that for each age, x_1, and number of miles driven, x_2, the prices of all Corvettes of that age that have been driven that number of miles are normally distributed with mean $\beta_0 + \beta_1 x_1 + \beta_2 x_2$ and standard deviation σ. Assumption 4 would mean that the observations are obtained independently.

b. $R^2 = 0.960$. 96.0% of the total variation in the observed values of price is accounted for by the linear regression equation in age and number of miles driven.

c. The multiple linear regression equation that uses both age and number of miles driven explains 96.0% of the total variation in the observed prices. This is an increase of 2.3% over the regression equation using age as the only predictor variable. The regression equation with both predictor variables explains more of the total variation in price than the regression equation with only age as a predictor. However, the increase is rather small and may not be statistically meaningful or practically useful.

d. $s_e = 12.11$. Our best estimate of the (common) conditional standard deviation, σ, of Corvette prices for any particular age and number of miles driven is $1211.

e. H_0: $\beta_1 = \beta_2 = 0$, H_a: At least one of the β_is is not zero; $\alpha = 0.05$; critical value $= 4.74$. Since $F = 84.07 > 4.74$, reject H_0; at the 5% significance level, the data provide sufficient evidence to conclude that at least one of the population regression coefficients (β_1, β_2) is not zero. Taken together, the predictor variables (age, miles driven) are useful in predicting the price of a Corvette. For the P-value approach, note that $F = 84.07$ has $P = 0.000 < 0.05$.

A.69

a. $SSE = SST$

b. $SSR = 0$

c. The sample multiple linear regression equation is not useful at all for making predictions.

Exercises A.4

A.75 If the partial slopes are all zero, a change in the value of any predictor will result in no change at all in the value of the response variable. Thus the predictor variables are useless in predicting the response variable if the partial slopes are all zero.

A.77 $t = b_1/s_{b_1}$. This test statistic has a t-distribution with $\text{df} = n - (k + 1)$.

A.79

a. H_0: $\beta_1 = 0$, H_a: $\beta_1 \neq 0$; $\alpha = 0.05$; critical values $= \pm 2.447$, $t = 5.01$; reject H_0; at the 5% significance level, the data provide sufficient evidence to conclude that the regression coefficient, β_1, of the population regression equation is not zero. Hence television advertising expenditure is a useful predictor of total sales in the regression equation that also contains the predictor variables magazine and radio advertising expenditures. For the P-value approach, note that $P = 0.002 < 0.05$.

b. H_0: $\beta_3 = 0$, H_a: $\beta_3 \neq 0$; critical values $= \pm 2.447$, $t = 1.22$; do not reject H_0; at the 5% significance level, the data do not provide sufficient evidence to conclude that the regression coefficient, β_3, of the population regression equation is not zero. Hence, we have insufficient evidence to conclude that radio advertising expenditure is a useful predictor of total sales in the regression equation that also contains the predictor variables television and magazine advertising expenditures. For the P-value approach, note that $P = 0.269 > 0.05$.

c. 3.438 to 10.016

d. We can be 95% confident that the change in mean total sales per $1 million increase in television advertising expenditures is somewhere between \$3.438 million and \$10.016 million.

e. −4.554 to 13.568

A.80

a. H_0: $\beta_1 = 0$, H_a: $\beta_1 \neq 0$; $\alpha = 0.05$; critical values = ±2.365, $t = -7.22$; reject H_0; at the 5% significance level, the data provide sufficient evidence to conclude that the regression coefficient, β_1, of the population regression equation is not zero. Hence age is a useful predictor of price of a Corvette in the regression equation that also contains the predictor variable miles driven. For the P-value approach, note that $P = 0.000 < 0.05$.

b. H_0: $\beta_2 = 0$, H_a: $\beta_2 \neq 0$; $\alpha = 0.05$; critical values = ±2.365, $t = 2.02$; do not reject H_0; at the 5% significance level, the data do not provide sufficient evidence to conclude that the regression coefficient, β_2, of the population regression equation is not zero. Hence we have insufficient evidence to conclude that miles driven is a useful predictor of the price of a Corvette in the regression equation that also contains the predictor variable age. For the P-value approach, note that $P = 0.083 > 0.05$. If the level of significance is changed to 10%, we would reject $H_0 : \beta_2 = 0$, and conclude that miles driven is a useful predictor.

c. −49.61 to −25.14

d. We can be 95% confident that the change in mean price of Corvettes per 1 year increase in age is somewhere between −\$4961 and −\$2514.

e. −0.28 to 3.56

Exercises A.5

A.87 Confidence interval for the conditional mean, prediction interval for the response. The four assumptions required are presented in Key Fact A.3 on page A-29.

A.89 87.62

A.91

a. \$367.58 million

b. \$361.24 million to \$373.92 million

c. \$367.58 million

d. \$355.05 million to \$380.12 million

e. Prediction seems appropriate based on the two-dimensional plots of expenditures. However it would be better to have a three-dimensional plot of the data points for the three expenditures and determine whether the specified set of three expenditures falls in the region of the observed data points.

f. No. This set of three expenditures appears to be outside the region of the observed data points of the three expenditures.

A.92

a. \$26,372

b. \$25,365 to \$27,380

c. \$26,372

d. \$23,335 to \$29,410

e. Prediction seems appropriate since the age and number of miles driven are within the region of values observed in the data set.

f. No. A 10-year-old Corvette that has been driven 55,000 miles is outside the region of age and miles driven for the Corvettes in our sample.

Exercises A.6

A.99 observed value of the response variable; predicted value of the response variable

A.101

a. outlier

b. influential observation

A.103

a. Each of the residual plots in Printouts A.13(a)–(d) shows a random scatter of points in a horizontal band centered and roughly symmetric about the horizontal axis. There do not appear to be any violations of the assumptions of linearity of the regression equation or constancy of the conditional standard deviation. The normal probability plot in Printout A.13(e) does not appear to be linear. Printout A.12 does not indicate any outliers or influential observations.

b. The analysis in part (a) indicates that the normality assumption might be violated because the normal probability plot is not roughly linear. However since there are only 10 data points, it is difficult to draw any firm conclusions.

A.104

a. Each of the residual plots in Printouts A.15(a)–(c) shows a random scatter of points in a horizontal band centered and roughly symmetric about the horizontal axis. There do not appear to be any violations of the assumptions of linearity of the regression equation or constancy of the conditional standard deviation. The normal probability plot in Printout A.15(d) is roughly linear except for one residual at the bottom left of the plot. Printout A.14 does not indicate any outliers or influential observations.

b. The analysis in part (a) may indicate a problem with the normality assumption. The most negative residual does not follow the roughly linear pattern of the other residuals in the normal probability plot. However with so few observations it is difficult to conclude that there is a violation of the normality assumption.

Review Test for Module A

1.

a. x_1, x_2, x_3
b. y
c. b_0
d. b_1, b_2, b_3

2.

a. 5
b. $x_1 = 0, x_2 = 0$
c. 4 and −3
d. 4
e. 6

3.

a. True.
b. False. y-values would decrease since the partial slope for x_1 is negative.
c. True.

4. scatterplot matrix

5. to predict the value of the response variable for a given set of values of the predictor variables

6.

a. predictor variables
b. the response variable

7.

a. $\Sigma(y - (b_0 + b_1x_1 + \cdots + b_kx_k))^2$
b. the sample regression plane
c. extrapolation

8. R^2 is the proportion of the total variation in the observed values of the response variable, y, explained by the multiple linear regression equation in the predictor variables.

9.

a. Total sum of squares. *SST* measures the total variation in the observed values of the response variable.
b. Regression sum of squares. *SSR* measures the variation in the observed values of the response variable that is accounted for by the linear regression.
c. Error sum of squares. *SSE* measures the variation in the observed values of the response variable that is not accounted for by the linear regression.

10.

a. False. A value of R^2 close to 1 only indicates a strong statistical relationship, not a causal one.
b. False. Although the linear model may not be the appropriate one, it still may explain a good portion of the variation in the observed values of the response variable and thereby yield a value of R^2 that is not close to 0.
c. False. A value of R^2 close to 1 does not imply that each predictor variable is highly correlated with the response variable.

11.

a. $y = 4.1 + 0.148x_1 + 0.046x_2$
b. y-intercept is $b_0 = 4.1$; partial slopes are $b_1 = 0.148$ and $b_2 = 0.046$
c. \$70.50

12.

a. conditional
b. See Key Fact A.3 on page A-30.

13.

a. *MSR*, *MSE*
b. $k, n - (k + 1)$
c. b_i, s_{b_i}
d. $n - (k + 1)$

14.

a. False. If the null hypothesis is rejected, we only know that at least one of the predictor variables is useful, but we do not know whether all of them are useful.
b. True.

15. (b) the 95% prediction interval

16. Residual plots. See Key Fact A.11 on page A-63.

17.

a. population regression plane
b. equal standard deviations
c. normal populations

18.

a. $\hat{y} = -40.9 + 0.772x_1 + 3.11x_2$
b. \$33,564
c. $R^2 = 0.766$. 76.6% of the total variation in the observed annual incomes is explained by the multiple linear regression equation in age and number of years of school completed.
d. $s_e = 7.886$. Our best estimate for the (common) conditional standard deviation, σ, of annual incomes for males for any specified age between 25 and 50 and number of years of education of at least 9 years is \$7886.
e. H_0: $\beta_1 = \beta_2 = 0$, H_a: At least one of the β_is is not zero; $\alpha = 0.05$; critical value $= 3.12$. Since $F = 117.84 > 3.12$, reject H_0; at the 5% significance level, the data provide sufficient evidence to conclude that at least one of the population regression coefficients (β_1, β_2) is not zero. Taken together, the predictor variables (age, number of years of school completed) are useful in predicting annual income of males between the ages of 25 and 50 with at least a ninth-grade education. For the P-value approach, note that $F = 117.84$ has $P = 0.000 < 0.05$.

f. H_0: $\beta_1 = 0$, H_a: $\beta_1 \neq 0$; $\alpha = 0.05$; critical values $= \pm 1.994$, $t = 6.62$; reject H_0; at the 5% significance level, the data provide sufficient evidence to conclude that the regression coefficient, β_1, of the population regression equation is not zero. Hence age is a useful predictor of annual income of males in the regression equation that also contains the predictor variable number of years of school completed. For the P-value approach, note that $P = 0.000 < 0.05$.

g. H_0: $\beta_2 = 0$, H_a: $\beta_2 \neq 0$; $\alpha = 0.05$; critical values $= \pm 1.994$, $t = 13.80$; reject H_0; at the 5% significance level, the data provide sufficient evidence to conclude that the regression coefficient, β_2, of the population regression equation is not zero. Hence number of years of school completed is a useful predictor of annual income of males in the regression equation that also contains the predictor variable age. For the P-value approach, note that $P = 0.000 < 0.05$.

19.

a. \$33,528
b. \$31,342 to \$35,714
c. \$33,528
d. \$17,656 to \$49,400

20. The plots of the residuals in Printouts A.17(a)–(c) show a random scatter of points in a horizontal band centered and symmetric about the horizontal axis. There do not appear to be any violations of the assumptions of linearity of the regression equation or constancy of the conditional standard deviation. There are no apparent outliers in the plots. The normal probability plot in Printout A.17(d) shows a roughly linear plot of the residuals indicating that the normality assumption is appropriate.