CHAPTER 8

Multiple Regression

CONTENTS

■ Example 8.1: What Wins Baseball Games?

The game of baseball generates an unbelievable amount of descriptive statistics. Although most of us give these statistics only casual scrutiny, baseball managers may find them quite useful tools for analyzing team performance and consequently implementing policies to improve their team's standing.

Table 8.1 shows some summary statistics about the 10 National League baseball teams for the 1965 through 1968 seasons (Reichler, 1985). The variables collected for this study are

YEAR: the season: 1965–1968,
WIN: the team's winning percentage,

Statistical Methods, DOI: 10.1016/B978-0-12-374970-3.00008-1

Table 8.1 Winning Baseball Games

OBS	YEAR	WIN	RUNS	BA	DP	WALK	SO
1	1965	0.599	608	0.245	135	425	1079
2	1965	0.586	682	0.252	124	408	1060
3	1965	0.556	675	0.265	189	469	882
4	1965	0.549	825	0.273	142	587	1113
5	1965	0.531	708	0.256	145	541	996
6	1965	0.528	654	0.250	153	466	1071
7	1965	0.497	707	0.254	152	467	916
8	1965	0.444	635	0.238	166	481	855
9	1965	0.401	569	0.237	130	388	931
10	1965	0.309	495	0.221	153	498	776
11	1966	0.586	606	0.256	128	356	1064
12	1966	0.578	675	0.248	131	359	973
13	1966	0.568	759	0.279	215	463	898
14	1966	0.537	696	0.258	147	412	928
15	1966	0.525	782	0.263	139	485	884
16	1966	0.512	571	0.251	166	448	892
17	1966	0.475	692	0.260	133	490	1043
18	1966	0.444	612	0.255	126	391	929
19	1966	0.410	587	0.239	171	521	773
20	1966	0.364	644	0.254	132	479	908
21	1967	0.627	695	0.263	127	431	956
22	1967	0.562	652	0.245	149	453	990
23	1967	0.540	702	0.251	143	463	888
24	1967	0.537	604	0.248	124	498	1065
25	1967	0.506	612	0.242	174	403	967
26	1967	0.500	679	0.277	186	561	820
27	1967	0.475	631	0.240	148	449	862
28	1967	0.451	519	0.236	144	393	967
29	1967	0.426	626	0.249	120	485	1060
30	1967	0.377	498	0.238	147	536	893
31	1968	0.599	583	0.249	135	375	971
32	1968	0.543	599	0.239	125	344	942
33	1968	0.519	612	0.242	149	392	894
34	1968	0.512	690	0.273	144	573	963
35	1968	0.500	514	0.252	139	362	871
36	1968	0.494	583	0.252	162	485	897
37	1968	0.469	470	0.230	144	414	994
38	1968	0.469	543	0.233	163	421	935
39	1968	0.451	473	0.228	142	430	1014
40	1968	0.444	510	0.231	129	479	1021

RUNS: the number of runs scored by the team,
BA: the team's overall batting average,
DP: the total number of double plays,
WALK: the number of walks given to the other team, and
SO: the number of strikeouts by the team's pitchers.

Obviously the study of the relationships among several variables is much more complicated than that between two variables discussed in Chapter 7. However, it is still useful to examine graphically the relationships among the pairs of variables in this example. Figure 8.1 is a "table" of scatterplots among all pairs of variables in Example 8.1 produced by SAS/INSIGHT. The entries in the diagonal elements (top left to bottom right) identify the variable in the scatterplots on the corresponding rows and columns and the numbers in the corners show the minimum and maximum values of those variables. For example, the first scatterplot in the first row is that between WIN on the vertical axis and RUNS on the horizontal

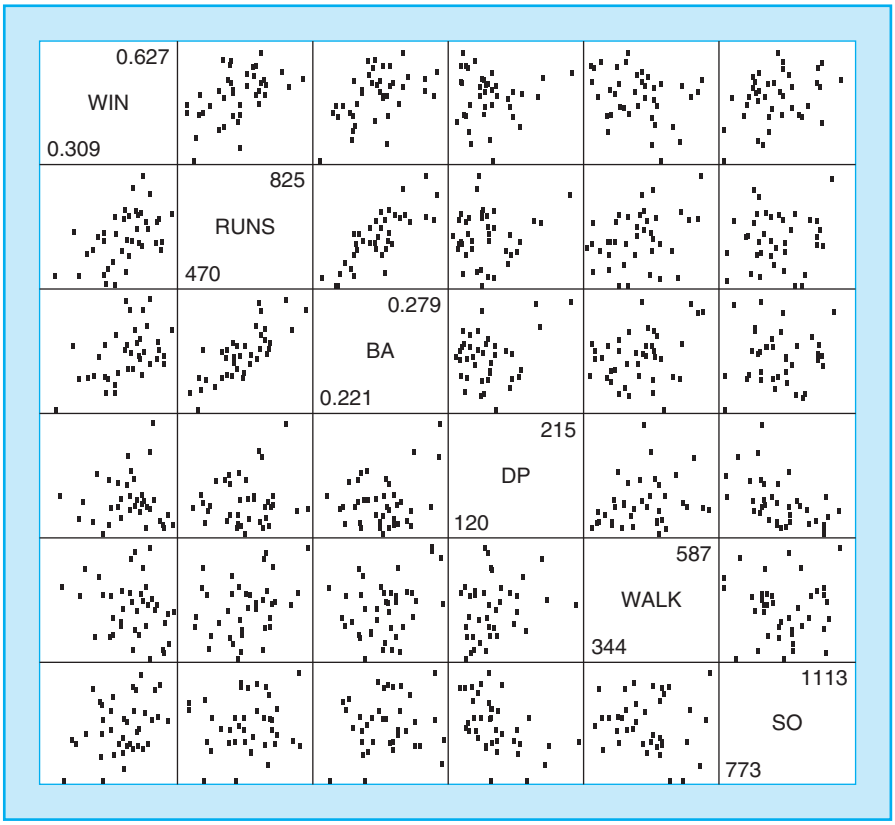

FIGURE 8.1
Scatterplots of Variables in Example 8.1.

axis, and the values of the variable WIN range from 0.309 to 0.627 and RUNS ranges from 470 to 825. Note that each scatterplot is reproduced twice with the axes interchanged.

In this example the focus is on determining the effects of the independent variables (RUNS, BA, DP, WALK, SO) on the winning percentages (WIN). This means that we are interested in the relationships depicted in the first row (or column) of scatterplots. These appear to indicate moderately strong positive relationships of WIN to RUNS, BA, and SO, which appear reasonable. However, looking at the other scatterplots, we see a very strong positive relationship between RUNS and BA. This raises the question whether either or both are responsible for increased winning percentages, since these two variables are closely related. There is also a relatively strong negative relationship between DP and SO. Could this relationship possibly change the effect of either on the winning percentages?

We will see that multiple regression analysis is designed to help answer these questions. However, because the interplay of so many variables can be very complex, the answers are not always as clear as we would like them to be. The solution to this example is provided in Section 8.10. ■

Notes on Exercises

Computations for all exercises in this chapter require statistical software. In most cases, the same program used for the exercises in Chapter 7 will suffice, the only difference being that more than one independent variable must be specified. After Section 8.2, Exercise 1 can be worked, using software options for the various outputs requested in that exercise. Referring to those outputs will help in understanding the material in Sections 8.1 through 8.4. Section 8.5 is a short review of the interpretation of computer outputs, after which all other assigned exercises except 8.7, 8.9, and 8.10 can be worked. These exercises can be worked after covering Section 8.6.

8.1 THE MULTIPLE REGRESSION MODEL

In Chapter 7 we observed that the simple linear regression model

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

which relates observed values of the dependent or response variable y to values of a single independent variable x, had limited practical application. The extension of this model to allow a number of independent variables is called a **multiple linear regression model**. The multiple regression model is written

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \varepsilon.$$

As in simple linear regression, y is the dependent or response variable, and the $x_i, i = 1, 2, \ldots, m$, are the m independent variables. The β_i are the (m) parameters or regression coefficients, one for each independent variable, and β_0 is the intercept. Also as in simple linear regression, ε is the random error.

The model is called linear regression because the model is linear in the parameters; that is, the coefficients (β_i) are simple (linear) multipliers of the independent variables and the error term (ε) is added (linearly) to the model. As we will see later, the model need not be linear in the independent variables. Although the model contains $(m + 1)$ parameters, it is often referred to as an m-variable model since the intercept coefficient does not correspond to a variable in the usual sense.

We have already alluded to applications of multiple regression models in Chapter 7. Some other applications include the following:

- A refinement of the fertilizer application example in Section 6.2, which relates yield to amounts applied of the three major fertilizer components: nitrogen, phosphorous, and potash.
- The number of "sick days" of school children is related to various characteristics such as waist circumference, height, weight, and age.
- Students' performances are related to scores on a number of different aptitude or mental ability tests.
- Amount of retail sales by an appliance manufacturer is related to expenditures for radio, television, newspaper, magazine, and direct mail advertising.
- Daily fuel consumption for home heating or cooling is related to temperature, cloud cover, and wind velocity.

In many ways, multiple regression is a relatively straightforward extension of simple linear regression. All assumptions and conditions underlying simple linear regression as presented in Chapter 7 remain essentially the same. The computations are more involved and tedious but computers have made these easier. The use of matrix notation and matrix algebra (Appendix B) makes the computations easier to understand and also illustrates the relationship between simple and multiple linear regression.

The potentially large number of parameters in a multiple linear regression model makes it useful to distinguish three different but related purposes for the use of this model:

1. To estimate the mean of the response variable (y) for a given set of values for the independent variables. This is the conditional mean, $\mu_{y|x}$, presented in Section 7.4, and estimated by $\hat{\mu}_{y|x}$. For example, we may want to estimate the mean fuel consumption for a day having a given set of values for the climatic variables. Associated with this purpose of a regression analysis is the question of whether all of the variables in the model are necessary to adequately estimate this mean.

2. To predict the response of a single unit for a given set of values of the independent variables. The point estimate is $\hat{\mu}_{y|x}$, but, because we are not estimating a mean, we will denote this predicted value by $\hat{y}$.
3. To evaluate the relationships between the response variable and the individual independent variables. That is, to make practical interpretations on the values of the regression coefficients, the β_i. For example, what would it mean if the coefficient for temperature in the above fuel consumption example were negative?

8.1.1 The Partial Regression Coefficient

The interpretation of the individual regression coefficients gives rise to an important difference between simple and multiple regression. In a multiple regression model the regression parameters, β_i, called **partial regression coefficients**, are not the same, either computationally or conceptually, as the so-called **total regression coefficients** obtained by individually regressing y on each x.

Definition 8.1 *The **partial regression coefficients** obtained in a multiple regression measure the change in the average value of y associated with a unit increase in the corresponding x, **holding constant all other variables**.*

This means that normally the individual coefficients of an m-variable multiple regression model will not have the same values nor the same interpretations as the coefficients for the m separate simple linear regressions involving the same variables. Many difficulties in using and interpreting the results of multiple regression arise from the fact that the definition of "holding constant," related to the concept of a partial derivative in calculus, is somewhat difficult to understand.

For example, in the application on estimating sick days of school children, the coefficient associated with the height variable measures the increase in sick days associated with a unit increase in height for a population of children all having identical waist circumference, weight, and age. In this application, the total and partial coefficients for height would differ because the total coefficient for height would measure not only the effect of height, but also indirectly measure the effect of the other related variables.

The application on estimating fuel consumption provides a similar scenario: The total coefficient for temperature would indirectly measure the effect of wind and cloud cover. Again this coefficient will differ from the partial regression coefficient because cloud cover and wind are often associated with lower temperatures.

We will see later that the inferential procedures for the partial coefficients are constructed to reflect this characteristic. We will also see that these inferences and associated interpretations are often made difficult by the existence of strong relationships among the several independent variables, a condition known as multicollinearity (Section 8.7).

Because the use of multiple regression models entails many different aspects, this chapter is quite long. Section 8.2 presents the procedures for estimating the coefficients, and Section 8.3 presents the procedure for obtaining the error variance and the inferences about model parameter and other estimates. Section 8.4 contains brief descriptions of correlations that describe the strength of linear relationships involving several variables. Section 8.5 provides some ideas on computer usage and presents computer outputs for examples used in previous sections. The last four sections deal with special models and problems that arise in a regression analysis.

8.2 ESTIMATION OF COEFFICIENTS

In Chapter 7, we showed that the least squares estimates of the parameters of the simple linear regression model are obtained by the solutions to the normal equations:

$$\beta_0 n + \beta_1 \sum x = \sum y,$$
$$\beta_0 \sum x + \beta_1 \sum x^2 = \sum xy.$$

Since there are only two equations in two unknowns, the solutions can be expressed in closed form, that is, as simple algebraic formulas involving the sums, sums of squares, and sums of products of the observed data values of the two variables x and y. These formulas are also used for the partitioning of sums of squares and the resulting inference procedures.

For the multiple regression model with m partial coefficients plus β_0 the least squares estimates are obtained by solving the following set of $(m+1)$ normal equations in $(m+1)$ unknown parameters:

$$\begin{array}{lllll}
\beta_0 n & + \beta_1 \sum x_1 & + \beta_2 \sum x_2 & + \cdots + \beta_m \sum x_m & = \sum y, \\
\beta_0 \sum x_1 & + \beta_1 \sum x_1^2 & + \beta_2 \sum x_1 x_2 & + \cdots + \beta_m \sum x_1 x_m & = \sum x_1 y, \\
\beta_0 \sum x_2 & + \beta_1 \sum x_2 x_1 & + \beta_2 \sum x_2^2 & + \cdots + \beta_m \sum x_2 x_m & = \sum x_2 y, \\
. & . \quad . & . \quad . & . \quad . \quad . & . \\
\beta_0 \sum x_m & + \beta_1 \sum x_m x_1 & + \beta_2 \sum x_m x_2 & + \cdots + \beta_m \sum x_m^2 & = \sum x_m y.
\end{array}$$

The solution to these normal equations provides the estimated coefficients, which are denoted by $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_m$. This set of equations is a straightforward extension of the set of two equations for the simple linear regression model. However, because of the large number of equations and variables, it is not possible to obtain simple formulas that directly compute the estimates of the coefficients as we did for the simple linear regression model in Chapter 7. In other words, the system of equations must be specifically solved for each application of this method. Although procedures are available for performing this task with handheld or desk calculators, the solution

is almost always obtained by computers using methods beyond the scope of this book. We do, however, need to represent symbolically the solutions to the set of equations. This is done with matrices and matrix notation.

Appendix B contains a brief introduction to matrix notation and the use of matrices for representing operations involving systems of linear equations. We will not actually be performing many matrix calculations; however, an understanding and appreciation of this material will make more understandable the material in the remainder of this chapter (as well as that of Chapter 11). Therefore, it is recommended Appendix B be reviewed before continuing.

8.2.1 Simple Linear Regression with Matrices

Estimating the coefficients of a simple linear regression produces a system of two equations in two unknowns, which can be solved explicitly and therefore do not require the use of matrix expressions. However, matrices can be used and we will do so here to illustrate this method.

Recall from Chapter 7 that the simple linear regression model for an individual observation is

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n.$$

Using matrix notation, the regression model is written

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E},$$

where $\mathbf{Y}$ is an $n \times 1$ matrix[1] of observed values of the dependent variable y; $\mathbf{X}$ is an $n \times 2$ matrix in which the first column consists of a column of ones[2] and the second column contains the values of the independent variable x; $\mathbf{B}$ is a 2×1 matrix of the two parameters β_0 and β_1; and $\mathbf{E}$ is an $n \times 1$ matrix of the n values of the random error ε_i.

Placing these matrices in the above expression results in the matrix equation

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

[1] We use the convention that a matrix is denoted by the capital letter of the elements of the matrix. Unfortunately, the capital letters corresponding to β and μ are almost indistinguishable from B and M.

[2] This column may be construed as representing values of an artificial or dummy variable associated with the intercept coefficient, β_0.

Using the principles of matrix multiplication, we can verify that any row of the resulting matrices reproduces the simple linear regression model for an observation:

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

We want to estimate the parameters of the regression model resulting in the estimating equation

$$\hat{\mathbf{M}}_{y|x} = \mathbf{X}\hat{\mathbf{B}},$$

where $\hat{\mathbf{M}}_{y|x}$ is an $n \times 1$ matrix of the $\hat{\mu}_{y|x}$ values, and $\hat{\mathbf{B}}$ is the 2×1 matrix of the estimated coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$. The set of normal equations that must be solved to obtain the least squares estimates is

$$(\mathbf{X}'\mathbf{X})\hat{\mathbf{B}} = \mathbf{X}'\mathbf{Y},$$

where

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \cdot \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum x \\ \sum x & \sum x^2 \end{bmatrix},$$

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum y \\ \sum xy \end{bmatrix}.$$

The equations can now be written

$$\begin{bmatrix} n & \sum x \\ \sum x & \sum x^2 \end{bmatrix} \cdot \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} \sum y \\ \sum xy \end{bmatrix}.$$

Again, using the principles of matrix multiplication, we can see that this matrix equation reproduces the normal equations for simple linear regression (Section 7.3). The matrix representation of the solution of the normal equations is

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

Since we will have occasion to refer to individual elements of the matrix $(\mathbf{X}'\mathbf{X})^{-1}$, we will refer to it as the matrix $\mathbf{C}$, with the subscripts of the elements corresponding to

the regression coefficients. Thus

$$\mathbf{C} = \begin{bmatrix} c_{00} & c_{01} \\ c_{10} & c_{11} \end{bmatrix}.$$

The solution can now be represented by the matrix equation

$$\hat{\mathbf{B}} = \mathbf{C}\mathbf{X}'\mathbf{Y}.$$

For the one-variable regression, the $\mathbf{X}'\mathbf{X}$ matrix is a 2×2 matrix and, as we have noted in Appendix B, the inverse of such a matrix is not difficult to compute. Define the matrix

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}.$$

Then the inverse is

$$\mathbf{A}^{-1} = \begin{bmatrix} \frac{a_{22}}{k} & \frac{-a_{12}}{k} \\ \frac{-a_{21}}{k} & \frac{a_{11}}{k} \end{bmatrix},$$

where $k = a_{11}a_{22} - a_{12}a_{21}$. Substituting the elements of $\mathbf{X}'\mathbf{X}$, we have

$$(\mathbf{X}'\mathbf{X}^{-1}) = \mathbf{C} = \begin{bmatrix} \frac{\sum x^2}{k} & \frac{-\sum x}{k} \\ \frac{-\sum x}{k} & \frac{n}{k} \end{bmatrix},$$

where $k = n\sum x^2 - (\sum x)^2 = nS_{xx}$. Multiplying the matrices to obtain the estimates,

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{bmatrix} \frac{\sum x^2 \sum y}{nS_{xx}} + \frac{-\sum x \sum xy}{nS_{xx}} \\ \frac{-\sum x \sum y}{nS_{xx}} + \frac{n\sum xy}{nS_{xx}} \end{bmatrix}.$$

The second element of $\hat{\mathbf{B}}$ is

$$\frac{n\sum xy - \sum x \sum y}{nS_{xx}} = \frac{\sum xy - (\sum x \sum y/n)}{S_{xx}} = \frac{S_{xy}}{S_{xx}},$$

which is the formula for $\hat{\beta}_1$ given in Section 7.3. A little more algebra (which is left as an exercise for those who are so inclined) shows that the first element is $(\bar{y} - \hat{\beta}_1\bar{x})$, which is the formula for $\hat{\beta}_0$.

We illustrate the matrix approach with the home price data used to illustrate simple linear regression in Chapter 7 (data in Table 7.2). The data matrices (abbreviated

to save space) are

$$\mathbf{X} = \begin{bmatrix} 1 & 0.951 \\ 1 & 1.036 \\ 1 & 0.676 \\ 1 & 1.456 \\ 1 & 1.186 \\ \vdots & \vdots \\ 1 & 1.920 \\ 1 & 2.949 \\ 1 & 3.310 \\ 1 & 2.805 \\ 1 & 2.553 \\ 1 & 2.510 \\ 1 & 3.627 \end{bmatrix} \qquad \mathbf{Y} = \begin{bmatrix} 30.0 \\ 39.9 \\ 46.5 \\ 48.6 \\ 51.5 \\ \vdots \\ 167.5 \\ 169.9 \\ 175.0 \\ 179.0 \\ 179.9 \\ 189.5 \\ 199.0 \end{bmatrix}.$$

Using the transpose and multiplication rules,

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 58 & 109.212 \\ 109.212 & 228.385 \end{bmatrix}, \quad \text{and} \quad \mathbf{X}'\mathbf{Y} = \begin{bmatrix} 6439.998 \\ 13401.788 \end{bmatrix}.$$

The elements of these matrices are the uncorrected or uncentered sums of squares and cross products of the variables x and y and the "variable" represented by the column of ones. For this reason the matrices $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{Y}$ are often referred to as the sums-of-squares and cross-products matrices. Note that $\mathbf{X}'\mathbf{X}$ is symmetric. The inverse is

$$(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{C} = \begin{bmatrix} 0.17314 & -0.08279 \\ -0.08279 & 0.04397 \end{bmatrix},$$

which can be verified using the special inversion method for a 2×2 matrix, or multiplying $\mathbf{X}'\mathbf{X}$ by $(\mathbf{X}'\mathbf{X})^{-1}$, which will result in an identity matrix (except for round-off error). Finally,

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 5.4316 \\ 56.0833 \end{bmatrix},$$

which reproduces the estimated coefficients obtained using ordinary algebra in Section 7.3.

8.2.2 Estimating the Parameters of a Multiple Regression Model

The use of matrix methods to estimate the parameters of a simple linear regression model may appear to be a rather cumbersome method for getting the same results

obtained in Section 7.3. However, if we define the matrices **X** and **B** as

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ . & . & . & \cdots & . \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}, \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ \beta_m \end{bmatrix},$$

then the multiple regression model,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \varepsilon,$$

can be expressed as

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E},$$

and the parameter estimates as

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

Note that these expressions are valid for a multiple regression with any number of independent variables. That is, for a regression with m independent variables, the **X** matrix has n rows and $(m + 1)$ columns. Consequently, matrices **B** and $\mathbf{X}'\mathbf{Y}$ are of order $[(m + 1) \times 1]$ and $\mathbf{X}'\mathbf{X}$ and $(\mathbf{X}'\mathbf{X})^{-1}$ are of order $[(m + 1) \times (m + 1)]$.

The procedure for obtaining the estimates of the parameters of a multiple regression model is thus a straightforward application of using matrices to show the solution of a set of linear equations. First compute the $\mathbf{X}'\mathbf{X}$ matrix

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & \sum x_1 & \sum x_2 & \cdots & \sum x_m \\ \sum x_1 & \sum x_1^2 & \sum x_1 x_2 & \cdots & \sum x_1 x_m \\ \sum x_2 & \sum x_2 x_1 & \sum x_2^2 & \cdots & \sum x_2 x_m \\ . & . & . & \cdots & . \\ \sum x_m & \sum x_m x_1 & \sum x_m x_2 & \cdots & \sum x_m^2 \end{bmatrix},$$

that is, the matrix of sums of squares and cross products of all the independent variables. Next compute the $\mathbf{X}'\mathbf{Y}$ matrix

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} \sum y \\ \sum x_1 y \\ \sum x_2 y \\ \vdots \\ \sum x_m y \end{bmatrix}.$$

The next step is to compute the inverse of $\mathbf{X}'\mathbf{X}$. As we indicated earlier, we do not present here a procedure for this task; instead we assume the inverse has been obtained by a computer, which also provides the estimates by the matrix multiplication

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{C}\mathbf{X}'\mathbf{Y},$$

where, as previously noted, $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$.

8.2.3 Correcting for the Mean, an Alternative Calculating Method

The numerical difficulty of inverting the matrix X′X is somewhat lessened if all variables are first centered, or "corrected" by subtracting the sample means. This yields the corrected sums-of-squares and cross-products matrices. After centering, the intercept is identically 0, and so the column of ones is not needed in the revised X. The values of the partial regression coefficients are unchanged, and the original intercept (for a model with uncentered variables) can be recovered as

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}_1 - \hat{\beta}_2\bar{x}_2 - \cdots - \hat{\beta}_m\bar{x}_m$$

This is easily seen as an extension of the formula given in Chapter 7.

■ Example 8.2

In Example 7.2 we showed how home prices can be estimated using information on sizes by the use of linear regression. We noted that although the regression was significant, the error of estimation was too large to make the model useful.

It was suggested that the use of other characteristics of houses could make such a model more useful.

Solution

In Chapter 7 we used `size` as the single independent variable in a simple linear regression to estimate `price`. To illustrate multiple regression we will estimate `price` using the following five variables:

`age`: age of home, in years,
`bed`: number of bedrooms,
`bath`: number of bathrooms,
`size`: size of home in 1000 ft^2, and
`lot`: size of lot in 1000 ft^2.

In terms of the mnemonic variable names, the model is written

$$\texttt{price} = \beta_0 + \beta_1(\texttt{age}) + \beta_2(\texttt{bed}) + \beta_3(\texttt{bath}) + \beta_4(\texttt{size}) + \beta_5(\texttt{lot}) + \varepsilon.$$

The data for this example are shown in Table 8.2. Note that there is one observation that has no data for `size` as well as several observations with no data on `lot`. Because these observations cannot be used for this regression, the model will be applied to the remaining 51 observations.

Table 8.2 Data on Home Prices for Multiple Regression

Obs	age	bed	bath	size	lot	price
1	21	3	3.0	0.951	64.904	30.000
2	21	3	2.0	1.036	217.800	39.900
3	7	1	1.0	0.676	54.450	46.500
4	6	3	2.0	1.456	51.836	48.600
5	51	3	1.0	1.186	10.857	51.500
6	19	3	2.0	1.456	40.075	56.990
7	8	3	2.0	1.368	.	59.900
8	27	3	1.0	0.994	11.016	62.500
9	51	2	1.0	1.176	6.256	65.500
10	1	3	2.0	1.216	11.348	69.000
11	32	3	2.0	1.410	25.450	76.900
12	2	3	2.0	1.344	.	79.000
13	25	2	2.0	1.064	218.671	79.900
14	31	3	1.5	1.770	19.602	79.950
15	29	3	2.0	1.524	12.720	82.900
16	16	3	2.0	1.750	130.680	84.900
17	20	3	2.0	1.152	104.544	85.000
18	18	4	2.0	1.770	10.640	87.900
19	28	3	2.0	1.624	12.700	89.900
20	27	3	2.0	1.540	5.679	89.900
21	8	3	2.0	1.532	6.900	93.500
22	19	3	2.0	1.647	6.900	94.900
23	3	3	2.0	1.344	43.560	95.800
24	5	3	2.0	1.550	6.575	98.500
25	5	4	2.0	1.752	8.193	99.500
26	27	3	1.5	1.450	11.300	99.900
27	33	2	2.0	1.312	7.150	102.000
28	4	3	2.0	1.636	6.097	106.000
29	0	3	2.0	1.500	.	108.900
30	36	3	2.5	1.800	83.635	109.900
31	5	4	2.5	1.972	7.667	110.000
32	0	3	2.0	1.387	.	112.290

(*Continued*)

Table 8.2 (*Continued*)

Obs	age	bed	bath	size	lot	price
33	27	4	2.0	2.082	13.500	114.900
34	15	3	2.0	.	269.549	119.500
35	23	4	2.5	2.463	10.747	119.900
36	25	3	2.0	2.572	7.090	119.900
37	24	4	2.0	2.113	7.200	122.900
38	1	3	2.5	2.016	9.000	123.938
39	34	3	2.0	1.852	13.500	124.900
40	26	4	2.0	2.670	9.158	126.900
41	26	3	2.0	2.336	5.408	129.900
42	31	3	2.0	1.980	8.325	132.900
43	24	4	2.5	2.483	10.295	134.900
44	29	5	2.5	2.809	15.927	135.900
45	21	3	2.0	2.036	16.910	139.500
46	10	3	2.0	2.298	10.950	139.990
47	3	3	2.0	2.038	7.000	144.900
48	9	3	2.5	2.370	10.796	147.600
49	29	5	3.5	2.921	11.992	149.990
50	8	3	2.0	2.262	.	152.550
51	7	3	3.0	2.456	.	156.900
52	1	4	2.0	2.436	52.000	164.000
53	27	3	2.0	1.920	226.512	167.500
54	5	3	2.5	2.949	11.950	169.900
55	32	4	3.5	3.310	10.500	175.000
56	29	3	3.0	2.805	16.500	179.000
57	1	3	3.0	2.553	8.610	179.900
58	1	3	2.0	2.510	.	189.500
59	33	3	4.0	3.627	17.760	199.000

Figure 8.2 is a scatterplot matrix of the variables involved in this regression using the same format as in Figure 8.1, except that the dependent variable is in the last row and column. The only strong relationship appears to be between `price` and `size`, and there are weaker relationships among `size`, `bed`, `bath`, and `price`.

The first step is to compute the sums of squares and cross products needed for the $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{Y}$ matrices. Note that for this purpose the $\mathbf{X}$ matrix must contain the column of ones, the dummy variable used for the intercept. Since most computer programs automatically generate this variable, it is not usually listed as part of the data. The results of these computations are shown in the top half of Table 8.3. Normally the intermediate calculations presented in this table are not printed by most software and are available with special options invoked here with `PROC REG` of the SAS System. In this table, each element is the sum of products of the variables listed in the row and column headings. For example, the sum of products of `lot`

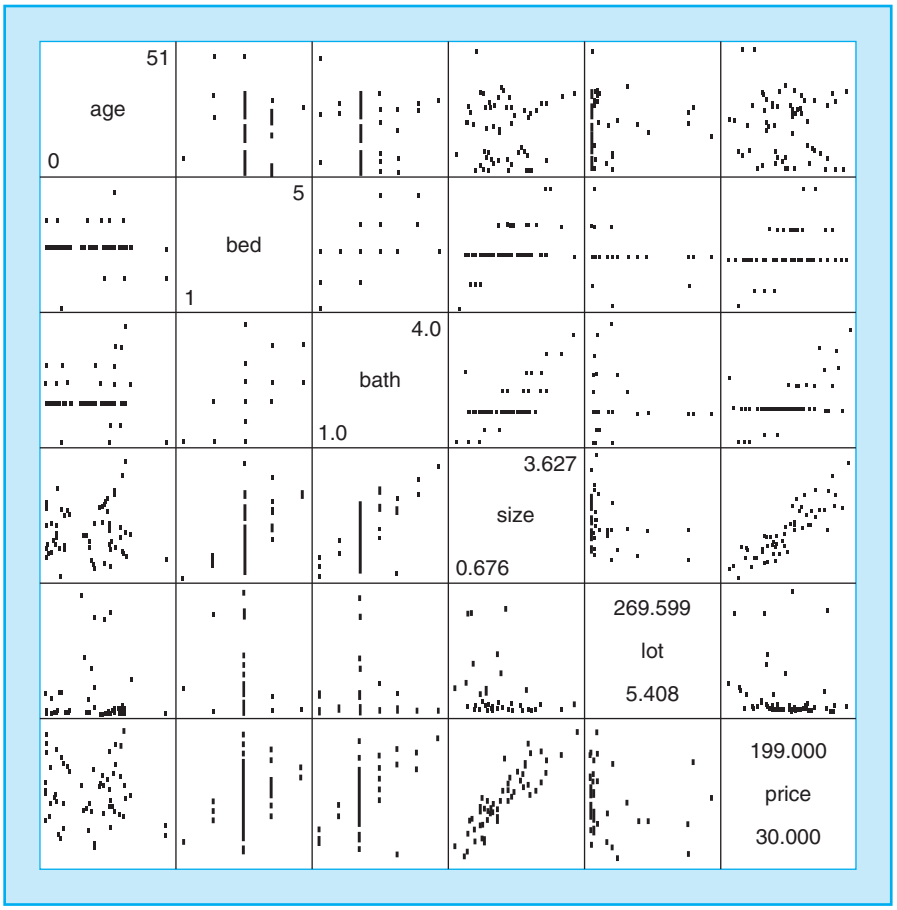

FIGURE 8.2
Scatterplot Matrix for Home Price Data.

and `size` is 3558.9235. Note that the first row and column, labeled `Intercept`, correspond to the column of ones used to estimate β_0, and the last row and column, labeled `price`, correspond to the dependent variable. Thus the first six rows and columns are $\mathbf{X}'\mathbf{X}$, the first six rows of the last column comprise $\mathbf{X}'\mathbf{Y}$, the first six columns of the last row comprise $\mathbf{Y}'\mathbf{X}$ while the last element is $\mathbf{Y}'\mathbf{Y}$, which is the sum of squares of the dependent variable `price`. Note also that the sum of products of `Intercept` and another variable is the sum of values of that variable; the first element is the number of observations used in the analysis, which we have noted is only 51 because of the missing data.

As we have noted, the elements of $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{Y}$ comprise the coefficients of the normal equations. Specifically, the first equation is

$$51\beta_0 + 1045\beta_1 + 162\beta_2 + 109\beta_3 + 96.385\beta_4 + 1708.838\beta_5 = 5580.958.$$

The other equations follow.

The inverse as well as the solution of the normal equations comprise the second half of Table 8.3. Again the row and column variable names identify the elements. The first six rows and columns are the elements of the inverse, $(\mathbf{X}'\mathbf{X})^{-1}$, which we also denote by $\mathbf{C}$. The first six rows of the last column are the matrix of the estimated coefficients ($\hat{\mathbf{B}}$), the first six columns of the last row are the transpose of the matrix of coefficient estimates ($\hat{\mathbf{B}}'$), and the last element corresponding to the row and column labeled with the dependent variable (`price`) is the residual sum of squares, which is defined in the next section.

A sharp-eyed reader will see the number −2.476418E−6 in the second column of row 6. This is shorthand for saying that the number is to be multiplied by 10^{-6}.

It is instructive to verify the calculation for the estimated coefficients. For example, the estimated coefficient for `age` is

$$\begin{aligned}\hat{\beta}_1 &= (-0.003058625)(5580.958) + (0.0001293154)(112308.608)\\ &\quad + (0.0000396856)(18230.154) + (0.0006649237)(12646.3950)\\ &\quad + (-0.000558371)(11688.513) + (-2.476418\text{E}{-6})(165079.37)\\ &= -0.349804.\end{aligned}$$

If you try to verify this on a calculator, the result may differ due to round-off. You may also wish to verify some of the other estimates.

We can now write the equations for the estimated regression:

$$\begin{aligned}\hat{price} &= 35.288 - 0.350(\texttt{age}) - 11.238(\texttt{bed})\\ &\quad - 4.540(\texttt{bath}) + 65.946(\texttt{size}) + 0.062(\texttt{lot}).\end{aligned}$$

This equation may be used to estimate the price for a home having specific values for the independent variables, with the caution that these values are in the range of the values observed in the data set. For example we can estimate the price of the first home shown in Table 8.2 as

$$\begin{aligned}\hat{price} &= 35.288 - 0.349(21) - 11.238(3) - 4.540(3)\\ &\quad + 65.946(0.951) + 0.062(64.904)\\ &= 47.349,\end{aligned}$$

or $47,349, compared to the actual price of $30,000.

The estimated coefficients are interpreted as follows:

- The intercept ($\hat{\beta}_0 = 35.288$) is the estimated mean price (in $1000) of a home for which the values of all independent variables are zero. As in many

Table 8.3 Matrices for Multiple Regression

The REG Procedure
Model Crossproducts X'X X'Y Y'Y

Variable	Intercept	age	bed	bath	size	lot	price
Intercept	51	1045	162	109	96.385	1708.838	5580.958
age	1045	29371	3313	2199.5	1981.721	36060.245	112308.608
bed	162	3313	538	355	318.762	4981.272	18230.154
bath	109	2199.5	355	250	219.4685	3558.9235	12646.395
size	96.385	1981.721	318.762	219.4685	203.085075	2683.133101	11688.513058
lot	1708.838	36060.245	4981.272	3558.9235	2683.133101	202858.09929	165079.36843
price	5580.958	112308.608	18230.154	12646.395	11688.513058	165079.36843	690197.14064

X'X Inverse, Parameter Estimates, and SSE

Intercept	0.6510931798	−0.003058625	−0.130725187	−0.097462177	0.0383208773	−0.000527955	35.287921644
age	−0.003058625	0.0001293154	0.0000396856	0.0006649237	−0.000558371	−2.476418E−6	−0.349804533
bed	−0.130725187	0.0000396856	0.0640254429	−0.007028134	−0.03218064	0.0000709189	−11.23820158
bath	−0.097462177	0.0006649237	−0.007028134	0.1314351128	−0.087657959	−0.00027108	−4.540152056
size	0.0383208773	−0.000558371	−0.03218064	−0.087657959	0.1328335042	0.0003475797	65.946466578
lot	−0.000527955	−2.476418E−6	0.0000709189	−0.00027108	0.0003475797	8.2341898E−6	0.0620508107
price	35.287921644	−0.349804533	−11.23820158	−4.540152056	65.946466578	0.0620508107	13774.049724

applications this coefficient has no practical value, but is necessary in order to specify the equation.

- The coefficient for `age` ($\hat{\beta}_1 = -0.350$) estimates a decrease of \$350 in the average price for each additional year of age, holding constant all other variables.
- The coefficient for `bed` ($\hat{\beta}_2 = -11.238$) estimates a decrease in price of \$11,238 for each additional bedroom, holding constant all other variables.
- The coefficient for `bath` ($\hat{\beta}_3 = -4.540$) estimates a decrease in price of \$4540 for each additional bathroom, holding constant all other variables.
- The coefficient for `size` ($\hat{\beta}_4 = 65.946$) estimates an increase in price of \$65.95 for each additional square foot of the home, holding constant all other variables.
- The coefficient for `lot` ($\hat{\beta}_5 = 0.062$) estimates an increase in price of 62 cents for each additional square foot of lot, holding constant all other variables.

The coefficients for `bed` and `bath` appear to contradict expectations, as one would expect additional bedrooms and bathrooms to increase the price of a home. However, because these are *partial* coefficients, the coefficient for `bed` estimates the change in `price` for an additional bedroom *holding constant* `size` (among others). Now if you increase the number of bedrooms without increasing the size of the home, the bedrooms are smaller and the home seems more crowded and less attractive, hence a lower price. The reason for a negative coefficient for `bath` is not as obvious.

The values of the partial coefficients are therefore generally different from the corresponding total coefficients obtained with simple linear regression. For example, the coefficient for `size` in the one variable regression in Chapter 7 was 56.083, which is certainly different from the value of 65.946 in the multiple regression. You may want to verify this for some of the other variables; for example, the coefficient for the regression of `price` on `bed` will almost certainly result in a positive coefficient.

Comparison of coefficients across variables can be made by the use of **standardized** coefficients. These are obtained by standardizing all variables to have mean zero and unit variance and using these to compute the regression coefficients. However, they are more easily computed by the formula

$$\hat{\beta}_i^* = \hat{\beta}_i \frac{s_{x_i}}{s_y},$$

where $\hat{\beta}_i$ are the usual coefficient estimates, s_{x_i} is the sample standard deviation of x_i, and s_y is the standard deviation of y. This relationship shows that the standardized coefficient is the usual coefficient multiplied by the ratio of the standard deviations of x_i and y. This coefficient shows the change in standard deviation units of y associated with a standard deviation change in x_i, holding constant all other variables.

Standardized coefficients are frequently used whenever the independent variables have very different scales. They are available in most regression programs, but are sometimes labeled BETA, which can be confused with the usual (unstandardized) coefficients. Unlike the unstandardized coefficients, the standardized coefficients are reporting the change in y for a unit change in x_j, where all the x_j have the same scales. Hence, independent variables with large absolute standardized coefficients are regarded as having more impact on y. This does not mean they necessarily have greater statistical significance.

The standardized coefficients for Example 8.2 are shown here as provided by the `STB` option of SAS System `PROC REG`:

Variable	Standardized Estimate
Intercept	0
age	−0.11070
bed	−0.19289
bath	−0.06648
size	1.07014
lot	0.08399

The intercept is zero, by definition. We can now see that `size` has by far the greatest effect, while `bath` and `lot` have the least. We will see, however, that this does not necessarily translate into degree of statistical significance (p value). ■

8.3 INFERENTIAL PROCEDURES

Having estimated the parameters of the regression model, the next step is to perform the associated inferential procedures. As in simple linear regression, the first step is to obtain an estimate of the variance of the random error ε, which is required for performing these inferences.

8.3.1 Estimation of σ^2 and the Partitioning of the Sums of Squares

As in the case of simple linear regression, the variance of the random error σ^2 is estimated from the residuals

$$s^2_{y|x} = \frac{\text{SSE}}{\text{df}} = \frac{\sum(y - \hat{\mu}_{y|x})^2}{(n - m - 1)},$$

where the denominator degrees of freedom $(n - m - 1) = [n - (m + 1)]$ results from the fact that the estimated values, $\hat{\mu}_{y|x}$, are based on $(m + 1)$ estimated parameters: $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_m$.

As in simple linear regression we do not compute the error sum of squares by direct application of the above formula. Instead we use a partitioning of sums of squares:

$$\sum y^2 = \sum \hat{\mu}_{y|x}^2 + \sum (y - \hat{\mu}_{y|x})^2.$$

Note that, unlike the partitioning of sums of squares for simple linear regression, the left-hand side is the uncorrected sum of squares for the dependent variable.[3] Consequently, the term corresponding to the regression sum of squares includes the contribution of the intercept and is therefore not normally used for inferences (see the next subsection).

As with simple linear regression, a shortcut formula is available for the sum of squares due to regression, which is then subtracted from $\sum y^2$ to provide the error sum of squares. Also as in simple linear regression, several equivalent forms are available for computing this quantity, which we will denote by SSR. The most convenient for manual computing is

$$\text{SSR} = \hat{\mathbf{B}}'\mathbf{X}'\mathbf{Y},$$

which results in the algebraic expression

$$\text{SSR} = \hat{\beta}_0 \sum y + \hat{\beta}_1 \sum x_1 y + \cdots + \hat{\beta}_m \sum x_m y.$$

Note that the individual terms are similar to SSR for the simple linear regression model; other equations for this quantity are

$$\text{SSR} = \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \hat{\mathbf{B}}'\mathbf{X}'\mathbf{X}\hat{\mathbf{B}}.$$

The quantities needed for the more convenient formula are available in Table 8.3 as

$$\begin{aligned}\sum y^2 &= 690{,}197.14,\\ \text{SSR} &= (35.288)(5580.958) + (-0.3498)(112308.6) + (-11.2382)(18230.1)\\ &\quad + (-4.5402)(12646.4) + (65.9465)(11688.5) + (0.06205)(165079.4)\\ &= 676{,}423.09;\end{aligned}$$

hence by subtraction

$$\text{SSE} = 690{,}197.14 - 676{,}423.09 = 13{,}774.05.$$

[3] This way of defining these quantities corresponds to the use of matrices consisting of uncorrected sums of squares and cross products with the column of ones for the intercept term. However, using matrices with corrected sums of squares and cross products results in defining TSS and SSR in a manner analogous to those shown in Chapter 7. These different definitions cause minor modifications in computational procedures but the ultimate results are the same.

This is the same quantity printed as the last element of the inverse matrix portion of the output in Table 8.3. As in simple linear regression, it can also be computed directly from the residuals, which are shown later in Table 8.6. The error degrees of freedom are

$$(n - m - 1) = 51 - 5 - 1 = 45,$$

and the resulting mean square error (MSE) provides the estimated variance

$$s^2_{y|x} = 13774.05/45 = 306.09,$$

resulting in an estimated standard deviation of 17.495. This is somewhat smaller than the value of 19.684, which was obtained in Chapter 7 using only `size` as the independent variable. This relatively small decrease suggests that the other variables may contribute only marginally to the fit of the regression equation. The formal test for this is presented in Section 8.3.3.

This estimated standard deviation is interpreted as it was in Section 1.5, and is an often overlooked statistic for assessing the goodness of fit of a regression model. Thus if the distribution of the residuals is reasonably bell shaped, approximately 95% of the residuals will be within two standard deviations of the regression estimates. In the house price data, the standard deviation is 17.495 ($17,495). Hence, using the empirical rule, it follows that approximately 95% of homes are within 2($17,495) or within approximately $35,000 of the values estimated by the regression model.

8.3.2 The Coefficient of Variation

In Section 1.5 we defined the **coefficient of variation** as the ratio of the standard deviation to the mean expressed as a percentage. This measure can also be applied as a measure of residual variation from an estimated regression model. For the 51 houses used in the house prices example, the mean price of homes is $109,431, and the estimated standard deviation is $17,495; hence the coefficient of variation is 0.1599, or 15.99%. Again, using the empirical rule, approximately 95% of homes have prices within 32% of the value estimated by the regression model. It should be noted that this statistic is useful primarily when the values of the dependent variable do not span a large range relative to the mean and is useless for variables that can take negative values.

8.3.3 Inferences for Coefficients

We have already noted that we do not get estimates of the partial coefficients by performing m simple linear regressions using the individual independent variables. Likewise we cannot do the appropriate inferences for the partial coefficients by direct application of simple linear regression methods for the individual coefficients.

Instead we will base our inferences on a general principle for testing hypotheses in a linear statistical model for which regression is a special case.

What we do is to define inferences for these parameters in terms of the effect on the model of imposing certain restrictions on the parameters. The following discussion explains this general principle, which is often called the "general linear test."

General Principle for Hypothesis Testing

Consider two models: a **full** or **unrestricted model** containing all parameters and a **reduced** or **restricted model**, which places some restrictions on the values of some of these parameters. The effects of these restrictions are measured by the decrease in the effectiveness of the restricted model in describing a set of data. In regression analysis the decrease in effectiveness is measured by the increase in the error sum of squares.

The most common inference is to test the null hypothesis that one or more of the coefficients are restricted to a value of 0. This is equivalent to saying that the corresponding independent variables are not used in the restricted model. The measure of the reduction in effectiveness of the restricted model is the increase in the error sum of squares (or, equivalently, the decrease in the model sum of squares) due to imposing the restriction, that is, due to leaving those variables out of the model.

In more specific terms the testing procedure is implemented as follows:

1. Divide the coefficients in $\mathbf{B}$ into two sets represented by matrices $\mathbf{B}_1$ and $\mathbf{B}_2$. That is,

$$\mathbf{B} = \begin{bmatrix} \mathbf{B}_1 \\ \text{- - -} \\ \mathbf{B}_2 \end{bmatrix}.$$

   We want to test the hypotheses

$$H_0\colon \mathbf{B}_2 = \mathbf{0},$$

$$H_1\colon \text{ at least one element of } \mathbf{B}_2 \neq \mathbf{0}.$$

   Denote the number of coefficients in $\mathbf{B}_1$ by q and the number of coefficients in $\mathbf{B}_2$ by p. Note that $p + q = m + 1$. Since the ordering of elements in the matrix of coefficients is arbitrary, $\mathbf{B}_2$ may contain any desired subset of the entire set of coefficients.[4]
2. Perform the regression using all coefficients, that is, using the full model $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$. The error sum of squares for the full model is SSE(B). As we have noted, this sum of squares has $(n - m - 1)$ degrees of freedom.
3. Perform the regression using only the coefficients in $\mathbf{B}_1$, that is, using the restricted model $\mathbf{Y} = \mathbf{X}_1\mathbf{B}_1 + \mathbf{E}$, which is the model specified by H_0. The error sum of squares for the restricted model is SSE(B_1). This sum of squares has $(n - q)$ degrees of freedom.

[4] We seldom perform inferences on β_0; hence this coefficient is normally included in $\mathbf{B}_1$.

4. The difference, $\text{SSE}(B_1) - \text{SSE}(B)$, is the increase in the error sum of squares due to the restriction that the elements in $\mathbf{B}_2$ are zero. This is defined as the **partial** contribution of the coefficients in $\mathbf{B}_2$. Since there are p coefficients in $\mathbf{B}_2$, this sum of squares has p degrees of freedom, which is the difference between the number of parameters in the full and reduced models. For any model TSS = SSR + SSE; hence this difference can also be described as the decrease in the regression (or model) sum of squares due to the deletion of the coefficients in $\mathbf{B}_2$. Dividing the resulting sum of squares by its degrees of freedom provides the corresponding mean square.
5. As before, the ratio of mean squares is the test statistic. In this case the mean square due to the partial contribution of $\mathbf{B}_2$ is divided by the mean square error for the full model. The resulting statistic is compared to the F distribution with $(p, n - m - 1)$ degrees of freedom.

We illustrate with the home price data. We have already noted that the mean square error for the five-variable multiple regression was not much smaller than that using only `size`. It is therefore reasonable to test the hypothesis that the additional four variables do not contribute significantly to the fit of the model. In other words, we want to test the hypothesis that the coefficients for `age`, `bed`, `bath`, and `lot` are all zero.

Formally,

$$H_0\text{: } \beta_{\text{age}} = 0, \quad \beta_{\text{bed}} = 0, \quad \beta_{\text{bath}} = 0, \quad \beta_{\text{lot}} = 0,$$
$$H_1\text{: at least one coefficient is not 0.}$$

Let

$$\mathbf{B}_1 = \begin{bmatrix} \beta_0 \\ \beta_{\text{size}} \end{bmatrix},$$

and

$$\mathbf{B}_2 = \begin{bmatrix} \beta_{\text{age}} \\ \beta_{\text{bed}} \\ \beta_{\text{bath}} \\ \beta_{\text{lot}} \end{bmatrix}.$$

We have already obtained the full model error sum of squares:

$$\text{SSE}(B) = 13774.05 \text{ with } 45 \text{ degrees of freedom.}$$

The restricted model is the one obtained for the example in Chapter 7 that used only `size` as the independent variable. However, we cannot use that result directly because that regression was based on 58 observations while the multiple regression

was based on the 51 observations that had data on `lot` and `size`. Redoing the simple linear regression with `size` using the 51 observations results in

$$\text{SSE}(B_1) = 17253.47 \text{ with } 49 \text{ degrees of freedom.}$$

The difference

$$\text{SSE}(B_1) - \text{SS}(B) = 17253.47 - 13774.05 = 3479.42 \text{ with } 4 \text{ degrees of freedom}$$

is the increase in the error sum of squares due to deleting `age`, `bed`, `bath`, and `lot` from the model and is therefore the partial sum of squares due to those four coefficients. The resulting mean square is 869.855. We use the mean square error for the full model as the denominator for testing the hypothesis that these coefficients are zero, resulting in $F(4, 45) = 869.855/306.09 = 2.842$. The 0.05 critical value for that distribution is 2.58; hence we can reject the hypothesis that all of these coefficients are zero.

8.3.4 Tests Normally Provided by Computer Outputs

Although most computer programs have provisions for requesting almost any kinds of inferences on the regression model, most provide two sets of hypothesis tests as default. These are as follows:

1. H_0: $(\beta_1, \beta_2, \ldots, \beta_m) = 0$, that is, the hypothesis that the entire set of coefficients associated with the m independent variables is zero, with the alternate being that any one or more of these coefficients are not zero. This test is often referred to as the test for the model.
2. H_{oj}: $\beta_j = 0, j = 1, 2, \ldots, m$, that is, the m separate tests that each partial coefficient is zero.

The Test for the Model

The null hypothesis is

$$H_0: (\beta_1, \beta_2, \ldots, \beta_m) = 0.$$

For this test then, the reduced model contains only β_0. The model is

$$y = \beta_0 + \varepsilon$$

or, equivalently,

$$y = \mu + \varepsilon.$$

The parameter μ is estimated by the sample mean $\bar{y}$, and the error sum of squares of this reduced model is

$$\text{SSE}(B_1) = \sum (y - \bar{y})^2 = \sum y^2 - \left(\sum y\right)^2 \Big/ n,$$

with $(n-1)$ degrees of freedom.[5] The error sum of squares for the full model is

$$\text{SSE}(B) = \sum y^2 - \hat{\mathbf{B}}'\mathbf{X}'\mathbf{Y}$$

and the difference yields

$$\text{SSR(regression model)} = \hat{\mathbf{B}}\mathbf{X}'\mathbf{Y} - \left(\sum y\right)^2 \Big/ n,$$

which has m degrees of freedom. Dividing by the degrees of freedom produces the mean square, which is then divided by the mean square error to provide the F statistic for the hypothesis test.

For the home price data the test for the model is

$$H_0: \begin{bmatrix} \beta_{\text{age}} \\ \beta_{\text{bed}} \\ \beta_{\text{bath}} \\ \beta_{\text{size}} \\ \beta_{\text{lot}} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

We have already computed the full model error sum of squares: 13,744.05. The error sum of squares for the restricted model using the information from Table 8.3 is

$$690197.14 - (5580.96)^2/51 = 690194.14 - 610727.74 = 79{,}469.40,$$

the difference

$$\text{SS(model)} = 79{,}469.40 - 13{,}774.05 = 65{,}695.36 \text{ with 5 degrees of freedom,}$$

resulting in a mean square of 13,139.07 with 5 degrees of freedom. Using the full model error mean square of 306.09,

$$F(5, 45) = 42.926,$$

which easily leads to rejection of the null hypothesis and we can conclude that at least one of the coefficients in the model is statistically significant.

Although we have presented this test in terms of the difference in error sums of squares, it is normally presented in terms of the partitioning of sums of squares as presented for simple linear regression in Chapter 7. In this presentation the total corrected sum of squares is partitioned into the model sum of squares and error sum of squares. The test is, of course, the same.

[5] We can now see that what we have called the correction factor for the mean (Section 1.5) is really a sum of squares due to the regression for the coefficient μ or, equivalently, β_0.

For our example then, the total corrected sum of squares is

$$\sum y^2 - \left(\sum y\right)^2 \Big/ n = 690197.14 - (5580.96)^2/51 = 690197.14 - 610727.74$$
$$= 79{,}469.40,$$

which is, of course, the error sum of squares for the restricted model with no coefficients (except the intercept). The full model error sum of squares is 13,774.05; hence the model sum of squares is the difference, 65,695.34. The results of this procedure are conveniently summarized in the familiar analysis of variance table, which, for this example, is shown in the section dealing with computer outputs (Table 8.6 in Section 8.5).

Tests for Individual Coefficients

The testing of hypotheses on the individual partial regression coefficients would seem to require the estimation of m models, each containing $(m-1)$ coefficients. Fortunately a shortcut exists.

It can be shown that the partial sum of squares due to a single partial coefficient, say, β_j, can be computed

$$\text{SSR}(\beta_j) = \hat{\beta}_j^2 / c_{jj}, \quad j = 1, 2, \ldots, m,$$

where c_{jj} is the element on the main diagonal of $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$ corresponding to the variable x_j. This sum of squares has 1 degree of freedom. This can be used for the test statistic

$$F = \frac{(\hat{\beta}_j^2 / c_{jj})}{\text{MSE}},$$

which has $(1, n-m-1)$ degrees of freedom.[6]

The estimated coefficients and diagonal elements of $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$ for the home price data are found in Table 8.3 as

`age`: $\hat{\beta}_1 = -0.3498, c_{11} = 0.0001293,$
`bed`: $\hat{\beta}_2 = -11.2383, c_{22} = 0.064025,$
`bath`: $\hat{\beta}_3 = -4.5401, c_{33} = 0.131435,$
`size`: $\hat{\beta}_4 = 65.9465, c_{44} = 0.132834,$
`lot`: $\hat{\beta}_5 = -0.0621, c_{55} = 8.2341\text{E}{-6}.$

The partial sums of squares and F statistics are

`age`: $\text{SS} = (-0.3498)^2/0.0001293 = 946.327,$
$F = 946.327/306.09 = 3.091,$

[6] As labeled in Section 8.2, the first row and column of $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$ correspond to β_0; hence the row and column corresponding to the jth independent variable will be the $(j+1)$st row and column, respectively. If the computer output uses the names of the independent variable (as in Table 8.3), the desired row and column are easily located.

$$\texttt{bed}: \quad \text{SS} = (-11.2383)^2/0.64025 = 1972.657,$$
$$F = 1972.657/306.09 = 6.445,$$
$$\texttt{bath}: \quad \text{SS} = (-4.5401)^2/0.131435 = 156.827,$$
$$F = 156.827/306.09 = 0.512,$$
$$\texttt{size}: \quad \text{SS} = (65.9465)^2/0.132834 = 32739.7,$$
$$F = 32739.7/306.09 = 106.961,$$
$$\texttt{lot}: \quad \text{SS} = (0.06205)^2/8.23418\text{E}-6 = 467.60,$$
$$F = 467.59/306.09 = 1.528.$$

The 0.05 critical value for $F(1, 45)$ is 4.06, and we reject the hypotheses that the coefficients for `bed` and `size` are zero, but cannot reject the corresponding hypotheses for the other variables. This means that the readily explained negative coefficient for `bed` really exists while evidence for the negative coefficient for `bath` is not necessarily confirmed. Note that we can use this same test for H_0: $\beta_0 = 0$, but because the intercept usually has no practical meaning, the test is not often used, although it is normally printed in computer output.

Note that these partial sums of squares do not constitute a partitioning of the model sum of squares. In other words, the sums of squares for the partial coefficients do not sum to the model sum of squares as was the case with orthogonal contrasts (Section 6.5). This means that, for example, simply because `lot` and `age` cannot individually be deemed significantly different from zero, it does not necessarily follow that the simultaneous addition of these coefficients will not significantly contribute to the model (although they do not in this example).

8.3.5 The Equivalent *t* Statistic for Individual Coefficients

We noted in Chapter 7 that the F test for the hypothesis that the coefficient is zero can be performed by an equivalent t test. The same relationship holds for the individual partial coefficients in the multiple regression model. The t statistic for testing H_0: $\beta_j = 0$ is

$$t = \frac{\hat{\beta}_j}{\sqrt{c_{jj}\text{MSE}}},$$

where c_{jj} is the jth diagonal element of $\mathbf{C}$, and the degrees of freedom are $(n - m - 1)$. It is easily verified that these statistics are the square roots of the F values obtained earlier and they will not be reproduced here. As in simple linear regression, the denominator of this expression is the standard error (or square root of the variance) of the estimated coefficient, which can be used to construct confidence intervals for the coefficients.

In Chapter 7 we noted that the use of the t statistic allowed us to test for specific (nonzero) values of the parameters, and allowed the use of one-tailed tests and the calculation of confidence intervals. For these reasons, most computers provide the

standard errors and t tests. A typical computer output for Example 8.2 is shown in Table 8.6. We can use this output to compute the confidence intervals for the coefficients in the regression equation as follows:

`age`: Std. error $= \sqrt{(0.0001293)(306.09)} = 0.199$
0.95 Confidence interval: $-0.3498 \pm (2.0141)(0.199)$: from -0.7506 to 0.051

`bed`: Std. error $= \sqrt{(0.64025)(306.09)} = 4.427$
0.95 Confidence interval: $-11.2382 \pm (2.0141)(4.427)$: from -20.1546 to -2.3218

`bath`: Std. error $= \sqrt{(0.131435)(306.09)} = 6.343$
0.95 Confidence interval: $-4.5401 \pm (2.0141)(6.343)$: from -17.3155 to 8.2353

`size`: Std. error $= \sqrt{(0.132834)(306.09)} = 6.376$
0.95 Confidence interval: $65.9465 \pm (2.0141)(6.376)$: from 53.1045 to 78.7884

`lot`: Std. error $= \sqrt{(8.234189\text{E}-6)(306.09)} = 0.0502$
0.95 Confidence interval: $0.06205 \pm (2.0141)(0.0502)$: from 0.0391 to 0.1632.

As expected, the confidence intervals of those coefficients deemed statistically significant at the 0.05 level do not include zero.

Finally, note that the tests we have presented are special cases of tests for any linear function of parameters. For example, we may wish to test

$$H_0: \beta_4 - 10\beta_5 = 0,$$

which for the home price data tests the hypothesis that the `size` coefficient is ten times larger than the `lot` coefficient. The methodology for these more general hypothesis tests is presented in Section 11.7.

8.3.6 Inferences on the Response Variable

As in the case of simple linear regression, we may be interested in the precision of the estimated conditional mean as well as predicted values of the dependent variable (see Section 7.5). The formulas for obtaining the variances needed for these inferences are obtained from matrix expressions, and are discussed in Section 11.7. Most computer programs have provisions for computing confidence and prediction intervals and also for providing the associated standard errors. A computer output showing 95% confidence intervals is presented in Section 8.5. A word of caution: Some computer program documentation may not be clear on which interval (confidence on the conditional mean or prediction) is being produced, so read instructions carefully!

The following example is provided as a review of the various steps for a multiple regression analysis.

Example 8.3

Example 7.3 provided a regression model to explain how the departure times (TIME) of lesser snow geese were affected by temperature (TEMP). Although the results were reasonably satisfactory, it is logical to expect that other environmental factors affect departure times.

Solution

Since information on other factors was also collected, we can propose a multiple regression model with the following additional environmental variables:

HUM, the relative humidity,
LIGHT, light intensity, and
CLOUD, percent cloud cover.

The data are given in Table 8.4.

An inspection of the data shows that two observations have missing values (denoted by .) for a variable. This means that these observations cannot be used for the regression analysis. Fortunately, most computer programs recognize missing values and will automatically ignore such observations. Therefore all calculations in this example will be based on the remaining 36 observations.

The first step is to compute $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{Y}$. We then compute the inverse and the estimated coefficients. As before, we will let the computer do this with the results given in Table 8.5 in the same format as that of Table 8.3.

The five elements in the last column, labeled TIME, of the inverse portion contain the estimated coefficients, providing the equation:

$$\widehat{\text{TIME}} = -52.994 + 0.9130(\text{TEMP}) + 0.1425(\text{HUM}) + 2.5160(\text{LIGHT}) + 0.0922(\text{CLOUD}).$$

Unlike the case of the regression involving only TEMP, the intercept now has no real meaning since zero values for HUM and LIGHT cannot exist. The remainder of the coefficients are positive, indicating later departure times for increased values of TEMP, HUM, LIGHT, and CLOUD. Because of the different scales of the independent variables, the relative magnitudes of these coefficients have little meaning and also are not indicators of relative statistical significance.

Note that the coefficient for TEMP is 0.9130 in the multiple regression model, while it was 1.681 for the simple linear regression involving only the TEMP variable. In this case, the so-called total coefficient for the simple linear regression model includes the indirect effect of other variables, while in the multiple regression model, the coefficient measures only the effect of TEMP by holding constant the effects of other variables.

Table 8.4 Snow Goose Departure Times Data

DATE	TIME	TEMP	HUM	LIGHT	CLOUD
11/10/87	11	11	78	12.6	100
11/13/87	2	11	88	10.8	80
11/14/87	−2	11	100	9.7	30
11/15/87	−11	20	83	12.2	50
11/17/87	−5	8	100	14.2	0
11/18/87	2	12	90	10.5	90
11/21/87	−6	6	87	12.5	30
11/22/87	22	18	82	12.9	20
11/23/87	22	19	91	12.3	80
11/25/87	21	21	92	9.4	100
11/30/87	8	10	90	11.7	60
12/05/87	25	18	85	11.8	40
12/14/87	9	20	93	11.1	95
12/18/87	7	14	92	8.3	90
12/24/87	8	19	96	12.0	40
12/26/87	18	13	100	11.3	100
12/27/87	−14	3	96	4.8	100
12/28/87	−21	4	86	6.9	100
12/30/87	−26	3	89	7.1	40
12/31/87	−7	15	93	8.1	95
01/02/88	−15	15	43	6.9	100
01/03/88	−6	6	60	7.6	100
01/04/88	−23	5	.	8.8	100
01/05/88	−14	2	92	9.0	60
01/06/88	−6	10	90	.	100
01/07/88	−8	2	96	7.1	100
01/08/88	−19	0	83	3.9	100
01/10/88	−23	−4	88	8.1	20
01/11/88	−11	−2	80	10.3	10
01/12/88	5	5	80	9.0	95
01/14/88	−23	5	61	5.1	95
01/15/88	−7	8	81	7.4	100
01/16/88	9	15	100	7.9	100
01/20/88	−27	5	51	3.8	0
01/21/88	−24	−1	74	6.3	0
01/22/88	−29	−2	69	6.3	0
01/23/88	−19	3	65	7.8	30
01/24/88	−9	6	73	9.5	30

Table 8.5 Regression Matrices for Snow Goose Departure Times

Model Crossproducts X′X X′Y Y′Y

X′X	INTERCEP	TEMP	HUM
INTERCEP	36	319	3007
TEMP	319	4645	27519
HUM	3007	27519	257927
LIGHT	326.2	3270.3	27822
CLOUD	2280	23175	193085
TIME	−157	1623	−9662

X′X	LIGHT	CLOUD	TIME
INTERCEP	326.2	2280	−157
TEMP	3270.3	23175	1623
HUM	27822	193085	−9662
LIGHT	3211.9	20079.5	−402.8
CLOUD	20079.5	194100	−3730
TIME	−402.8	−3730	9097

X′X Inverse, Parameter Estimates, and SSE

	INTERCEPT	TEMP	HUM
INTERCEP	1.1793413621	0.0085749149	−0.010464297
TEMP	0.0085749149	0.0010691752	0.0000605688
HUM	−0.010464297	0.0000605688	0.0001977643
LIGHT	−0.028115838	−0.00192403	−0.000581237
CLOUD	−0.001558842	−0.000089595	−0.000020914
TIME	−52.99392938	0.9129810924	0.1425316971

	LIGHT	CLOUD	TIME
INTERCEP	−0.028115838	−0.001558842	−52.99392938
TEMP	−0.00192403	−0.000089595	0.9129810924
HUM	−0.000581237	−0.000020914	0.1425316971
LIGHT	0.0086195605	0.0002464973	2.5160019069
CLOUD	0.0002464973	0.0000294652	0.0922051991
TIME	2.5160019069	0.0922051991	2029.6969929

For the second step we compute the partitioning of the sums of squares. The residual sum of squares

$$\begin{aligned}\text{SSE} &= \sum y^2 - \hat{\mathbf{B}}'\mathbf{X}'\mathbf{Y} \\ &= 9097 - [(-52.994)(-157) + (0.9123)(1623) + (0.1425)(-9662) \\ &\quad + (2.5160)(-402.8) + (0.09221)(-3730)],\end{aligned}$$

which is available in the computer output as the last element of the inverse portion and is 2029.70. The estimated variance is MSE $= 2029.70/(36-5) = 65.474$, and the estimated standard deviation is 8.092. This value is somewhat smaller than the 9.96 obtained for the simple linear regression involving only TEMP.

The model sum of squares is

$$\begin{aligned}\text{SSR(regression model)} &= \hat{\mathbf{B}}'\mathbf{X}'\mathbf{Y} - \left(\sum y\right)^2 \Big/ n \\ &= 7067.30 - 684.69 = 6382.61.\end{aligned}$$

The degrees of freedom for this sum of squares is 4; hence the model mean square is $6382.61/4 = 1595.65$. The resulting F statistic is $1595.65/65.474 = 24.371$, which clearly leads to the rejection of the null hypothesis of no regression. These results are summarized in an analysis of variance table shown in Table 8.7 in Section 8.5.

In the final step we use the standard errors and t statistics for inferences on the coefficients. For the TEMP coefficient, the estimated variance of the estimated coefficient is

$$\begin{aligned}\hat{\text{var}}(\hat{\beta}_{\text{TEMP}}) &= c_{\text{TEMP,TEMP}}\text{MSE} \\ &= (0.001069)(65.474) \\ &= 0.0700,\end{aligned}$$

which results in an estimated standard error of 0.2646. The t statistic for the null hypothesis that this coefficient is zero is

$$t = 0.9130/0.2646 = 3.451.$$

Assuming a desired significance level of 0.05, the hypothesis of no temperature effect is clearly rejected. Similarly, the t statistics for HUM, LIGHT, and CLOUD are 1.253, 3.349, and 2.099, respectively. When compared with the tabulated two-tailed 0.05 value for the t distribution with 31 degrees of freedom of 2.040, the coefficient for HUM is not significant, while LIGHT and CLOUD are. The p values are shown later in Table 8.7, which presents computer output for this problem. Basically this means that departure times appear to be affected by increasing levels of temperature, light, and cloud cover, but there is insufficient evidence to state that humidity affects the departure times. ■

8.4 CORRELATIONS

In Section 7.6 we noted that the correlation coefficient provides a convenient index of the strength of the linear relationship between two variables. In multiple

regression, two types of correlations describe strengths of linear relationships among the variables in a regression model:

1. multiple correlation, which describes the strength of the linear relationship of the dependent variable with the set of independent variables, and
2. partial correlation, which describes the strength of the linear relationship associated with a partial regression coefficient.

Other types of correlations used in some applications but not presented here are multiple partial and part (or semipartial) correlations (Kleinbaum *et al.*, 1998, Chapter 10).

8.4.1 Multiple Correlation

Definition 8.2 ***Multiple correlation*** *describes the maximum strength of a linear relationship of one variable with a linear function of a set of variables.*

In Section 7.6, the sample correlation between two variables x and y was defined as

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}.$$

With the help of a little algebra it can be shown that the absolute value of this quantity is equal to the correlation between the observed values of y and $\hat{\mu}_{y|x}$, the values of the variable y estimated by the linear regression of y on x. Thus, for example, the correlation coefficient can also be calculated using the values in the columns labeled `size` and `Predict` in Table 7.3. This definition of the correlation coefficient can be applied to a multiple linear regression and the resulting correlation coefficient is called the **multiple correlation coefficient**, which is usually denoted by R. Also, as in simple linear regression, the square of R, the **coefficient of determination**, is

$$R^2 = \frac{\text{SS due to regression model}}{\text{total SS for } y \text{ corrected for the mean}}.$$

In other words, the coefficient of determination measures the proportional reduction in variability about the mean resulting from the fitting of the multiple regression model. As in simple linear regression there is a correspondence between the coefficient of determination and the F statistic for testing the existence of the model:

$$F = \frac{(n - m - 1)R^2}{m\,(1 - R^2)}.$$

Also as in simple linear regression, the coefficient of determination must take values between and including 0 and 1 where a value of 0 indicates the linear relationship is nonexistent, and a value of 1 indicates a perfect linear relationship.

8.4.2 How Useful Is the R^2 Statistic?

The apparent simplicity of this statistic, which is often referred to as "R-square," makes it a popular and convenient descriptor of the effectiveness of a multiple regression model. This very simplicity has, however, made the coefficient of determination an often abused statistic. There is no rule or guideline as to what value of this statistic signifies a good regression. For some data, especially that from the social and behavioral sciences, coefficients of determination of 0.3 are often considered quite good, while in fields where random fluctuations are of smaller magnitudes, for example, engineering, coefficients of determination of less than 0.95 may imply an unsatisfactory fit. Incidentally, for the home prices model, the coefficient of determination is 0.8267. This is certainly considered to be high for many applications, yet the residual standard deviation of \$17,495 leaves much to be desired.

As more independent variables are added to a regression model, R^2 will increase even if the new variables are simply noise! This is because there is almost always some tiny chance correlation that least squares can use to explain the dependent variable. In fact, if there are $(n-1)$ independent variables in a regression with n observations, R^2 will be unity. To compare models with different numbers of independent variables, it is slightly safer to use the **adjusted R-square**, which is the proportional reduction in the mean squared error rather than in the sum of squared errors. This statistic has some interpretive problems (it can actually be negative in some situations with low R^2). However, it captures the idea that good fit should be balanced against the complexity of the model, as indexed by the number of independent variables. There are a number of such statistics, including Mallows' $C(p)$, discussed in Section 8.8.

As noted in Section 8.3, the residual standard deviation may be a better indicator of the fit of the model.

8.4.3 Partial Correlation

Definition 8.3 *A **partial correlation** coefficient describes the strength of a linear relationship between two variables, holding constant a number of other variables.*

As noted in Section 7.6, the strength of the linear relationship between x and y was measured by the simple correlation between these variables, and the simple linear regression coefficient described their relationship. Just as a partial regression coefficient shows the relationship of y to one of the independent variables, holding constant the other variables, a **partial correlation coefficient** measures the strength of the relationship between y and one of the independent variables, holding constant all other variables in the model. This means that the partial correlation measures the strength of the linear relationship between two variables after "adjusting" for relationships involving all the other variables.

Suppose independent variables $x_1, x_2, \ldots, x_m$ are already in a regression and we are considering new candidate independent variables $x^*_{m+1}, \ldots, x^*_{m+k}$. Let e be the residuals from the current regression of y on $x_1, x_2, \ldots, x_m$, and $f_{m+1}, \ldots, f_{m+k}$ be the

residuals from regressing each of the candidate variables on the same $x_1, x_2, \ldots, x_m$. The residuals e represent the portion of y that we have not yet succeeded in explaining. The residuals f represent the portion of each candidate variable that is not redundant with the current set of x. It makes sense, then, that the most promising new independent variable is the one having the strongest correlation coefficient between e and $f_{m+j}, j = 1, 2, \ldots, k$. This correlation coefficient is exactly the **partial correlation** of y with x_{m+j} given $x_1, x_2, \ldots, x_m$.

As with all correlations, there is an exact relationship to the test statistic of the corresponding regression coefficient. For example, suppose we wanted to know whether x_{m+j} would significantly improve a regression that already contained $x_1, x_2, \ldots, x_m$, and we had computed the partial correlation coefficient r. The t statistic for testing whether the regression coefficient of the new variable is zero is

$$|t| = \sqrt{\frac{(n - m - 1)r^2}{(1 - r^2)}}$$

Far less cumbersome methods exist for computing partial correlation coefficients, and these are implemented in most computer packages. We present the ideas simply to justify partial correlation coefficients as a means of identifying good candidates for new variables to include in a regression.

As an illustration of the use of partial correlation coefficients, consider the data in Example 8.2, and a regression model for `price` that already includes the independent variable `size`. The `PROC CORR` in the SAS System gives the partial correlation coefficients of `age`, `bed`, `bath`, and `lot` with `price` (after adjusting for `size`) as -0.206, -0.353, -0.042, and $+0.165$, respectively. We would select `bed` as the most promising additional independent variable.

8.5 USING THE COMPUTER

Almost all regressions are performed using statistical software packages. Reputable packages will have at least one very powerful module designed for multiple regression. As we will see later, outputs from these packages always contain some common information. The information may be arranged differently, but despite minor variations is usually easy to identify. For example, the coefficient of determination is labeled `R-Square`, and given as a proportion in the SAS System's `PROC REG`, but labeled `R-Sq` and given as a percentage in Minitab. These variations are generally simple to spot.

Some variations in labeling are more extreme. Be aware that p values are labeled in a variety of ways. The SAS System commonly uses `Prob`, reminding us that a p value is a probability of a test statistic value as or more extreme than that actually observed. Minitab often simply uses `p`. SPSS, on the other hand, often labels the values `Sign.`, an abbreviation for "observed significance level," and some modules

of the SAS System do the same. Standardized regression coefficients sometimes are labeled `B` and sometimes `BETA`, and a few packages use the same for the unstandardized coefficients! Fortunately, most packages offer voluminous documentation including annotated samples of output with all elements carefully defined. Learning to navigate the documentation is an essential skill.

Example 8.4 : Example 8.2 Revisited

Table 8.6 contains the output from `PROC REG` of the SAS System for the multiple regression model for the home price data we have been using as an example (we have omitted some of the output to save space). The implementation of this program required the following specifications:

1. The name of the program; in this case it is `PROC REG`.
2. The name of the dependent and independent variables; in this case `price` is the dependent variable and `age`, `bed`, `bath`, `size`, and `lot` are the independent variables. The intercept is not specified since most computer programs automatically assume that an intercept will be included in the model.
3. Options to print, in addition to the standard or default output, the predicted and residual values, the standard errors of the estimated mean, and the 95% confidence intervals for the estimated means.

Although much of the output in Table 8.6 is self-explanatory, a brief summary is presented here. The reader should verify all results that compare with those presented in the previous sections. Also useful are comparisons with output from other computer packages, if available.

Solution

The output begins by giving the name of the dependent variable. This identifies the output in case several analyses have been run in one computer job. The first tabular presentation contains the overall partitioning of the sums of squares and the F test for the model. The notation `Corrected Total` is used to denote that this is the total sum of squares corrected for the mean; hence the model sum of squares is presented in the manner we used for simple linear regression. That is, it is the sum of squares due to the regression after the mean has already been estimated.

The next section gives some miscellaneous statistics. `Root MSE` is the residual standard deviation, which is the square root of the mean square error. `Dependent Mean` is $\bar{y}$ and `R-Square` is the coefficient of determination. `Adj R-Sq` is the adjusted coefficient of determination. `Coeff Var` is the coefficient of variation (in %) as defined in Section 8.3.

The third portion contains the parameter (coefficient) estimates and associated statistics: the standard errors and t statistics and their p values, which are labeled

Table 8.6 Output for Multiple Regression

The REG Procedure
Model: MODEL1
Dependent Variable: price

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	65696	13139	42.93	<.0001
Error	45	13774	306.08999		
Corrected Total	50	79470			

Root MSE	17.49543	R-Square	0.8267
Dependent Mean	109.43055	Adj R-Sq	0.8074
Coeff Var	15.98770		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	35.28792	14.11712	2.50	0.0161
age	1	−0.34980	0.19895	−1.76	0.0855
bed	1	−11.23820	4.42691	−2.54	0.0147
bath	1	−4.54015	6.34279	−0.72	0.4778
size	1	65.94647	6.37644	10.34	<.0001
lot	1	0.06205	0.05020	1.24	0.2229

Output Statistics

Obs	Dep Var price	Predicted Value	Std Error Mean Predict	95% CL Mean		Residual
1	30.0000	47.3494	10.2500	26.7049	67.9939	−17.3494
2	39.9000	66.9823	9.0854	48.6834	85.2812	−27.0823
3	46.5000	65.0194	8.9813	46.9302	83.1087	−18.5194
4	48.6000	89.6287	4.1333	81.3039	97.9535	−41.0287
.	.	.	.	.	.	.
.	.	.	(Observations Omitted)	.	.	.
.	.	.	.	.	.	.
57	179.9000	156.4986	6.3606	143.6877	169.3096	23.4014
58	189.5000	.	.	.	.	.
59	199.0000	212.1590	10.5356	190.9392	233.3788	−13.1590

Sum of Residuals	0
Sum of Squared Residuals	13774
Predicted Residual SS (PRESS)	19927

Pr>|t|. The parameter estimates are identified by the names of the corresponding independent variables, and the estimate of β_0 is labeled Intercept.

The last portion contains some optional statistics for the individual observations. The values in the columns labeled Dep Var price and Predicted Value are self-explanatory. The column labeled Std Error Mean Predict contains the standard errors of the estimated conditional means. The column 95% CL Mean contains the 0.95 confidence limits of the conditional mean.

Finally the sum and sum of squares of the actual residuals are given. The Sum of Residuals should be zero, which it is, and the Sum of Squared Residuals should be equal to the error sum of squares obtained in the analysis of variance table.[7] ■

■ Example 8.5 : Example 8.3 Revisited

Table 8.7 shows the results of implementing the lesser snow geese departure regression on Minitab using the REGRESS command. This command required the specification of the name of the dependent variable and the number of independent variables in the model followed by a listing of names of these variables. No additional options were requested.

Solution

As we have noted before, the output is somewhat similar to that obtained with the SAS System, and the results are the same as those presented in Example 8.3. This output actually gives the estimated model in equation form as well as a listing of coefficients and their inference statistics. Also the output states that two observations could not be used because of missing values. In the SAS System, this information is given in output we did not present for that example.

In addition, the Minitab output contains two items that were not in the SAS output: a set of sequential sums of squares (SEQ SS) and a listing of two unusual observations. The sequential sums of squares are not particularly useful for this example but will be used in polynomial regression, which is presented in Section 8.6. Because these have a special purpose, they must be specifically requested when using the SAS System.

The two unusual observations are identified as having large "Studentized residuals," which are residuals that have been standardized to look like t statistics; hence values exceeding a critical value of t are deemed to be unusual. A discussion of unusual observations is presented in Section 8.9.

Listings of all predicted and residual values, confidence intervals, etc., can be obtained as options for both of these computer programs. In general, we can see

[7] If there is more than a minimal difference between the two, severe round-off errors have probably occurred.

Table 8.7 Snow Goose Regression with Minitab

The regression equation is time = -53.0 + 0.913 temp + 0.143 hum + 2.52 light + 0.0922 cloud 36 cases used 2 cases contain missing values

Predictor	Coef	Stdev	t-ratio	p
Constant	−52.994	8.787	−6.03	0.000
temp	0.9130	0.2646	3.45	0.002
hum	0.1425	0.1138	1.25	0.220
light	2.5160	0.7512	3.35	0.002
cloud	0.09221	0.04392	2.10	0.044

s = 8.092 R-sq = 75.9% R-sq(adj) = 72.8%

Analysis of Variance

SOURCE	df	SS	MS	F	p
Regression	4	6382.6	1595.7	24.37	0.000
Error	31	2029.7	65.5		
Total	35	8412.3			

SOURCE	df	SEQ SS
temp	1	4996.6
hum	1	633.3
light	1	464.2
cloud	1	288.5

Unusual Observations

Obs.	temp	time	Fit Stdev.	Fit	Residual	St. Resid
4	20.0	−11.00	12.40	2.84	−23.40	−3.09R
12	18.0	25.00	8.93	2.65	16.07	2.10R

R denotes an obs. with a large st. resid.

that different computer packages generally provide equivalent results, although they may provide different automatic and optional outputs. ■

8.6 SPECIAL MODELS

It is rather well known that straight line relationships of the type described by a multiple linear regression model do not often occur in the real world. Nevertheless, such models enjoy wide use, primarily because they are relatively easy to implement, but also because they provide useful approximations for other functions, especially over a limited range of values of the independent variables. However, strictly linear regression models are not always effective; hence we present in this section some

methods for implementing regression models that do not necessarily imply straight line relationships.

As we have noted a linear regression model is constrained to be linear in the **parameters**, that is, the β_i and ε, but not necessarily linear in the independent variables. Thus, for example, the independent variables may be nonlinear functions of observed variables that describe curved responses, such as $x^2, 1/x, \sqrt{x}$, etc.

8.6.1 The Polynomial Model

The most popular such function is the **polynomial** model, which involves powers of the independent variables. Fitting a polynomial model is usually referred to as "curve fitting" because it is used to fit a curve rather than to explain the relationship between the dependent and independent variable(s). That is, the interest is in the nature of the fitted response curve rather than in the partial regression coefficients. The polynomial model is very useful for this purpose, as it is easy to implement and provides a reasonable approximation to virtually any function within a limited range.

Given observations on a dependent variable y and two independent variables x_1 and x_2, we can estimate the parameters of the polynomial model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \varepsilon,$$

by redefining variables

$$\begin{aligned} w_1 &= x_1, \\ w_2 &= x_1^2, \\ w_3 &= x_2, \\ w_4 &= x_2^2, \\ w_5 &= x_1 x_2, \end{aligned}$$

and performing a multiple linear regression using the model

$$y = \beta_0 + \beta_1 w_1 + \beta_2 w_2 + \beta_3 w_3 + \beta_4 w_4 + \beta_5 w_5 + \varepsilon.$$

This is an ordinary multiple linear regression model using the w's as independent variables.

■ Example 8.6

Biologists are interested in the characteristics of growth curves, that is, finding a model for describing how organisms grow with time. Relationships of this type

tend to be curvilinear in that the rate of growth decreases with age and eventually stops altogether. A polynomial model is sometimes used for this purpose.

This example concerns the growth of rabbit jawbones. Measurements were made on lengths of jawbones for rabbits of various ages. The data are given in Table 8.8, and the plot of the data is given in Fig. 8.3 where the line is the estimated polynomial regression line described below. Two points for much older rabbits are shown on the plot but not used in the regression.

Table 8.8 Rabbit Jawbone Length

AGE	LENGTH	AGE	LENGTH	AGE	LENGTH
0.01	15.5	0.41	29.7	2.52	49.0
0.20	26.1	0.83	37.7	2.61	45.9
0.20	26.3	1.09	41.5	2.64	49.8
0.21	26.7	1.17	41.9	2.87	49.4
0.23	27.5	1.39	48.9	3.39	51.4
0.24	27.0	1.53	45.4	3.41	49.7
0.24	27.0	1.74	48.3	3.52	49.8
0.25	26.0	2.01	50.7	3.65	49.9
0.26	28.6	2.12	50.6		
0.34	29.8	2.29	49.2		

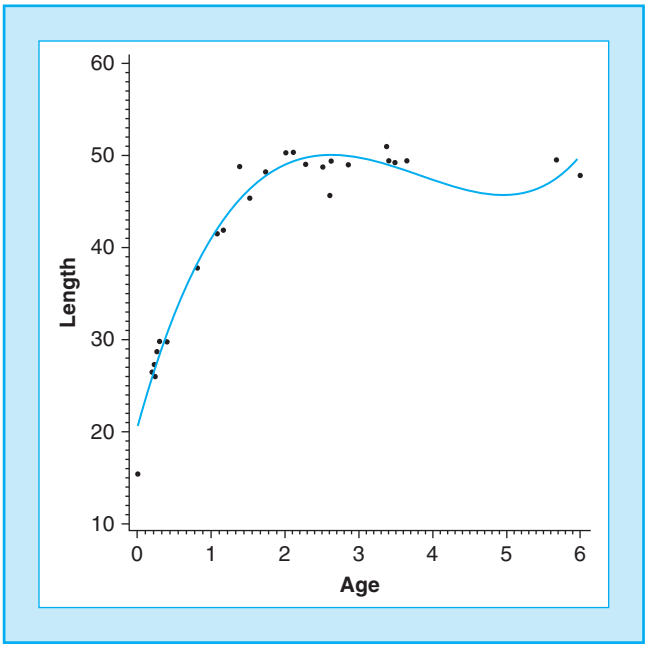

FIGURE 8.3
Polynomial Regression Plot.

Solution

We will use a fourth-degree polynomial model for estimating the relationship of LENGTH to AGE. This model contains as independent variables the first four powers of the variable AGE. Since we will use computer output to show the results, we use the following variable names:

LENGTH, the dependent variable, is the length (in mm) of the jawbone.
AGE is the age (in days) of the rabbits divided by 100. The computations for a polynomial regression model may be subject to considerable round-off error, especially when the independent variable contains both very large and small numbers. Round-off error is reduced if the independent variable can be scaled so that values lie between 0.1 and 10. In this example only one scaled value is outside that recommended range.
A2 = $(\text{AGE})^2$.
A3 = $(\text{AGE})^3$.
A4 = $(\text{AGE})^4$.

In terms of the computer,[8] the linear regression model now is

$$\text{LENGTH} = \beta_0 + \beta_1(\text{AGE}) + \beta_2(\text{A2}) + \beta_3(\text{A3}) + \beta_4(\text{A4}) + \varepsilon.$$

The results of the regression analysis using this model, again obtained by PROC REG of the SAS System, are shown in Table 8.9. The overall statistics for the model in the top portion of the output clearly show that the model is statistically significant, $F(4, 23) = 291.35, p$ value < 0.0001. The estimated polynomial equation is

$$\widehat{\text{LENGTH}} = 18.58 + 36.38(\text{AGE}) - 15.69(\text{AGE})^2 + 2.86(\text{AGE})^3 - 0.175(\text{AGE})^4.$$

The individual coefficients in a polynomial equation usually have no practical interpretation; hence the test statistics for these coefficients also have little use. In fact, a pth-degree polynomial should always include all terms with lower powers. It is of interest, however, to ascertain the lowest degree of polynomial required to describe the relationship adequately. To assist in answering this question, many computer programs provide a set of **sequential** sums of squares, which show how the model sum of squares is increased (or error sum of squares is decreased) as higher order polynomial terms are added to the model.[9] In the computer output

[8] The powers of AGE are computed in the data input stage. Some computer programs allow the specifications of polynomial terms as part of the regression program.

[9] Sequential sums of squares of this type are automatically provided by orthogonal polynomial contrasts as discussed in Section 6.5. Of course, they cannot be used here because the values of the independent variable are not equally spaced. Furthermore, the ease of direct implementation of polynomial regression on computers make orthogonal polynomials a relatively unattractive alternative except for small experiments such as those presented in Section 6.5 and also Chapter 9.

Table 8.9 Polynomial Regression

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob > F
Model	4	3325.65171	831.41293	291.346	0.0001
Error	23	65.63507	2.85370		
C Total	27	3391.28679			
Root MSE		1.68929		R-square	0.9806
Dep mean		39.26071		Adj R-sq	0.9773
C.V.		4.30275			

Parameter Estimates

Variable	df	Parameter Estimate	Standard Error	T for H0: Parameter = 0	Prob > |T|
INTERCEP	1	18.583478	1.27503661	14.575	0.0001
AGE	1	36.380515	6.44953987	5.641	0.0001
A2	1	−15.692308	7.54002073	−2.081	0.0487
A3	1	2.860487	3.13335286	0.913	0.3708
A4	1	−0.175485	0.42335354	−0.415	0.6823

Variable	df	Type I SS
INTERCEP	1	43159
AGE	1	2715.447219
A2	1	552.468707
A3	1	57.245461
A4	1	0.490324

in Table 8.9, these sequential sums of squares are called `Type I SS`.[10] Since these are 1 degree of freedom sums of squares, we can use them to build the most appropriate model by sequentially using an F statistic to test for the significance of each added polynomial term. For this example these tests are as follows:

1. The sequential sum of squares for `INTERCEP` is the correction for the mean of the dependent variable. This quantity can be used to test the hypothesis that the mean of this variable is zero; this is seldom a meaningful test.
2. The sequential sum of squares for `AGE` (2715.4) is divided by the mean square error (2.8537) to get an F ratio of 951.55. We use this to test the hypothesis that a linear regression does not fit the data better than the mean. This hypothesis is rejected.

[10] Remember that these were automatically printed with Minitab, while `PROC REG` of the SAS System required a special option. Also in the Minitab output they were called `SEQ SS`. This should serve as a reminder that not all computer programs produce the same default output or use identical terminology!

3. The sequential sum of squares for A2, the quadratic term in AGE, is divided by the mean square error to test the hypothesis that the quadratic term is not needed. The resulting F ratio of 193.60 rejects this hypothesis.
4. In the same manner, the sequential sums of squares for A3 and A4 produce F ratios that indicate that the cubic term is significant but the fourth-degree term is not.

Sequential sums of squares are additive: They add to the sum of squares for a model containing all coefficients. Therefore they can be used to reconstruct the model and error sums of squares for any lower order model. For example, if we want to compute the mean square error for the third-degree polynomial, we can subtract the sequential sums of squares for the linear, quadratic, and cubic coefficients from the corrected total sum of squares,

$$3391.29 - 2715.44 - 552.47 - 57.241 = 66.12,$$

and divide by the proper degrees of freedom ($n - 1 - 3 = 24$). The result for our example is 2.755.[11] It is of interest to note that this is actually smaller than the mean square error for the full fourth-degree model (2.8537 from Table 8.9). For this reason it is appropriate to reestimate the equation using only the linear, quadratic, and cubic terms. This results in the equation

$$\widehat{\text{LENGTH}} = 18.97 + 33.99(\text{AGE}) - 12.67(\text{AGE})^2 + 1.57(\text{AGE})^3.$$

This equation can be used to estimate the average jawbone length for any age within the range of the data. For example, for AGE = 0.01 (one day) the estimated jawbone length is 19.2, compared with the observed value of 15.5. The plot of the estimated jawbone lengths is shown as the solid line in Fig. 8.3. The estimated curve is reasonably close to the observed values with the possible exception of the first observation where the curve overestimates the jawbone length. The nature of the fit can be examined by a residual plot, which is not reproduced here.

We have repeatedly warned that estimated regression equations should not be used for extrapolation. This is especially true of polynomial models, which may exhibit drastic fluctuations in the estimated response beyond the range of the data. For example, using the estimated polynomial regression equation, estimated jawbone lengths for rabbits aged 500 and 700 days are 68.31 and 174.36 mm, respectively!

Although polynomial models are frequently used to estimate responses that cannot be described by straight lines, they are not always useful. For example, the cubic polynomial for the rabbit jawbone lengths shows a "hook" for the older

[11] Equivalently, the sequential sum of squares for the fourth power coefficient may be added to the full model error sum of squares.

ages, a characteristic not appropriate for growth curves. For this reason, other types of response models are available.

8.6.2 The Multiplicative Model

Another model that describes a curved line relationship is the **multiplicative model**

$$y = e^{\beta_0} x_1^{\beta_1} x_2^{\beta_2} \ldots x_m^{\beta_m} e^{\varepsilon},$$

where e refers to the Naperian constant used as the basis for natural logarithms. This model is quite popular and has many applications. The coefficients, sometimes called **elasticities**, indicate the *percent* change in the dependent variable associated with a *one-percent* change in the independent variable, holding constant all other variables.

Note that the error term e^{ε} is a multiplicative factor. That is, the value of the deterministic portion is *multiplied* by the error. The expected value of this error, when $\varepsilon = 0$, is one. When the random error is positive the multiplicative factor is greater than 1; when negative it is less than 1. This type of error is quite logical in many applications where variation is proportional to the magnitude of the values of the variable.

The multiplicative model can be made linear by the logarithmic transformation,[12] that is,

$$\log(y) = \beta_0 + \beta_1 \log(x_1) + \beta_2 \log(x_2) + \cdots + \beta_m \log(x_m) + \varepsilon.$$

This model is easily implemented. Most statistical software have provisions for making transformations on the variables in a set of data. ■

■ Example 8.7

We illustrate the multiplicative model with a biological example. It is desired to study the size range of squid eaten by sharks and tuna. The beak (mouth) of squid is indigestible hence it is found in the digestive tracts of harvested fish; therefore, it may be possible to predict the total squid weight with a regression that uses various beak dimensions as predictors. The beak measurements and their computer names are

RL = rostral length,
WL = wing length,
RNL = rostral to notch length,

[12] The logarithm base e is used here. The logarithm base 10 (or any other base) may be used; the only difference will be in the intercept.

NWL = notch to wing length,
W = width.

The dependent variable WT is the weight of squid.

Data are obtained on a sample of 22 specimens. The data are given in Table 8.10. The specific definitions or meaning of the various dimensions are of little importance for our purposes except that all are related to the total size of the squid.

Table 8.10 Squid Data

Obs	RL	WL	RNL	NWL	W	WT
1	1.31	1.07	0.44	0.75	0.35	1.95
2	1.55	1.49	0.53	0.90	0.47	2.90
3	0.99	0.84	0.34	0.57	0.32	0.72
4	0.99	0.83	0.34	0.54	0.27	0.81
5	1.05	0.90	0.36	0.64	0.30	1.09
6	1.09	0.93	0.42	0.61	0.31	1.22
7	1.08	0.90	0.40	0.51	0.31	1.02
8	1.27	1.08	0.44	0.77	0.34	1.93
9	0.99	0.85	0.36	0.56	0.29	0.64
10	1.34	1.13	0.45	0.77	0.37	2.08
11	1.30	1.10	0.45	0.76	0.38	1.98
12	1.33	1.10	0.48	0.77	0.38	1.90
13	1.86	1.47	0.60	1.01	0.65	8.56
14	1.58	1.34	0.52	0.95	0.50	4.49
15	1.97	1.59	0.67	1.20	0.59	8.49
16	1.80	1.56	0.66	1.02	0.59	6.17
17	1.75	1.58	0.63	1.09	0.59	7.54
18	1.72	1.43	0.64	1.02	0.63	6.36
19	1.68	1.57	0.72	0.96	0.68	7.63
20	1.75	1.59	0.68	1.08	0.62	7.78
21	2.19	1.86	0.75	1.24	0.72	10.15
22	1.73	1.67	0.64	1.14	0.55	6.88

For simplicity we illustrate the multiplicative model by using only RL and W to estimate WT (the remainder of the variables are used later). First we perform the linear regression with the results in Table 8.11 and the residual plot in Fig. 8.4.

The regression appears to fit well and both coefficients are significant, although the p value for RL is only 0.032. However, the residual plot reveals some problems:

- The residuals have a curved pattern: positive at the extremes and negative in the center. This pattern suggests a curved response.
- The residuals are less variable with smaller values of the predicted value and then become increasingly dispersed as values increase. This pattern reveals a

Table 8.11 Linear Regression for Squid Data

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F value	Pr > F
Model	2	206.74216	103.37108	213.89	<.0001
Error	19	9.18259	0.48329		
Corrected Total	21	215.92475			

Root MSE	0.69519	R-Square	0.9575
Dependent Mean	4.19500	Adj R-Sq	0.9530
Coeff Var	16.57196		

Parameter Estimates

Variable	df	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	−6.83495	0.76476	−8.94	<.0001
RL	1	3.27466	1.41606	2.31	0.0321
W	1	13.40078	3.38003	3.96	0.0008

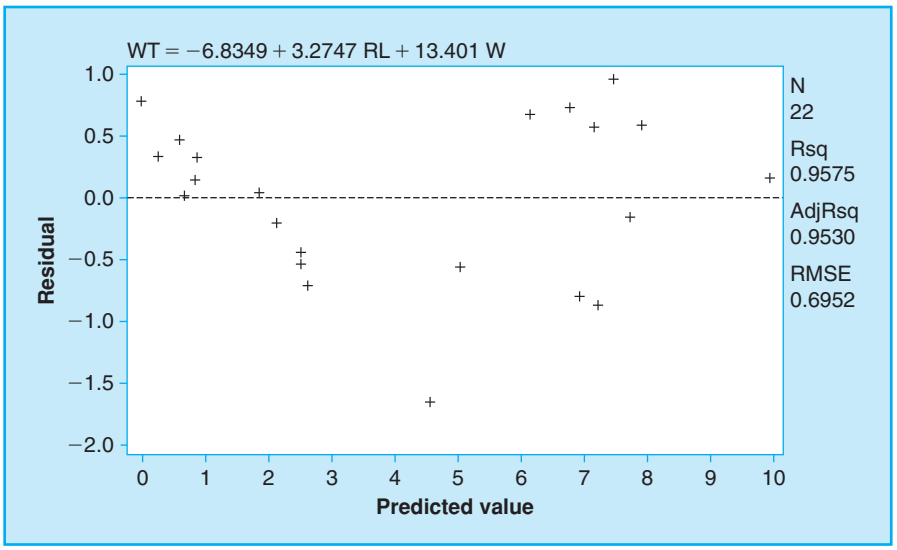

FIGURE 8.4
Residual Plot for Linear Regression.

heteroscedasticity problem of the type discussed in Section 6.4 where we noted that the logarithmic transformation should be used when the standard deviation is proportional to the mean.

The pattern of residuals for the linear regression would appear to suggest that the variability is proportional to the size of the squid. This type of variability is logical for variables related to sizes of biological specimens, which suggests a multiplicative error. In addition, the multiplicative model itself is appropriate for this example. The dependent variable, the weight of squid, is related to volume, which is a *product* of its dimension. For example, the volume of a cube is d^3, where d is the dimension of a side. The basic shape of a squid is in the form of a cylinder for which the volume is $\pi r^2 l$, where r is the radius and l is the length.

To fit the multiplicative model we first create the variables LWT, LW, and LRL to be the logarithms of WT, W, and RL, respectively, and do a linear regression. The results of fitting the two-variable model using logarithms for the squid data are shown in Table 8.12 and the residual plot is shown in Fig. 8.5.

Table 8.12 Multiplicative Model for Squid Data

Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	17.82601	0.91301	400.52	<.0001
Error	19	0.42281	0.02225		
Corrected Total	21	18.24883			
Root MSE		0.14918	R-Square	0.9768	
Dependent Mean		1.07156	Adj R-Sq	0.9744	
Coeff Var		13.92142			

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.16889	0.47827	2.44	0.0245
LRL	1	2.27849	0.49330	4.62	0.0002
LW	1	1.10922	0.37361	2.97	0.0079

This model certainly fits better and both coefficients are highly significant. The multiplicative model is

$$\hat{\mathrm{WT}} = e^{1.169}(\mathrm{RL})^{2.278}(\mathrm{W})^{1.109}.$$

Note that the estimated exponents are close to 2 and unity, which are suggested by the formula for the volume of a cylinder. Finally the residuals appear to have a uniformly random pattern. ■

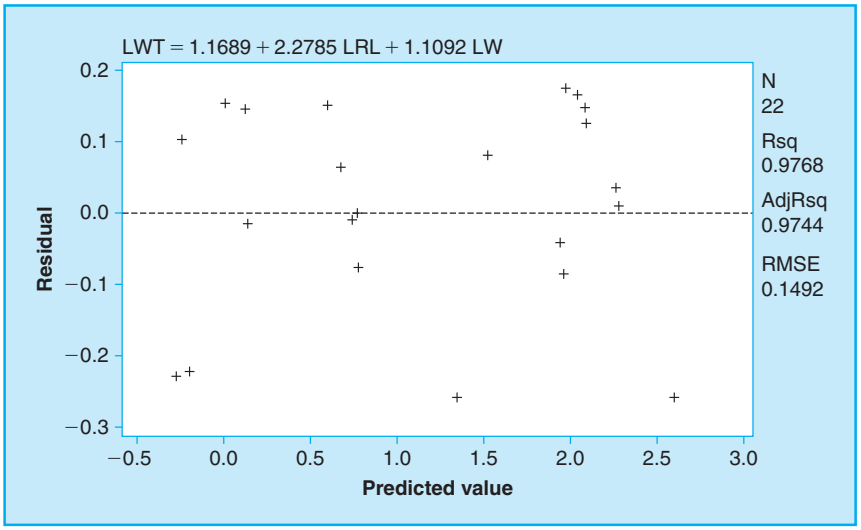

FIGURE 8.5
Residual Plot for Model Using Logarithms.

8.6.3 Nonlinear Models

In some cases no models that are linear in the parameters can be found to provide an adequate description of the data. One such model is the negative exponential model, which is, for example, used to describe the decay of a radioactive substance

$$y = \alpha + \beta e^{\delta t} + \varepsilon,$$

where y is the remaining weight of the substance at time t. According to the model, $(\alpha + \beta)$ is the initial weight when $t = 0, \alpha$ is the ultimate weight of the nondecaying portion of the substance at $t = \infty$, and δ indicates the speed of the decay and is related to the half-life of the substance. Implementation of nonlinear models such as these require specialized methodology introduced in Chapter 13.

8.7 MULTICOLLINEARITY

Often in a multiple regression model, several of the independent variables are measures of similar phenomena. This can result in a high degree of correlation among the set of independent variables. This condition is known as **multicollinearity**. For example, a model used to estimate the total biomass of a plant may include independent variables such as the height, stem diameter, root depth, number of branches, density of canopy, and aerial coverage. Many of these measures are related to the overall size of the plant. All tend to have larger values for larger plants and smaller values for smaller plants and will therefore tend to be highly correlated.

Naively, we might hope that we could create a large number of independent variables, including products and polynomial terms, then use the computing power

CASE STUDY 8.1

Simple plots of predicted values are important tools in understanding the results, particularly when one of the independent variables is an interaction; that is, a product of other independent variables. Consider a study by Robinson *et al.* (2008) where the dependent variable is $y =$ the number of knocks a subject makes on a door when requesting admittance. (The subject does not know this is being measured.) Each subject previously had been scored for Extraversion and Neuroticism. Since the scales of these two variables are quite arbitrary, these were converted to z-scores to form the independent variables $z_1 =$ Extraversion and $z_2 =$ Neuroticism. The fitted regression equation was approximately

$$\hat{y} = 3.95 + .09z_1 - .18z_2 - .36z_1z_2$$

The regression coefficients for z_1 and z_2 did not differ significantly from zero, but this does not mean that these variables are not related to y. Since the coefficient corresponding to z_1z_2 did differ significantly from zero, we know that the relation of y to z_1 differs according to the value of z_2.

A simple plot can show this. Somewhat low and somewhat high values of each independent variable would correspond to -1 and $+1$, assuming a roughly normal distribution. We can plot the four fitted values corresponding to each combination of $-1/+1$ by inserting these values into the equation, resulting in Figure 8.6, where the plotting symbol is the value of Extraversion (Low or High). Based on the significant interaction and the plot, the authors reasonably conclude that being "high in one trait and low in the other is associated with more assertive behavior."

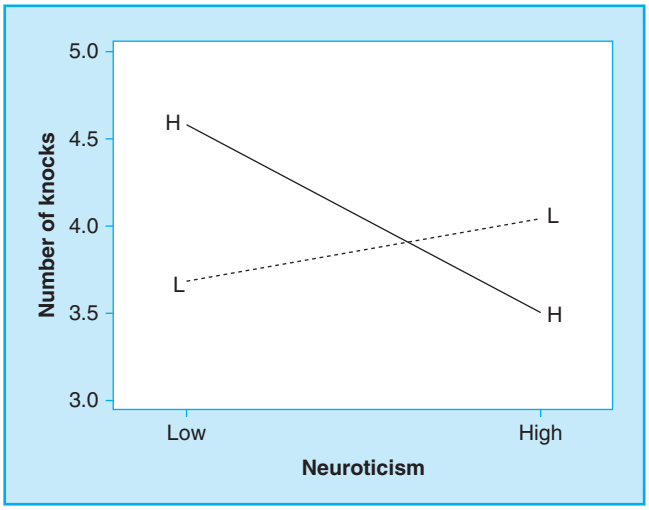

FIGURE 8.6
Fitted Number of Knocks, Applying Values from Case Study 8.1.

of automated software to find the most relevant variables. Sophisticated variable selection routines, such as those discussed in Section 8.8, will certainly attempt this task. Unfortunately, the presence of multicollinearity causes this process to yield ambiguous results. In essence, when independent variables are closely related, relevance cannot be clearly assigned to one variable and not another.[13]

[13] In a polynomial regression (Section 8.6), the powers of x are often highly correlated. Technically, this also leads to multicollinearity, which in this case does not have the same implications.

Remember that a partial coefficient is the change in the dependent variable associated with the change in one of the independent variables, holding constant all other variables. If several variables are closely related it is, by definition, difficult to vary one while holding the others constant. In such cases the partial coefficient is attempting to estimate a phenomenon not exhibited by the data. In a sense such a model is extrapolating beyond the reach of the data.

This extrapolation is reflected by large variances (hence standard errors) of the estimated regression coefficients and a subsequent reduction in the ability to detect statistically significant partial coefficients. A typical result of a regression analysis of data exhibiting multicollinearity is that the overall model is highly significant (has small p value) while few, if any, of the individual partial coefficients are significant (have large p values).

A number of statistics are available for measuring the degree of multicollinearity in a data set. An obvious set of statistics for this purpose is the pairwise correlations among all the independent variables. Large magnitudes of these correlations certainly do signify the existence of multicollinearity; however, the lack of large-valued correlations does not guarantee the absence of multicollinearity and for this reason these correlations are not often used to detect multicollinearity.

A very useful set of statistics for detecting multicollinearity is the set of **variance inflation factors (VIF)**, which indicate, for each independent variable, how much larger the variance of the estimated coefficient is than it would be if the variable were uncorrelated with the other independent variables. Specifically, the VIF for a given independent variable, say, x_j, is $1/(1 - R_j^2)$, where R_j^2 is the coefficient of determination of the regression of x_j on all other independent variables. If R_j^2 is zero, the VIF value is unity and the variable x_j is not involved in any multicollinearity. Any nonzero value of R_j^2 causes the VIF value to exceed unity and indicates the existence of some degree of multicollinearity. For example, if the coefficient of determination for the regression of x_j on all other variables is 0.9, the variance inflation factor will be 10.

There is no universally accepted criterion for establishing the magnitude of a VIF value necessary to identify serious multicollinearity. It has been proposed that VIF values exceeding 10 serve this purpose. However, in cases where the model R^2 is small, smaller VIF values may create problems and vice versa. Finally, if any R_j^2 is 1, indicating an exact linear relationship, VIF $= \infty$, which indicates that $\mathbf{X}'\mathbf{X}$ is singular and thus there is no unique estimate of the regression coefficients.

■ Example 8.8: Example 8.5 Revisited

We illustrate multicollinearity with the squid data, using the logarithms of all variables. Because all of these variables are measures of size, they are naturally

correlated, suggesting that multicollinearity may be a problem. Figure 8.7 shows the matrix of pairwise scatterplots among the logarithms of the variables. Obviously all variables are highly correlated, and in fact, the correlations with the dependent variable appear no stronger than those among the independent variables. Obviously multicollinearity is a problem with this data set.

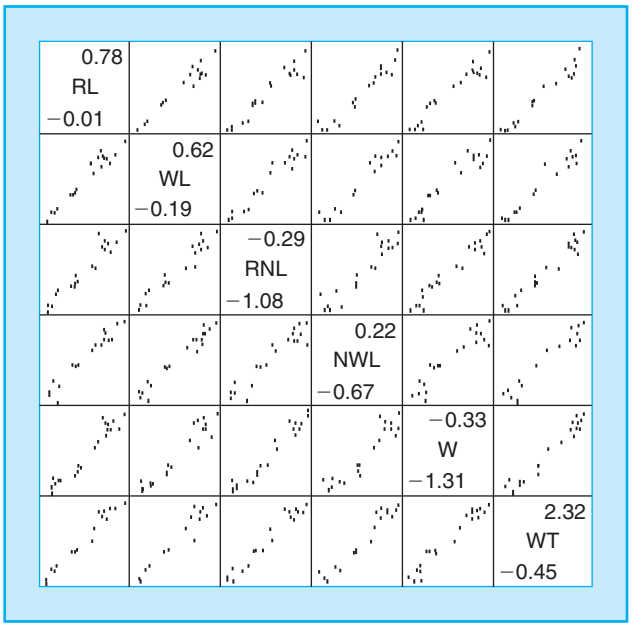

FIGURE 8.7
Scatterplots among Variables in Example 8.5.

We request `PROC REG` of the SAS System to compute the logarithm-based regression using all beak measurements, adding the option for obtaining the variance inflation factors. The results of the regression are shown in Table 8.13. The results are typical of a regression where multicollinearity exists. The test for the model gives a p value of less than 0.0001, while none of the partial coefficients has a p value of less than 0.05. Also, one of the partial coefficient estimates is negative, which is certainly an unexpected result. The variance inflation factors, in the column labeled `VARIANCE INFLATION`, are all in excess of 20 and thus exceed the proposed criterion of 10. The variance inflation factor for the intercept is by definition zero. ■

The course of action to be taken when multicollinearity is found depends on the purpose of the analysis. The presence of multicollinearity is not a violation of assumptions and therefore does not, in general, inhibit our ability to obtain a good

Table 8.13 Regression for Squid Data

DEP VARIABLE: WT

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB > F
MODEL	5	17.927662	3.585532	178.627	0.0001
ERROR	16	0.321163	0.020073		
C TOTAL	21	18.248825			

ROOT MSE	0.141678	R-SQUARE	0.9824
DEP MEAN	1.071556	ADJR-SQ	0.9769
C.V.	13.22173		

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR HO: PARAMETER = 0	PROB > |T|	VARIANCE INFLATION
INTERCEP	1	2.401917	0.727617	3.301	0.0045	0.000000
RL	1	1.192555	0.818469	1.457	0.1644	43.202506
WL	1	−0.769314	0.790315	−0.973	0.3448	45.184233
RNL	1	1.035553	0.666790	1.553	0.1400	31.309370
NWL	1	1.073729	0.582517	1.843	0.0839	27.486102
W	1	0.843984	0.439783	1.919	0.0730	21.744851

fit for the model. This can be seen in the above example by the large R-square value and the small residual mean square. Furthermore, the presence of multicollinearity does not affect the inferences about the mean response or prediction of new observations as long as these inferences are made within the range of the observed data. Thus, if the purpose of the analysis is to estimate or predict, then one or more of the independent variables may be dropped from the analysis, using the procedures presented in Section 8.8, to obtain a more efficient model. The purpose of the analysis of the squid data has this objective in mind, and therefore the equation shown in Table 8.10 or the equation resulting from variable selection (Table 8.12) could be effectively used, although care must be taken to avoid any hint of extrapolation.

On the other hand, if the purpose of the analysis is to determine the effect of the various independent variables, then a procedure that simply discards variables is not effective. After all, an important variable may have been discarded because of multicollinearity.

8.7.1 Redefining Variables

One procedure for counteracting the effects of multicollinearity is to redefine some of the independent variables. This procedure is commonly applied in the analysis

of national economic statistics collected over time, where variables such as income, employment, savings, etc., are affected by inflation and increases in population and are therefore correlated. Deflating these variables by a price index and converting them to a per capita basis greatly reduces the multicollinearity.

Example 8.9 : Example 8.5 Revisited

In the squid data, all measurements are related to overall size of the beak. It may be useful to retain one measurement of size, say, `W`, and express the rest as ratios to `W`. The resulting ratios may then measure shape characteristics and exhibit less multicollinearity. Since the variables used in the regression are logarithms, the logarithms of the ratios are differences. For example, $\log(\texttt{RL}/\texttt{W}) = \log(\texttt{RL}) - \log(\texttt{W})$. Using these redefinitions and keeping $\log(\texttt{W})$ as is, we obtain the results shown in Table 8.14.

Solution

A somewhat unexpected result is that the overall model statistics—the F test for the model, R^2, and the mean square error—have not changed. This is because a linear regression model is not really changed by a linear transformation that retains the same number of variables, as demonstrated by the following simple example. Assume a two-variable regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon.$$

Define $x_3 = x_1 - x_2$, and use the model

$$y = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_3 + \varepsilon.$$

In terms of the original variables, this model is

$$y = \gamma_0 + (\gamma_1 + \gamma_2) x_1 - \gamma_2 x_2 + \varepsilon,$$

which is effectively the same model where $\beta_1 = (\gamma_1 + \gamma_2)$ and $\beta_1 = -\gamma_2$.

In the new model for the squid data, we see that the overall width variable (`W`) clearly stands out as the main contributor to the prediction of weight, and the degree of multicollinearity has been decreased. At the bottom is a test of the hypothesis that all other variables contribute nothing to the regression involving `W`. This test shows that hypothesis to be rejected, indicating the need for at least one of these other variables, although none of the individual coefficients in this set are significant (all p values > 0.05). Variable selection (Section 8.8) may be useful for determining which additional variable(s) may be needed. ■

Table 8.14 Regression with Redefined Variables

Model: MODEL 1
Dependent Variable: WT

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob > F
Model	5	17.92766	3.58553	178.627	0.0001
Error	16	0.32116	0.02007		
C Total	21	18.24883			

Root MSE	0.14168	R-square	0.9824
Dep Mean	1.07156	Adj R-sq	0.9769
C.V.	13.22173		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter = 0	Prob > |T|
INTERCEP	1	2.401917	0.72761686	3.301	0.0045
RL	1	1.192555	0.81846940	1.457	0.1644
WL	1	−0.769314	0.79031542	−0.973	0.3448
RNL	1	1.035553	0.66679027	1.553	0.1400
NWL	1	1.073729	0.58251746	1.843	0.0839
W	1	3.376507	0.17920582	18.842	0.0001

Variable	DF	Variance Inflation
INTERCEP	1	0.00000000
RL	1	8.53690485
WL	1	7.15487734
RNL	1	4.35395220
NWL	1	4.94314166
W	1	3.61063657

Dependent Variable: WT

Test: ALLOTHER	Numerator:	0.1441	df:	4	F value:	7.1790
	Denominator:	0.020073	df:	16	Prob > F:	0.0016

8.7.2 Other Methods

Another approach is to perform multivariate analyses such as principal components or factor analysis on the set of independent variables to obtain ideas on the nature of the multicollinearity. These methods are beyond the scope of this book (see Freund *et al.*, 2006, Section 5.4).

An entirely different approach is to modify the method of least squares to allow biased estimators of the regression coefficients. Some biased estimators effectively reduce the effect of multicollinearity so that, although the estimates are biased, they have a much smaller variance and therefore have a larger probability of being close to the true parameter value. One such biased regression procedure is called ridge regression (see Freund *et al.*, 2006, Section 5.4).

8.8 VARIABLE SELECTION

One of the benefits of modern computers is the ability to handle large data sets with many variables. One objective of many experiments is to "filter" these variables to identify those that are most important in explaining a process. In many applications this translates into obtaining a good regression using a minimum number of independent variables. Although the search for this set of variables should use knowledge about the process and its variables, the power of the computer may be useful in implementing a data-driven search for a subset of independent variables that provides adequately precise estimation with a minimum number of variables, which may incidentally provide for less multicollinearity than the full set.

Finding such a model may be accomplished by means of one of a number of **variable selection** techniques. Unfortunately, variable selection is not the panacea it is sometimes ascribed to be. Rather, variable selection is a sort of data dredging that may provide results of spurious validity. Furthermore, if the purpose of the regression analysis is to establish the partial regression relationships, discarding variables may be self-defeating. In other words, variable selection is not always appropriate for the following reasons:

1. It does not help to determine the structure of the relationship among the variables.
2. It uses the power of the computer as a substitute for intelligent study of the problem.
3. The decisions on whether to keep or drop an independent variable from the model are based on the test statistics of the estimated coefficients. Such a procedure is generating hypotheses based on the data, which we have already indicated plays havoc with the specified significance levels. Therefore, just as it is preferable to use preplanned contrasts to automatic post hoc comparisons in the analysis of variance, it is preferable to use knowledge-based selection instead of automatic data-driven selection in regression.

However, despite all these shortcomings, variable selection is widely used, primarily because computers have made it so easy to do. Often there seems to be no reasonable alternative and it actually can produce useful results. For these reasons we present here some variable selection methods together with some aids that may be useful in selecting a useful model.

The purpose of variable selection is to find that subset of the variables in the original model that will in some sense be "optimum." There are two interrelated factors in determining that optimum:

1. For any given subset size (number of variables in the model) we want the subset of independent variables that provides the minimum residual sum of squares. Such a model is considered "optimum" for that subset size.
2. Given a set of such optimum models, select the most appropriate subset size.

One aspect of this problem is that to **guarantee** optimum subsets, all possible subsets must be examined. Hypothetically this method requires that the error sum of squares be computed for 2^m subsets! For example, if $m = 10$, there will be 1024 subsets; for $m = 20$, there will be 1,048,576 subsets!

Modern computers and highly efficient computational algorithms allow some shortcuts, so this problem is not as insurmountable as it may seem. Thus, for example, using the SAS System, the guaranteed optimum subset method can be used for models containing as many as 30 variables. Useful alternatives for models that exceed available computing power are discussed at the end of this selection.

We illustrate the guaranteed optimum subset method with the squid data using the logarithms of the original variables. The program used is `PROC REG` from the SAS System, implementing the `RSQUARE` selection option. The results are given in Table 8.15.

This procedure has examined 31 subsets (not including the null subset), but we have requested that it print results for only the best five for each subset size, which are listed in order from best (optimum) to fifth best. Although we focus on the optimum subsets, the others may be useful, for example, if the second best is almost optimum and contains variables that cost less to measure. For each of these subsets, the procedure prints the R^2 values, the $C(p)$ statistic that is discussed below, and the listing of variables in each selected model.

There are no truly objective criteria for choosing subset size. Statistical significance tests are inappropriate since we generate hypotheses from data. The usual procedure is to plot the behavior of some goodness-of-fit statistic against the number of variables and choose the minimum subset size before the statistic indicates a deterioration of the fit. Virtually any statistic such as MSE or R^2 can be used, but the most popular one currently in use is the **$C(p)$ statistic**.

The $C(p)$ statistic, proposed by Mallows (1973), is a measure of total squared error for a model containing $p(<m)$ independent variables. This total squared error is a measure of the error variance plus a bias due to an underspecified model, that is, a model that excludes variables that should be in the "true" model. Thus, if $C(p)$ is "large" then there is bias due to an underspecified model. The formula for $C(p)$ is of little interest but it is structured so that for a p-variable model:

Table 8.15 Variable Selection for Squid Data

Dependent Variable: WT
R-Square Selection Method

Number in Model	R-Square	C(p)	Variables in Model
1	0.9661	12.8361	RL
1	0.9517	25.8810	RNL
1	0.9508	26.7172	W
1	0.9461	30.9861	WL
1	0.9399	36.6412	NWL
2	0.9768	5.0644	RL W
2	0.9763	5.5689	NWL W
2	0.9752	6.5661	RL RNL
2	0.9732	8.3275	RNL NWL
2	0.9682	12.9191	RL NWL
3	0.9797	4.4910	RL NWL W
3	0.9796	4.5603	RNL NWL W
3	0.9786	5.4125	RL RNL W
3	0.9775	6.4971	RL RNL NWL
3	0.9770	6.8654	RL WL W
4	0.9814	4.9478	RL RNL NWL W
4	0.9801	6.1232	WL RNL NWL W
4	0.9797	6.4120	RL WL NWL W
4	0.9787	7.3979	RL WL RNL W
4	0.9783	7.6831	RL WL RNL NWL
5	0.9824	6.0000	RL WL RNL NWL W

- if $C(p) > (p + 1)$, the model is underspecified, and
- if $C(p) < (p + 1)$, the model is overspecified; that is, it most likely contains unneeded variables.

By definition, when $p = m$ (the full model), $C(p) = m + 1$. The plot of $C(p)$ values for the variable selections in Table 8.15 is shown in Fig. 8.8; the line plots $C(p)$ against $(p + 1)$, which is the boundary between over- and underspecified models.

The $C(p)$ plot shows that the four-variable model is slightly overspecified, the three-variable model is slightly underspecified, and the two-variable model is underspecified (the $C(p)$ values for the one-variable model are off the scale). The choice would seem to be the three-variable model. However, note that there are two almost identically fitting "optimum" three-variable models, suggesting that there is still too much multicollinearity. Thus the two-variable model would appear to be a better choice, which is the one used to illustrate the multiplicative model (Table 8.12 and

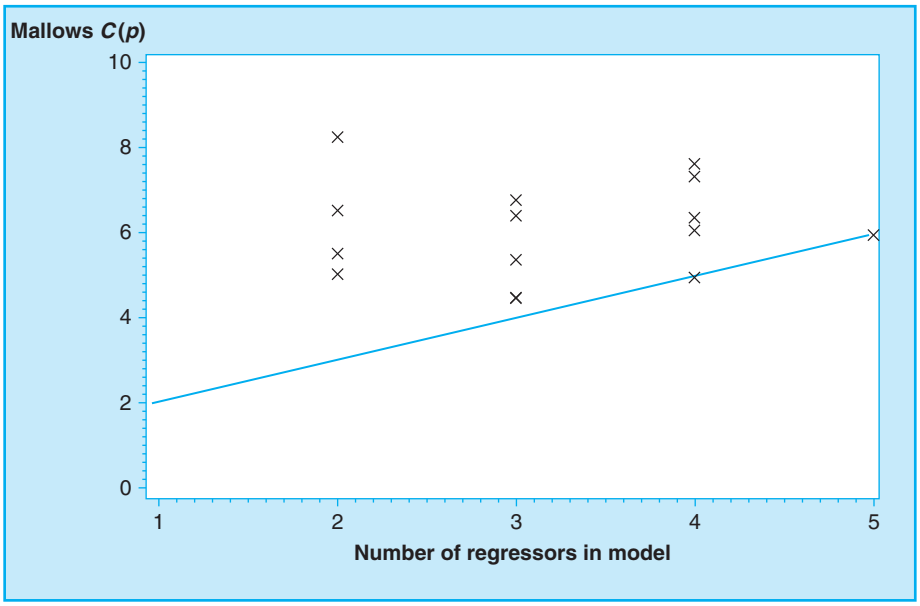

FIGURE 8.8
$C(p)$ Plot for Variable Selection.

Fig. 8.6). This decision is, of course, somewhat subjective and the researcher can examine the two competing three-variable models and use the one which makes the most sense relative to the problem being addressed.

8.8.1 Other Selection Procedures

We have noted that the guaranteed optimum subset method can be quite expensive to perform. For this reason several alternative procedures that provide nearly optimum models by combining the two aspects of variable selection into a single process exist. Actually these procedures do provide optimum subsets in many cases, but it is not possible to know whether this has actually occurred.

These alternative procedures are also useful as screening devices for models with many independent variables. For example, applying one of these for a 30-variable case may indicate that only about 5 or 6 variables are needed. It is then quite feasible to perform the guaranteed optimum subset method for subsets of size 5 or 6.

The most frequently used alternative methods for variable selection are as follows:

1. *Backward elimination*: Starting with the full model, delete the variable whose coefficient has the smallest partial sum of squares (or smallest magnitude t statistic). Repeat with the resulting $(m-1)$ variable equation, and so forth. Stop deleting variables when all variables contribute some specified minimum partial sum of squares (or have some minimum magnitude t statistic).

2. *Forward selection*: Start by selecting the variable that, by itself, provides the best-fitting equation. Add the second variable whose additional contribution to the regression sum of squares is the largest, and so forth. Continue to add variables, one at a time, until any variable when added to the model contributes less than some specified amount to the regression sum of squares.
3. *Stepwise*: This is an adaptation of forward selection in which, each time a variable has been added, the resulting model is examined to see whether any variable included makes a sufficiently small contribution so that it can be dropped (as in backward elimination).

None of these methods is demonstrably superior for all applications and do not, of course, provide the power of the "all possible" search method.

Although the step methods are usually not recommended for problems with a small number of variables, we illustrate the forward selection method with the transformed squid data, using the forward selection procedure in SPSS Windows. The output is shown in Table 8.16.

The first box in the output summarizes the forward selection procedure. It indicates that two "steps" occurred resulting in two models. The first contained only the variable RL. The second model added W. The box also specifies the method and the criteria used for each step. The next box contains the Model Summary for each model. This box indicates that the R Square for model 1 had a value of 0.966 and that adding the variable W increased the R Square only to 0.977.

The third box contains the ANOVA results for both models. Both are significant with a p value (labeled Sig.) listed as .000, which is certainly less than 0.05.[14] The next box lists the coefficients for the two regression models and the t test for them. Notice that the values of the coefficients for model 2 are the same as those in Table 8.12.

The final box lists the variables excluded from each model and some additional information about these variables. This table displays information about the variables not in the model at each step. Beta in is the standardized regression coefficient that would result if the variable were entered into the equation at the next step. For example, if we used the model that only contained RL, the variable RNL would result in a regression that had a coefficient for RNL with a value of 0.382 resulting in a p value of 0.016. However, the forward procedure dictated that a better two-variable model would be RL and W. Then when RNL was considered for bringing into the model, it would have a coefficient of 0.211 but the p value would be 0.232.

The last box also includes the partial correlation coefficients (with WT), and something called the "tolerance," which is the reciprocal of the VIF. If the criteria for the VIF is anything larger than 10 then the criteria for the tolerance would be anything less than 0.10.

[14] Remember that this is not a "true" significance level!

Table 8.16 Results of Forward Selection

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	RL		Forward (Criterion: Probability-of-F-to-enter <= .050)
2	W		Forward (Criterion: Probability-of-F-to-enter <= .050)

a. Dependent Variable: WT

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.983[a]	.966	.964	.17592
2	.988[b]	.977	.974	.14918

a. Predictors: (Constant), RL
b. Predictors: (Constant), RL, W

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	17.630	1	17.630	569.664	.000[a]
	Residual	.619	20	.031		
	Total	18.249	21			
2	Regression	17.826	2	8.913	400.524	.000[b]
	Residual	.423	19	.022		
	Total	18.249	21			

a. Predictors: (Constant), RL
b. Predictors: (Constant), RL, W
c. Dependent Variable: WT

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.241	.067		-3.622	.002
	RL	3.690	.155	.983	23.868	.000
2	(Constant)	1.169	.478		2.444	.024
	RL	2.279	.493	.607	4.619	.000
	W	1.109	.374	.390	2.969	.008

a. Dependent Variable: WT

Excluded Variables[c]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	WL	.230[a]	1.099	.285	.245	3.835E-02
	RNL	.382[a]	2.639	.016	.518	6.237E-02
	NWL	.212[a]	1.122	.276	.249	4.698E-02
	W	.390[a]	2.969	.008	.563	7.064E-02
2	WL	.079[b]	.414	.684	.097	3.493E-02
	RNL	.211[b]	1.238	.232	.280	4.087E-02
	NWL	.246[b]	1.583	.131	.350	4.675E-02

a. Predictors in the Model: (Constant), RL
b. Predictors in the Model: (Constant), RL, W
c. Dependent Variable: WT

The forward selection procedure resulted in two "steps" and terminated with a model that contained the variables RL and W. This is, of course, consistent with previous analyses. Normally two different variable selection procedures will result in the same conclusion, but not always, particularly if there is a great deal of multicollinearity present.

In conclusion we emphasize again that variable selection, although very widely used, should be employed with caution. There is no substitute for intelligent, nondata-based variable choices.

8.9 DETECTION OF OUTLIERS, ROW DIAGNOSTICS

We have repeatedly emphasized that failures of assumptions about the nature of the data may invalidate statistical inferences. For this reason we have encouraged the use of exploratory data analysis of observed or residual values to aid in the detection of failures in assumptions and the use of alternate methods if such failures are found.

As data and models become more complex, opportunities increase for undetected violations and inappropriate analyses. For example, in regression analysis the misspecification of the model, such as leaving out important independent variables or neglecting the possibility of curvilinear responses, may lead to estimates of parameters exhibiting large variances. The fact that data for regression analysis are usually observed, rather than the result of carefully designed experiments, makes the existence of misspecification, violation of assumptions, and inappropriate analysis more difficult to detect.

For these types of data it is also more difficult to detect outliers. We first discuss the basic reason for this and subsequently present some methodologies that may aid in overcoming the problem.

A Physical Analogue to Least Squares

A fundamental law of physics, called Hooke's law, specifies that the tension of a coil spring is proportional to the square of the length that the spring has been stretched (assuming a perfect spring). The least squares estimate of a one-variable regression line is equivalent to hooking a set of springs, perpendicular to the x axis, from the data points to a rigid rod. The equilibrium position of the rod represents the minimum total tension of the springs and thus represents the least squares line (assuming no gravity). This is illustrated in Fig. 8.9.

This analogue is useful for illustrating a number of characteristics of least squares estimation. For example, the amount of force required to pull the rod into a horizontal position ($\beta_1 = 0$) represents the strength or statistical significance of the linear regression of y on x. Remember the estimated variance of β_1 is $(s^2_{y|x}/S_{xx})$, which increases in magnitude as the x values span a narrower range (Section 7.5). Similarly, the force required to pull the rod into the horizontal position is lower if the data values

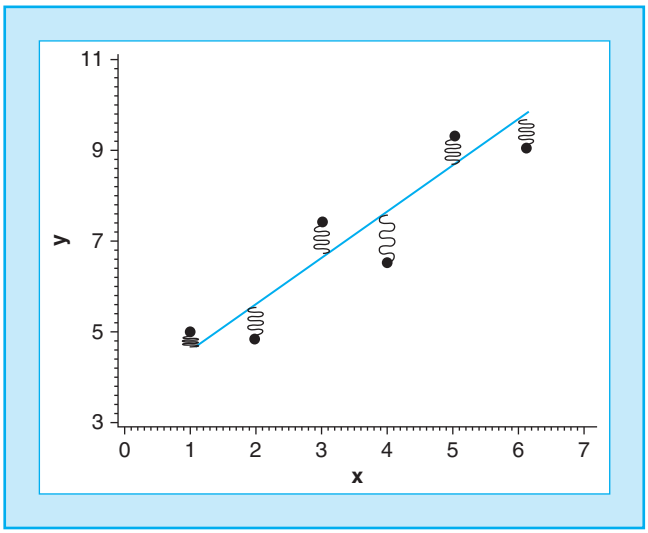

FIGURE 8.9
Illustration of Hooke's Law.

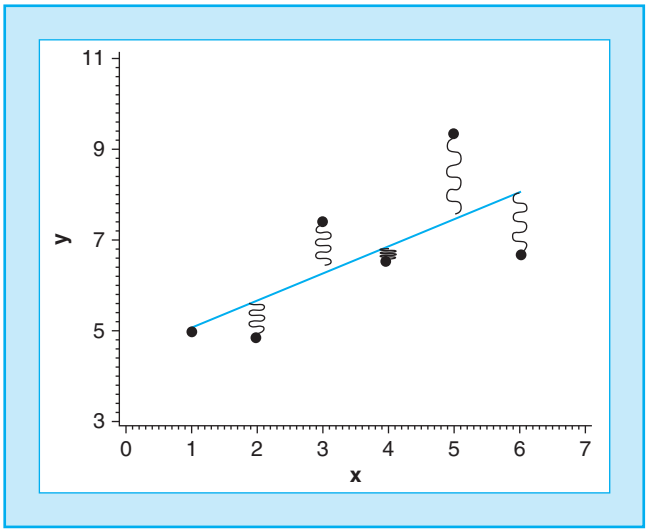

FIGURE 8.10
Effect of an Outlier.

occupy a narrow range in the x direction when the springs are close to the center of the rod.

The spring analogue also illustrates the effect of the location of individual observations on the estimated coefficients. For example, an unusual or extreme value for the dependent variable y will tend to exert a relatively large influence or **leverage** on the equilibrium location of the regression line as illustrated in Fig. 8.10, where the data are identical to those Fig. 8.9 except that the response for $x = 6$ has been decreased by 3 units, making this an outlier. In this case the outlier occurs at the extreme of

the range of the x values; hence, the point exerts extreme *leverage* so that the line is forced to pass quite close to that point. Hence the largest observed residual is actually for $x = 5$, which is not an outlier. On the other hand an outlier at the center of the range of x values will not exert such a large leverage on the location of the line. However, the outlier may create a large residual, which when squared, contributes to an overestimate of the variance.

This example shows that the effect of outliers in the response variable depends on where the observation lies in the space of the independent variable(s). This effect is relatively easy to visualize in the case of a simple linear regression but is, obviously, more difficult to "see" when there are many independent variables. While outlier detection statistics tend to focus on outliers in the dependent variable, other statistics focus on outliers in the independent variable, which we have identified as observations having a high degree of leverage. Yet other statistics provide information on both of these aspects. While examining a large number of such statistics can be quite useful, the scope of this book limits our presentation to one of the most frequently used combination statistics. A more complete discussion can be found in Belsley *et al.* (1980).

One important class of statistics that investigate the combined effects of outliers and leverage is known as **influence statistics**. These statistics are based on the question: "What happens if the regression is estimated using the data without a particular observation?" We present one such influence statistic and give an example of how it may be useful. The statistic, known as the DFFITS statistic, is the difference between the predicted value for each observation using the model estimated with all data and that using the model estimated with that observation omitted (Belsley *et al.*, 1980). This difference is standardized, using the residual variance as estimated with the observation omitted. Large values of this statistic may indicate suspicious observations. Generally, values exceeding $2\sqrt{(m+1)/n}$ are considered large for this purpose.[15] Actually this criterion is not often needed since outliers having serious effects on model estimates usually have DFFITS values greatly exceeding this criterion.

The DFFITS statistics are closely related to the **studentized deleted residuals**, also called the **jackknifed residuals**. For each observation, we calculate the difference between the actual observation and the fitted value from a regression that drops that observation from the data set. (This deletion process is called **jackknifing**.) If the observation is an outlier, it will not have a chance to corrupt the parameter estimates, and so its deleted residual will stand out as abnormally large. These residuals are then studentized; that is, divided by an estimate of their standard error.

[15] Fortunately, it is not necessary to recompute the regression equation omitting each observation in turn. Special algorithms are available that make these computations quite feasible even for rather large problems. We also emphasize that other outlier detection statistics are available and that the DFFITS statistic is not necessarily the best. However, this statistic is quite popular, and to present other statistics at this point may confuse the issue.

This puts them on a familiar t distribution scale. As a quick rule-of-thumb, absolute values for jackknifed residuals above 3.0 are suspicious. Jackknifed residuals are superior to ordinary residuals for detecting outliers. However they are not infallible—they will fail to catch multiple outliers, especially when these are located close together.

Example 8.10

The production levels of a finished product (produced from sheets of stainless steel) have varied quite a bit, and management is trying to devise a method for predicting the daily amount of finished product. The ability to predict production is useful for scheduling labor, warehouse space, and shipment of raw materials and also to suggest a pricing strategy.

The number of units of the product (Y) that can be produced in a day depends on the width (X1) and the density (X2) of the sheets being processed, and the tensile strength of the steel (X3). The data are taken from 20 days of production. The observations are given in Table 8.17.

Table 8.17 Data for Outlier Detection

OBS	Y	X1	X2	X3
1	763	19.8	128	86
2	650	20.9	110	72
3	554	15.1	95	62
4	742	19.8	123	82
5	470	21.4	77	52
6	651	19.5	107	72
7	756	25.2	123	84
8	563	26.2	95	83
9	681	26.8	116	76
10	579	28.8	100	64
11	716	22.0	110	80
12	650	24.2	107	71
13	761	24.9	125	81
14	549	25.6	89	61
15	641	24.7	103	71
16	606	26.2	103	67
17	696	21.0	110	77
18	795	29.4	133	83
19	582	21.6	96	65
20	559	20.0	91	62

Solution

We perform a linear regression of Y on X1, X2, and X3, using PROC REG of the SAS System. The analysis, including the residuals and DFFITS statistics, is shown in Table 8.18. The results appear to be quite reasonable. The regression is certainly

Table 8.18 Analysis of Steel Data

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB > F
MODEL	3	146684.105	48894.702	133.750	0.0001
ERROR	16	5849.095	365.568		
C TOTAL	19	152533.200			
ROOT MSE		19.119844	R-SQUARE	0.9617	
DEP MEAN		648.200	ADJ R-SQ	0.9545	
C.V.		2.949683			

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER = 0	PROB > |T|	VARIANCE INFLATION
INTERCEP	1	6.383762	40.701546	0.157	0.8773	0.000000
X1	1	−0.916131	1.243010	−0.737	0.4718	1.042464
X2	1	5.409022	0.595196	9.088	0.0001	3.906240
X3	1	1.157731	0.909244	1.273	0.2211	3.896413

OBS	Y	RESIDUALS	DFFITS
1	763	−17.164	−0.596
2	650	−15.586	−0.259
3	554	−24.187	−1.198
4	742	−6.488	−0.175
5	470	6.525	0.263
6	651	0.359	0.007
7	756	10.144	0.218
8	563	−29.330	−12.535
9	681	−16.266	−0.334
10	579	−15.996	−0.592
11	716	42.160	1.138
12	650	4.822	0.064
13	761	7.524	0.167
14	549	14.045	0.380
15	641	17.916	0.261
16	606	−11.078	−0.230
17	696	24.717	0.450
18	795	0.059	0.003
19	582	0.886	0.015
20	559	6.938	0.155

significant. Only one coefficient appears to be important and there is little multicollinearity. Thus one would be inclined to suggest a model that includes only X2 and would probably show increased production with increased values of X2. The residuals, given in the column labeled RESIDUALS, also show no real surprises. The residual for observation 11 appears quite large, but the residual plot (not reproduced here) does not show it as an extreme value. However, the DFFITS statistics show a different story. The value of that statistic for observation 8 is about 10 times that for any other observation. Figure 8.11 (top) shows the plot of the ordinary residuals, and Figure 8.11 (bottom) shows the jackknifed residuals. Clearly, the outlier is easier to detect using the jackknifed residuals. By any criterion this observation is certainly a suspicious candidate.

The finding of a suspicious observation does not, however, suggest what the proper course of action should be. Simply discarding such an observation is usually not recommended. Serious efforts should be made to verify the validity of the data values or to determine whether some unusual event did occur. However, for purposes of illustration here, we do reestimate the regression without that observation. The results of the analysis are given in Table 8.19, where it becomes evident that omitting observation number 8 has greatly changed the results of the regression analysis. The residual variance has decreased from 366 to 106, the F statistic for testing the model has increased from 134 to 448, the estimated coefficients and their p values have changed drastically so that now X3 is the dominant independent variable, and the degree of multicollinearity between X2 and X3 has also increased. In other words, the conclusions about the factors affecting production have changed by eliminating one observation.

The change in the degree of multicollinearity provides a clue to the reasons for the apparent outlier. Figure 8.12 shows the matrix of scatterplots for these variables. The plotting symbol is a period except for observation 8, whose symbol is "8." These plots clearly show that the observed values for X2 and X3 as well as Y and X3 are highly correlated *except* for observation 8. However, that observation appears not to be unusual with respect to the other variables. The conclusion to be reached is that the unusual combination of values X2 and X3 that occurred in observation 8 is a combination that does not conform to the normal operating conditions. Or it could be a recording error. ■

Finding and identifying outliers or influential observations does not answer the question of what to do with such observations. Simply discarding or changing such observations is bad statistical practice since it may lead to self-fulfilling prophesies. Sometimes, when an outlier can be traced to sloppiness or mistakes, deletion or modification may be justified. In the above example, the outlier may have resulted from an unusual product mix that does not often occur. In this case, omission may be justified, but only if the conclusions state that the equation may only be used for the usual product mix and that a close watch must be posted to detect unusual mixes

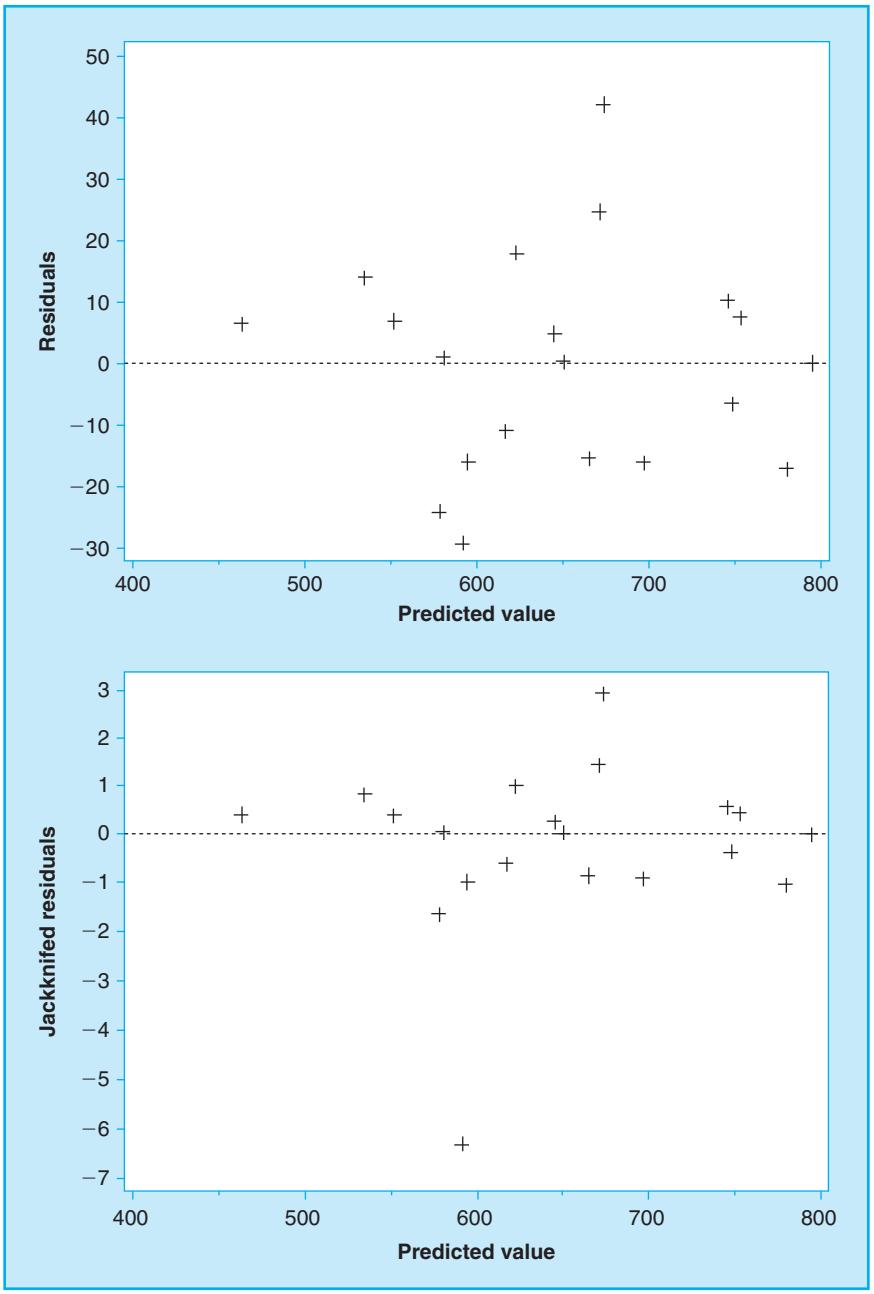

FIGURE 8.11
Residual Plots for Steel Data in Example 8.10.

Table 8.19 Results when Outlier is Omitted

DEP VARIABLE: Y

SOURCE	df	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB > F
MODEL	3	143293.225	47764.408	448.105	0.0001
ERROR	15	1598.880	106.592		
C TOTAL	18	144892.105			

ROOT MSE	10.324340	R-square	0.9890
DEP MEAN	652.684	Adj R-sq	0.9868
C.V.	1.581828		

VARIABLE	df	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER = 0	PROB > |T|	VARIANCE INFLATION
INTERCEP	1	−42.267607	23.289383	−1.815	0.0896	0.000000
X1	1	0.982466	0.735468	1.336	0.2015	1.202123
X2	1	1.738216	0.664253	2.617	0.0194	16.053214
X3	1	6.738637	1.011032	6.665	0.0001	15.420233

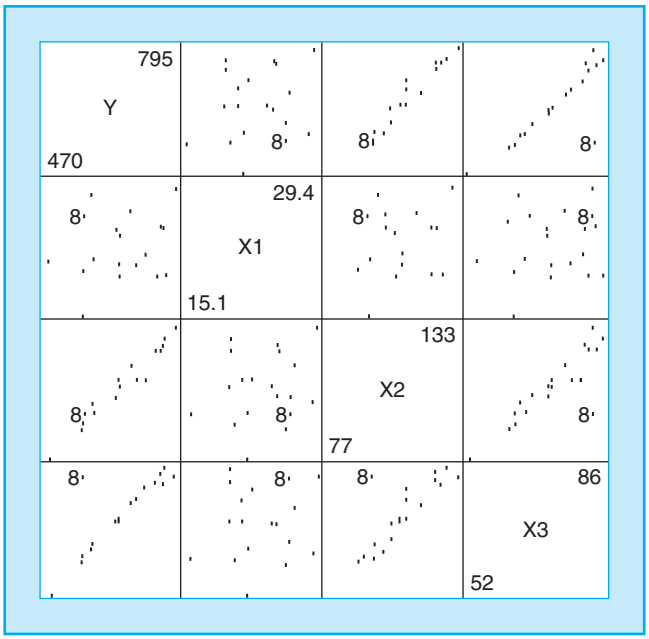

FIGURE 8.12
Scatterplots of Steel Data.

whose costs cannot be predicted by that model. In the previous example, predicting the number of units produced for day 8 without using that day's values provides a predicted value of 702.9, certainly a very bad prediction!

8.10 CHAPTER SUMMARY

Solution to Example 8.1

The effect of performance factors on winning percentages of baseball teams can be studied by a multiple regression using WIN as the dependent variable and the team performance factors as independent variables. Although the data are certainly not random, it is reasonable to assume that the residuals from the model are random and otherwise adhere reasonably to the required assumptions. The output for the regression as produced by PROC REG of the SAS System is shown in Table 8.20.

Table 8.20 Regression for Winning Baseball Games

Model: MODEL1
Dependent Variable: WIN

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob > F
Model	5	0.12324	0.02465	12.410	0.0001
Error	34	0.06753	0.00199		
C Total	39	0.19076			
Root MSE		0.04457	R-square		0.6460
Dep Mean		0.50000	Adj R-sq		0.5940
C.V.		8.91323			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter = 0	Prob > |T|
INTERCEP	1	−0.277675	0.19131508	−1.451	0.1558
RUNS	1	0.000278	0.00014660	1.895	0.0666
BA	1	1.741999	0.92847059	1.876	0.0692
DP	1	0.000737	0.00045021	1.637	0.1108
WALK	1	−0.000590	0.00012916	−4.566	0.0001
SO	1	0.000346	0.00010441	3.315	0.0022

Variable	DF	Variance Inflation
INTERCEP	1	0.00000000
RUNS	1	2.93207465
BA	1	3.05405561
DP	1	1.61208141
WALK	1	1.21916888
SO	1	1.46815334

Starting at the top, it is evident that the regression is certainly significant, although the coefficient of determination may not be considered particularly large. The residual standard deviation of 0.045 indicates that about 95% of observed proportion of wins are within 0.09 of the predicted values, which indicates that there are obviously some other factors affecting the winning percentages. The coefficients all have the expected signs, but it appears that the only important factors relate to pitching. The variance inflation factors are relatively small, although there appears to be an expected degree of correlation between number of runs and batting average.

It is interesting to investigate the relative importance of the offensive (`RUNS, BA`) and defensive (`DP, WALK, SO`) factors. These questions can be answered with this computer program by the so-called `TEST` commands. The first test, labeled `OFFENSE`, tests the hypothesis that the coefficients for `RUNS` and `BA` are both zero, and the second, labeled `DEFENSE`, tests the null hypothesis that the coefficients of `DP`, `WALK`, and `SO` are all zero. These commands produce the following results:

```
Test: OFFENSE  Numerator:    0.0304    DF:   2  F value:   15.3263
               Denominator:  0.001986  DF:  34  Prob > F:   0.0001
Test: DEFENSE  Numerator:    0.0226    DF:   3  F value:   11.3990
               Denominator:  0.001986  DF:  34  Prob > F:   0.0001
```

It appears that both offense and defense contribute to winning, but offense may be more important. This conclusion is not quite consistent with the tests on individual coefficients, a result that may be due to the existence of some correlation among the variables.

Since a number of the individual factors appear to have little effect on the winning percentage, variable selection may be useful. The `RSQUARE` selection of `PROC REG` provides the results shown in Table 8.21. The selection process indicates little loss in the mean square error associated with dropping double plays and runs; hence the remaining three variables may provide a good model. The resulting regression is summarized in Table 8.22.

The model with the three remaining variables fits almost as well as the one with all five variables, and now the effects of the performance factors are more definitive. Additional analysis includes the residual plot, which is shown in Fig. 8.13. Although one team has a rather large negative residual, the overall pattern of residuals shows no major cause for concern about assumptions. ■

The multiple linear regression model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \varepsilon$$

is the extension of the simple linear regression model to more than one independent variable. The basic principles of a multiple regression analysis are the same as for the simple case, but many of the details are different.

Table 8.21 Variable Selection For Baseball Regression

The REG Procedure
Model: MODEL1
Dependent Variable: WIN
R-Square Selection Method

Number in Model	R-Square	C(p)	Variables in Model
1	0.2625	34.8352	BA
1	0.2606	35.0174	RUNS
1	0.1793	42.8268	SO
1	0.0691	53.4079	WALK
2	0.4829	15.6621	BA WALK
2	0.4769	16.2464	RUNS WALK
2	0.4069	22.9662	BA SO
2	0.3882	24.7608	RUNS SO
3	0.5856	7.8051	BA WALK SO
3	0.5612	10.1473	RUNS WALK SO
3	0.5313	13.0186	RUNS BA WALK
3	0.4852	17.4423	BA DP WALK
4	0.6181	6.6800	RUNS BA WALK SO
4	0.6094	7.5201	RUNS DP WALK SO
4	0.6086	7.5919	BA DP WALK SO
4	0.5316	14.9882	RUNS BA DP WALK
5	0.6460	6.0000	RUNS BA DP WALK SO

The least squares principle for obtaining estimates of the regression coefficients requires the solution of a set of linear equations that can be represented symbolically by matrices and is solved numerically, usually by computers.

As in simple linear regression, the variance of the random error is based on the sum of squares of residuals and is computed through a partitioning of sums of squares.

Because the partial regression coefficients in a multiple regression model measure the effect of a variable in the presence of all other variables in the model, estimates and inferences for these coefficients are different from the total regression coefficients obtained by the corresponding simple linear regressions. Inference procedures for the partial regression coefficients are therefore based on the comparison of the full model, which includes all coefficients and the restricted model, with the restrictions relating to the inference on specific coefficients.

Inferences for the response have the same connotation as they have for the simple linear regression model.

Table 8.22 Selected Model for Baseball Regression

Model: MODEL1
Dependent Variable: WIN

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob > F
Model	6	0.11171	0.03724	16.955	0.0001
Error	36	0.07906	0.00220		
C Total	39	0.19076			
Root MSE		0.04686	R-Square	0.5856	
Dep Mean		0.50000	Adj R-sq	0.5510	
C.V.		9.37245			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter = 0	Prob > |T|
INTERCEP	1	−0.356943	0.15890423	−2.246	0.0309
BA	1	3.339829	0.60054220	5.561	0.0001
WALK	1	−0.000521	0.00013230	−3.940	0.0004
SO	1	0.000274	0.00009178	2.986	0.0051

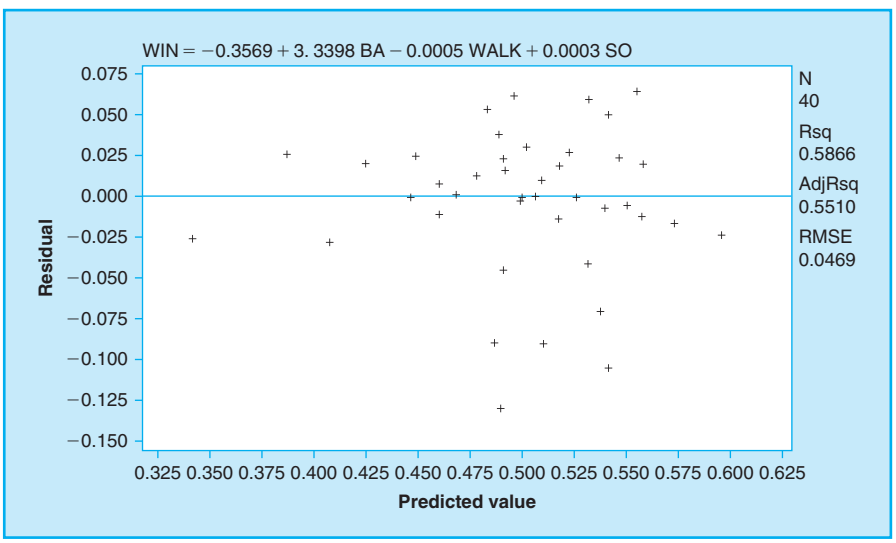

FIGURE 8.13
Residual Plot for Baseball Regression.

The multiple correlation coefficient is a measure of the strength of a multiple regression model. The square of the multiple regression coefficient is the ratio of the regression to total sum of squares, as it was for the simple linear regression model. A partial correlation coefficient is a measure of the strength of the relationship associated with a partial regression coefficient.

Although the multiple regression model must be linear in the model parameters, it may be used to describe curvilinear relationships. This is accomplished primarily by polynomial regression, but other forms may be used. A regression linear in the logarithms of the variables has special uses.

Often a proposed regression model has more independent variables than necessary for an adequate description of the data. A side effect of such model specification is that of multicollinearity, which is defined as the existence of large correlations among the independent variables. This phenomenon causes the individual regression coefficients to have large variances, often resulting in an estimated model that has good predictive power but with little statistical significance for the regression coefficients.

One possible solution to an excessive number of independent variables is to select a subset of independent variables for use in the model. Although this is very easy to do, it should be done with caution, because such procedures generate hypotheses with the data.

As in all statistical analyses, it is important to check assumptions. Because of the complexity of multiple regression, simple residual plots may not be adequate. Some additional methods for checking assumptions are presented.

8.11 CHAPTER EXERCISES

Concept Questions

1. Given that SSR $= 50$ and SSE $= 100$, calculate R^2.
2. The multiple correlation coefficient can be calculated as the simple correlation between_____and_____.
3. **(a)** What value of R^2 is required so that a regression with five independent variables is significant if there are 30 observations? [*Hint:* Use the 0.05 critical value for $F(5,24)$].
   (b) Answer part (a) if there are 500 observations.
   (c) What do these results tell us about the R^2 statistic?
4. If x is the number of inches and y is the number of pounds, what is the unit of measure of the regression coefficient?
5. What is the common feature of most "influence" statistics?
6. Under what conditions is least squares not the best method for estimating regression coefficients?

7. What is the interpretation of the regression coefficient when using logarithms of all variables?
8. What is the basic principle underlying inferences on partial regression coefficients?
9. Why is multicollinearity a problem?
10. List some reasons why variable selection is not always an appropriate remedial method when multicollinearity exists.
11. _____ (True/False) When all VIF are less than 10, then multicollinearity is not a problem.
12. _____ (True/False) The adjusted *R*-square attempts to balance good fit against model complexity.
13. _____ (True/False) The t statistic for an individual coefficient measures the contribution of the corresponding independent variable, after controlling for the other variables in the model.
14. _____ (True/False) Because polynomials are smooth functions, it is permissible to extrapolate slightly beyond the range of the independent variable when fitting quadratic models.
15. You fit a full regression model with five independent variables, obtaining an SSE with 40 df. Then you fit a reduced model that has only three of the independent variables, but now you obtain an SSE with 46 df. Does this make sense? What is the most likely explanation? What should you do?
16. The null hypothesis for the test for the model (Section 8.3) does not include the intercept term β_0. Give the interpretation of a null hypothesis that did include β_0, $H_0 : \beta_0 = \beta_1 = \ldots = \beta_m = 0$. Explain why this hypothesis would rarely be of interest.

Exercises

1. This exercise is designed to provide a review of the mechanics for performing a regression analysis. The data are:

OBS	X1	X2	Y
1	1	5	5.4
2	2	6	8.5
3	4	6	9.4
4	6	5	11.5
5	6	4	9.4
6	8	3	11.8
7	10	3	13.2
8	11	2	12.1

First we compute $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{Y}$, the sums of squares and cross products as in Table 8.3. Verify at least two or three of these elements.

MODEL CROSSPRODUCTS X'X X'Y Y'Y

X'X	INTERCEP	X1	X2	Y
INTERCEP	8	48	34	81.3
X1	48	378	171	544.9
X2	34	171	160	328.7
Y	81.3	544.9	328.7	870.27

Next we invert $\mathbf{X}'\mathbf{X}$ and compute $\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, again as in Table 8.3.

X'X INVERSE B, SSE

INVERSE	INTERCEP	X1	X2	Y
INTERCEP	12.76103	−0.762255	−1.89706	−1.44424
X1	−0.762255	0.05065359	0.1078431	1.077859
X2	−1.89706	0.1078431	0.2941176	1.209314
Y	−1.44424	1.077859	1.209314	2.859677

Verify that at least two elements of the matrix product $(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}$ are elements of an identity matrix. We next perform the partitioning of sums of squares and perform the tests for the model and the partial coefficients. Verify these computations.

DEP VARIABLE: Y

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB > F
MODEL	2	41.199073	20.599536	36.017	0.0011
ERROR	5	2.859677	0.571935		
C TOTAL	7	44.058750			

ROOT MSE	0.756264	R-SQUARE	0.9351
DEP MEAN	10.162500	ADJ R-SQ	0.9091
C.V.	7.441714		

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR HO: PARAMETER = 0	PROB > |T|
INTERCEPT	1	−1.444240	2.701571	−0.535	0.6158
X1	1	1.077859	0.170207	6.333	0.0014
X2	1	1.209314	0.410142	2.949	0.0319

Finally, we compute the predicted and residual values:

OBS	ACTUAL	PREDICT VALUE	RESIDUAL
1	5.400	5.680	−.280188
2	8.500	7.967	0.532639
3	9.400	10.123	−.723080
4	11.500	10.069	0.430515
5	9.400	9.860	−.460172
6	11.800	10.807	0.993423
7	13.200	12.962	0.237704
8	12.100	12.831	−.730842
	SUM OF RESIDUALS		1.e−14
SUM OF SQUARED RESIDUALS			2.859677

Verify at least two of the predicted and residual values and also that the sum of residuals is zero and that the sum of squares of the residuals is the ERROR sum of squares given in the partitioning of the sums of squares.

2. The complete data set on energy consumption given for Exercise 7 in Chapter 7 contains other factors that may affect power consumption. The following have been selected for this exercise:

TMAX: maximum daily temperature,
TMIN: minimum daily temperature,
WNDSPD: windspeed, coded "0" if less than 6 knots and "1" if 6 or more knots,
CLDCVR: cloud cover coded as follows:
0.0—clear
1.0—less than 0.6 covered
2.0—0.6 to 0.9 covered
3.0—cloudy (increments of 0.5 are used to denote variable cloud cover between indicated codes), and
KWH: electricity consumption.

The data are given in Table 8.23.

Table 8.23 Data for Exercise 2

OBS	MO	DAY	TMAX	TMIN	WNDSPD	CLDCVR	KWH
1	9	19	87	68	1	2.0	45
2	9	20	90	70	1	1.0	73
3	9	21	88	68	1	1.0	43
4	9	22	88	69	1	1.5	61
5	9	23	86	69	1	2.0	52

(*Continued*)

Table 8.23 (*Continued*)

OBS	MO	DAY	TMAX	TMIN	WNDSPD	CLDCVR	KWH
6	9	24	91	75	1	2.0	56
7	9	25	91	76	1	1.5	70
8	9	26	90	73	1	2.0	69
9	9	27	79	72	0	3.0	53
10	9	28	76	63	0	0.0	51
11	9	29	83	57	0	0.0	39
12	9	30	86	61	1	1.0	55
13	10	1	85	70	1	2.0	55
14	10	2	89	69	0	2.0	57
15	10	3	88	72	1	1.5	68
16	10	4	85	73	0	3.0	73
17	10	5	84	68	1	3.0	57
18	10	6	83	69	0	2.0	51
19	10	7	81	70	0	1.0	55
20	10	8	89	70	1	1.5	56
21	10	9	88	69	1	0.0	72
22	10	10	88	76	1	2.5	73
23	10	11	77	66	1	3.0	69
24	10	12	75	65	1	2.5	38
25	10	13	72	64	1	3.0	50
26	10	14	68	65	1	3.0	37
27	10	15	71	67	0	3.0	43
28	10	16	75	66	1	3.0	42
29	10	17	74	52	1	0.0	25
30	10	18	77	51	0	0.0	31
31	10	19	79	50	0	0.0	31
32	10	20	80	50	0	0.0	32
33	10	21	80	53	0	0.0	35
34	10	22	81	53	1	0.0	32
35	10	22	80	53	0	0.0	34
36	10	24	81	54	1	2.0	35
37	10	25	83	67	0	2.0	41
38	10	26	84	67	1	1.5	51
39	10	27	80	63	1	3.0	34
40	10	28	73	53	1	1.0	19
41	10	29	71	49	0	0.0	19
42	10	30	72	56	1	3.0	30
43	10	31	72	53	1	0.0	23
44	11	1	79	48	1	0.0	35
45	11	2	84	63	1	1.0	29
46	11	3	74	62	0	3.0	55
47	11	4	83	72	1	2.5	56

Perform a regression analysis to determine how the factors affect fuel consumption (KWH). Include checking for multicollinearity, variable selection (if appropriate), and outlier detection. Finally, interpret the results and assess their usefulness.

3. The data in Table 8.24 represent the results of a test for the strength of an asphalt concrete mix. The test consisted of applying a compressive force on the top of different sample specimens. Two responses occurred: the stress and strain at which a sample specimen failed. The factors relate to mixture proportions, rates of speed at which the force was applied, and ambient temperature. Higher values of the response variables indicate stronger materials.

Table 8.24 Data for Exercise 3: Asphalt Data

Obs	X1	X2	X3	Y1	Y2
1	5.3	0.02	77	42	3.20
2	5.3	0.02	32	481	0.73
3	5.3	0.02	0	543	0.16
4	6.0	2.00	77	609	1.44
5	7.8	0.20	77	444	3.68
6	8.0	2.00	104	194	3.11
7	8.0	2.00	77	593	3.07
8	8.0	2.00	32	977	0.19
9	8.0	2.00	0	872	0.00
10	8.0	0.02	104	35	5.86
11	8.0	0.02	77	96	5.97
12	8.0	0.02	32	663	0.29
13	8.0	0.02	0	702	0.04
14	10.0	2.00	77	518	2.72
15	12.0	0.02	77	40	7.35
16	12.0	0.02	32	627	1.17
17	12.0	0.02	0	683	0.14
18	12.0	0.02	104	22	15.00
19	14.0	0.02	77	35	11.80

The variables are:

X1: percent binder (the amount of asphalt in the mixture),
X2: loading rate (the speed at which the force was applied),
X3: ambient temperature,
Y1: the stress at which the sample specimen failed, and
Y2: the strain at which the specimen failed.

Perform separate regressions to relate stress and strain to the factors of the experiment. Check the residuals for possible specification errors. Interpret all results.

4. The data in Table 8.25 were collected in order to study factors affecting the supply and demand for commercial air travel. Data on various aspects of commercial air

Table 8.25 Data for Exercise 4

CITY1	CITY2	PASS	MILES	INM	INS	POPM	POPS	AIRL
ATL	AGST	3.546	141	3.246	2.606	1270	279	3
ATL	BHM	7.016	139	3.246	2.637	1270	738	4
ATL	CHIC	13.300	588	3.982	3.246	6587	1270	5
ATL	CHST	5.637	226	3.246	3.160	1270	375	5
ATL	CLBS	3.630	193	3.246	2.569	1270	299	4
ATL	CLE	3.891	555	3.559	3.246	2072	1270	3
ATL	DALL	6.776	719	3.201	3.245	1359	1270	2
ATL	DC	9.443	543	3.524	3.246	2637	1270	5
ATL	DETR	5.262	597	3.695	3.246	4063	1270	4
ATL	JAX	8.339	285	3.246	2.774	1270	505	4
ATL	LA	5.657	1932	3.759	3.246	7079	1270	3
ATL	MEM	6.286	336	3.246	2.552	1270	755	3
ATL	NO	7.058	424	3.245	2.876	1270	1050	4
ATL	NVL	5.423	214	3.246	2.807	1270	534	3
ATL	ORL	4.259	401	3.246	2.509	1270	379	3
ATL	PHIL	6.040	666	3.243	3.246	4690	1270	5
ATL	PIT	3.345	521	3.125	3.246	2413	1270	2
ATL	RAL	3.371	350	3.246	2.712	1270	198	3
ATL	SF	4.624	2135	3.977	3.246	3075	1270	3
ATL	SVNH	3.669	223	3.246	2.484	1270	188	1
ATL	TPA	7.463	413	3.246	2.586	1270	881	5
DC	NYC	150.970	205	3.962	2.524	11698	2637	12
LA	BOSTN	16.397	2591	3.759	3.423	7079	3516	4
LA	CHIC	55.681	1742	3.759	3.982	7079	6587	5
LA	DALL	18.222	1238	3.759	3.201	7079	1359	3
LA	DC	20.548	2296	3.759	3.524	7079	2637	5
LA	DENV	22.745	830	3.759	3.233	7079	1088	4
LA	DETR	17.967	1979	3.759	3.965	7079	4063	4
LA	NYC	79.450	2446	3.962	3.759	11698	7079	5
LA	PHIL	14.705	2389	3.759	3.243	7079	4690	5
LA	PHNX	29.002	356	3.759	2.841	7079	837	5
LA	SACR	24.896	361	3.759	3.477	7079	685	3
LA	SEAT	33.257	960	3.759	3.722	7079	1239	2
MIA	ATL	14.242	605	3.246	3.024	1270	1142	4
MIA	BOSTN	21.648	1257	3.423	3.024	3516	1142	5
MIA	CHIC	39.316	1190	3.982	3.124	6587	1142	5
MIA	CLE	13.669	1089	3.559	3.124	2072	1142	4
MIA	DC	14.499	925	3.524	3.024	2637	1142	6
MIA	DETR	18.537	1155	3.695	3.024	4063	1142	5
MIA	NYC	126.134	1094	3.962	3.024	11698	1142	7
MIA	PHIL	21.117	1021	3.243	3.024	4690	1142	7

(*Continued*)

Table 8.25 (*Continued*)

CITY1	CITY2	PASS	MILES	INM	INS	POPM	POPS	AIRL
MIA	TPA	18.674	205	3.024	2.586	1142	881	7
NYC	ATL	26.919	748	3.962	3.246	11698	1270	5
NYC	BOSTN	189.506	188	3.962	3.423	11698	3516	8
NYC	BUF	43.179	291	3.962	3.155	11698	1325	4
NYC	CHIC	140.445	711	3.962	3.982	11698	6587	7
NYC	CLE	53.620	404	3.962	3.559	11698	2072	7
NYC	DETR	66.737	480	3.962	3.695	11698	4063	8
NYC	PIT	53.580	315	3.962	3.125	11698	2413	7
NYC	RCH	31.681	249	3.962	3.532	11698	825	3
NYC	STL	27.380	873	3.962	3.276	11698	2320	5
NYC	SYR	32.502	193	3.962	2.974	11698	515	3
SANDG	CHIC	6.162	1731	3.982	3.149	6587	1173	3
SANDG	DALL	2.592	1181	3.201	3.149	1359	1173	2
SANDG	DC	3.211	2271	3.524	3.149	2637	1173	4
SANDG	LA	21.642	111	3.759	3.149	7079	1173	4
SANDG	LVEG	2.760	265	3.149	3.821	1173	179	5
SANDG	MINP	2.776	1532	3.621	3.149	1649	1173	2
SANDG	NYC	6.304	2429	3.962	3.149	11698	1173	4
SANDG	PHNX	6.027	298	3.149	2.841	1173	837	3
SANDG	SACR	2.603	473	3.149	3.477	1173	685	3
SANDG	SEAT	4.857	1064	3.722	3.149	1239	1173	2
SF	BOSTN	11.933	2693	3.423	3.977	3516	3075	4
SF	CHIC	33.946	1854	3.982	3.977	6587	3075	4
SF	DC	16.743	2435	3.977	3.524	3075	2637	5
SF	DENV	14.742	947	3.977	3.233	3075	1088	3
SF	LA	148.366	347	3.759	3.977	7079	3075	7
SF	LVEG	16.267	416	3.977	3.821	3075	179	6
SF	LVEG	9.410	458	3.977	3.149	3075	1173	5
SF	NYC	57.863	2566	3.962	3.977	11698	3075	5
SF	PORT	23.420	535	3.977	3.305	3075	914	4
SF	RENO	18.400	185	3.977	3.899	3075	109	3
SF	SEAT	41.725	679	3.977	3.722	3075	1239	3
SF	SLC	11.994	598	3.977	2.721	3075	526	3

travel for an arbitrarily chosen set of 74 pairs of cities were obtained from a 1966 (before deregulation) CAB study. Other data were obtained from a standard atlas. The variables are:

CITY1 and CITY2: a pair of cities,
PASS: the number of passengers flying between the cities in a sample week,
MILES: air distance between the pair of cities,
INM: per capita income in the larger city,

INS: per capita income in the smaller city,
POPM: population of the larger city,
POPS: population of the smaller city, and
AIRL: the number of airlines serving that route.

(a) Perform a regression relating the number of passengers to the other variables. Check residuals for possible specification errors. Do the results make sense?

(b) Someone suggests using the logarithms of all variables for the regression. Does this recommendation make sense? Perform the regression using logarithms; answer all questions as in part (a).

(c) Another use of the data is to use the number of airlines as the dependent variable. What different aspect of the demand or supply of airline travel is related to this model? Implement that model and relate the results to those of parts (a) and (b).

5. It is beneficial to be able to estimate the yield of useful product of a tree based on measurements of the tree taken before it is harvested. Measurements on four such variables were taken on a sample of trees, which subsequently was harvested and the actual weight of product determined. The variables are:

DBH: diameter at breast height (about 4′ from ground level), in inches,
HEIGHT: height of tree, in feet,
AGE: age of tree, in years,
GRAV: specific gravity of the wood, and
WEIGHT: the harvested weight of the tree (in lb).

The first two variables (DBH and HEIGHT) are logically the most important and are also the easiest to measure. The data are given in Table 8.26.

Table 8.26 Data for Exercise 5: Estimating Tree Weights

OBS	DBH	HEIGHT	AGE	GRAV	WEIGHT
1	5.7	34	10	0.409	174
2	8.1	68	17	0.501	745
3	8.3	70	17	0.445	814
4	7.0	54	17	0.442	408
5	6.2	37	12	0.353	226
6	11.4	79	27	0.429	1675
7	11.6	70	26	0.497	1491
8	4.5	37	12	0.380	121
9	3.5	32	15	0.420	58
10	6.2	45	15	0.449	278
11	5.7	48	20	0.471	220
12	6.0	57	20	0.447	342

(*Continued*)

Table 8.26 (*Continued*)

OBS	DBH	HEIGHT	AGE	GRAV	WEIGHT
13	5.6	40	20	0.439	209
14	4.0	44	27	0.394	84
15	6.7	52	21	0.422	313
16	4.0	38	27	0.496	60
17	12.1	74	27	0.476	1692
18	4.5	37	12	0.382	74
19	8.6	60	23	0.502	515
20	9.3	63	18	0.458	766
21	6.5	57	18	0.474	345
22	5.6	46	12	0.413	210
23	4.3	41	12	0.382	100
24	4.5	42	12	0.457	122
25	7.7	64	19	0.478	539
26	8.8	70	22	0.496	815
27	5.0	53	23	0.485	194
28	5.4	61	23	0.488	280
29	6.0	56	23	0.435	296
30	7.4	52	14	0.474	462
31	5.6	48	19	0.441	200
32	5.5	50	19	0.506	229
33	4.3	50	19	0.410	125
34	4.2	31	10	0.412	84
35	3.7	27	10	0.418	70
36	6.1	39	10	0.470	224
37	3.9	35	19	0.426	99
38	5.2	48	13	0.436	200
39	5.6	47	13	0.472	214
40	7.8	69	13	0.470	712
41	6.1	49	13	0.464	297
42	6.1	44	13	0.450	238
43	4.0	34	13	0.424	89
44	4.0	38	13	0.407	76
45	8.0	61	13	0.508	614
46	5.2	47	13	0.432	194
47	3.7	33	13	0.389	66

(a) Perform a linear regression relating weight to the measured quantities. Plot residuals. Is the equation useful? Is the model adequate?

(b) If the results appear to not be very useful, suggest and implement an alternate model. (*Hint*: Weight is a product of dimensions.)

6. Data were collected to discern environmental factors affecting health standards. For 21 small regions we have data on the following variables:

   POP: population (in thousands),
   VALUE: value of all residential housing, in millions of dollars; this is the proxy for economic conditions,
   DOCT: the number of doctors,
   NURSE: the number of nurses,
   VN: the number of vocational nurses, and
   DEATHS: number of deaths due to health-related causes (i.e., not accidents); this is the proxy for health standards.

   The data are given in Table 8.27.

Table 8.27 Data for Exercise 6

POP	VALUE	DOCT	NURSE	VN	DEATHS
100	141.83	49	76	221	661
110	246.80	103	250	378	1149
130	238.06	76	140	207	1333
142	265.90	95	150	381	1321
202	397.63	162	324	554	2418
213	464.32	194	282	560	2039
246	409.95	130	211	465	2518
280	556.03	205	383	942	3088
304	711.61	222	461	723	1882
316	820.52	304	469	598	2437
328	709.86	267	525	911	2177
330	829.84	245	639	739	2593
337	465.15	221	343	541	2295
379	839.11	330	714	330	2119
434	792.02	420	865	894	4294
434	883.72	384	601	1158	2836
436	939.71	363	530	1219	4637
447	1141.80	511	180	513	3236
1087	2511.53	1193	1792	1922	7768
2305	6774.16	3450	5357	4125	14590
2637	8318.92	3131	4630	4785	19044

   (a) Perform a regression relating DEATHS to the other variables, excluding POP. Compute the variance inflation factors; interpret all results.

   (b) Obviously multicollinearity is a problem for these data. What is the cause of this phenomenon? It has been suggested that all variables should be converted to a per capita basis. Why should this solve the multicollinearity problem?

(c) Perform the regression using per capita variables. Compare results with those of part (a). Is it useful to compare R^2 values? Why or why not?

7. We have data on the distance covered by irrigation water in a furrow of a field. The data are to be used to relate the distance covered to the time since watering began. The data are given in Table 8.28.

Table 8.28 Distance Covered by Irrigation Water

Obs	Distance	Time
1	85	0.15
2	169	0.48
3	251	0.95
4	315	1.37
5	408	2.08
6	450	2.53
7	511	3.20
8	590	4.08
9	664	4.93
10	703	5.42
11	831	7.17
12	906	8.22
13	1075	10.92
14	1146	11.92
15	1222	13.12
16	1418	15.78
17	1641	18.83
18	1914	21.22
19	1864	21.98

(a) Perform a simple linear regression relating distance to time. Plot the residuals against time. What does the plot suggest?

(b) Perform a regression using time and the square of time. Interpret the results. Are they reasonable?

(c) Plot residuals from the quadratic model. What does this plot suggest?

8. Twenty-five volunteer athletes participated in a study of cross-disciplinary athletic abilities. The group was comprised of athletes from football, baseball, water polo, volleyball, and soccer. None had ever played organized basketball, but did acknowledge interest and some social participation in the game.

Height, weight, and speed in the 100-yard dash were recorded for each subject. The basketball test consisted of the number of field goals that could be made in a 60-min. period. The data are given in Table 8.29.

Table 8.29 Basket Goals Related to Physique

OBS	WEIGHT	HEIGHT	DASH100	GOALMADE
1	130	71	11.50	15
2	149	74	12.23	19
3	170	70	12.26	11
4	177	71	12.65	15
5	188	69	10.26	12
6	210	73	12.76	17
7	223	72	11.89	15
8	170	75	12.32	19
9	145	72	10.77	16
10	132	74	11.31	18
11	211	71	12.91	13
12	212	72	12.55	15
13	193	73	11.72	17
14	146	72	12.94	16
15	158	71	12.21	15
16	154	75	11.81	20
17	193	71	11.90	15
18	228	75	11.22	19
19	217	78	10.89	22
20	172	79	12.84	23
21	188	72	11.01	16
22	144	75	12.18	20
23	164	76	12.37	21
24	188	74	11.98	19
25	231	70	12.23	13

(a) Perform the regression relating GOALMADE to the other variables. Comment on the results.
(b) Is there multicollinearity?
(c) Check for outliers.
(d) If appropriate, develop and implement an alternative model.

9. In an effort to estimate the plant biomass in a desert environment, field measurements on the diameter and height and laboratory determination of oven dry weight were obtained for a sample of plants in a sample of transects (area). Collections were made at two times, in the warm and cool seasons. The data are to be used to see how well the weight can be estimated by the more easily determined field observations, and further whether the model for estimation is the same for the two seasons. The data are given in Table 8.30.

(a) Perform separate linear regressions for estimating weight for the two seasons. Plot residuals. Interpret results.

Table 8.30 Data for Exercise 9

COOL			WARM		
Width	Height	Weight	Width	Height	Weight
4.9	7.6	0.420	20.5	13.0	6.840
8.6	4.8	0.580	10.0	6.2	0.400
4.5	3.9	0.080	10.1	5.9	0.360
19.6	19.8	8.690	10.5	27.0	1.385
7.7	3.1	0.480	9.2	16.1	1.010
5.3	2.2	0.540	12.1	12.3	1.825
4.5	3.1	0.400	18.6	7.2	6.820
7.1	7.1	0.350	29.5	29.0	9.910
7.5	3.6	0.470	45.0	16.0	4.525
10.2	1.4	0.720	5.0	3.1	0.110
8.6	7.4	2.080	6.0	5.8	0.200
15.2	12.9	5.370	12.4	20.0	1.360
9.2	10.7	4.050	16.4	2.1	1.720
3.8	4.4	0.850	8.1	1.2	1.495
11.4	15.5	2.730	5.0	23.1	1.725
10.6	6.6	1.450	15.6	24.1	1.830
7.6	6.4	0.420	28.2	2.2	4.620
11.2	7.4	7.380	34.6	45.0	15.310
7.4	6.4	0.360	4.2	6.1	0.190
6.3	3.7	0.320	30.0	30.0	7.290
16.4	8.7	5.410	9.0	19.1	0.930
4.1	26.1	1.570	25.4	29.3	8.010
5.4	11.8	1.060	8.1	4.8	0.600
3.8	11.4	0.470	5.4	10.6	0.250
4.6	7.9	0.610	2.0	6.0	0.050
			18.2	16.1	5.450
			13.5	18.0	0.640
			26.6	9.0	2.090
			6.0	10.7	0.210
			7.6	14.0	0.680
			13.1	12.2	1.960
			16.5	10.0	1.610
			23.1	19.5	2.160
			9.0	30.0	0.710

(b) Transform width, height, and weight using the natural logarithm transform discussed in Section 8.6. Perform separate regressions for estimating log–weight for the two seasons. Plot residuals. Interpret results. Compare results with those from part (a). (A formal method for comparing the regressions for the two seasons is presented in Chapter 11 and is applied to this exercise in Exercise 10, Chapter 11.)

10. In this problem we are trying to estimate the survival of liver transplant patients using information on the patients collected before the operation. The variables are:

CLOT: a measure of the clotting potential of the patient's blood,
PROG: a subjective index of the patient's prospect of recovery,
ENZ: a measure of a protein present in the body,
LIV: a measure relating to white blood cell count and the response, and
TIME: a measure of the survival time of the patient.

The data are given in Table 8.31.

Table 8.31 Survival of Liver Transplant Patients

OBS	CLOT	PROG	ENZ	LIV	TIME
1	3.7	51	41	1.55	34
2	8.7	45	23	2.52	58
3	6.7	51	43	1.86	65
4	6.7	26	68	2.10	70
5	3.2	64	65	0.74	71
6	5.2	54	56	2.71	72
7	3.6	28	99	1.30	75
8	5.8	38	72	1.42	80
9	5.7	46	63	1.91	80
10	6.0	85	28	2.98	87
11	5.2	49	72	1.84	95
12	5.1	59	66	1.70	101
13	6.5	73	41	2.01	101
14	5.2	52	76	2.85	109
15	5.4	58	70	2.64	115
16	5.0	59	73	3.50	116
17	2.6	74	86	2.05	118
18	4.3	8	119	2.85	120
19	6.5	40	84	3.00	123
20	6.6	77	46	1.95	124
21	6.4	85	40	1.21	125
22	3.7	68	81	2.57	127
23	3.4	83	53	1.12	136
24	5.8	61	73	3.50	144
25	5.4	52	88	1.81	148
26	4.8	61	76	2.45	151
27	6.5	56	77	2.85	153
28	5.1	67	77	2.86	158
29	7.7	62	67	3.40	168
30	5.6	57	87	3.02	172
31	5.8	76	59	2.58	178

(*Continued*)

Table 8.31 (*Continued*)

OBS	CLOT	PROG	ENZ	LIV	TIME
32	5.2	52	86	2.45	181
33	5.3	51	99	2.60	184
34	3.4	77	93	1.48	191
35	6.4	59	85	2.33	198
36	6.7	62	81	2.59	200
37	6.0	67	93	2.50	202
38	3.7	76	94	2.40	203
39	7.4	57	83	2.16	204
40	7.3	68	74	3.56	215
41	7.4	74	68	2.40	217
42	5.8	67	86	3.40	220
43	6.3	59	100	2.95	276
44	5.8	72	93	3.30	295
45	3.9	82	103	4.55	310
46	4.5	73	106	3.05	311
47	8.8	78	72	3.20	313
48	6.3	84	83	4.13	329
49	5.8	83	88	3.95	330
50	4.8	86	101	4.10	398
51	8.8	86	88	6.40	483
52	7.8	65	115	4.30	509
53	11.2	76	90	5.59	574
54	5.8	96	114	3.95	830

(a) Perform a linear regression for estimating survival times. Plot residuals. Interpret and critique the model used.

(b) Because the distributions of survival times are often quite skewed, a logarithmic model is often used for such data. Perform the regression using such a model. Compare the results with those of part (a).

11. Considerable variation occurs among individuals in their perception of what specific acts constitute a crime. To obtain an idea of factors that influence this perception, 45 college students were given the following list of acts and asked how many of these they perceived as constituting a crime. The acts were:

aggravated assault	armed robbery	arson
atheism	auto theft	burglary
civil disobedience	communism	drug addiction
embezzlement	forcible rape	gambling
homosexuality	land fraud	Nazism
payola	price fixing	prostitution
sexual abuse of child	sex discrimination	shoplifting
striking	strip mining	treason
vandalism		

The number of activities perceived as crimes is measured by the variable CRIMES. Variables describing personal information that may influence perception are:

AGE: age of interviewee,
SEX: coded 0: female, 1: male,
COLLEGE: year of college, coded 1 through 4, and
INCOME: income of parents ($1000).

Perform a regression to estimate the relationship between the number of activities perceived as crimes and the personal characteristics of the interviewees. Check assumptions and perform any justifiable remedial actions. Interpret the results. The data are given in Table 8.32.

Table 8.32 Crimes Perception Data–Exercise 11

OBS	AGE	SEX	COLLEGE	INCOME	CRIMES
1	19	0	2	56	13
2	19	1	2	59	16
3	20	0	2	55	13
4	21	0	2	60	13
5	20	0	2	52	14
6	24	0	3	54	14
7	25	0	3	55	13
8	25	0	3	59	16
9	27	1	4	56	16
10	28	1	4	52	14
11	38	0	4	59	20
12	29	1	4	63	25
13	30	1	4	55	19
14	21	1	3	29	8
15	21	1	2	35	11
16	20	0	2	33	10
17	19	0	2	27	6
18	21	0	3	24	7
19	21	1	2	53	15
20	16	1	2	63	23
21	18	1	2	72	25
22	18	1	2	75	22
23	18	0	2	61	16
24	19	1	2	65	19
25	19	1	2	70	19
26	20	1	2	78	18
27	19	0	2	76	16

(*Continued*)

Table 8.32 (*Continued*)

OBS	AGE	SEX	COLLEGE	INCOME	CRIMES
28	18	0	2	53	12
29	31	0	4	59	23
30	32	1	4	62	25
31	32	1	4	55	22
32	31	0	4	57	25
33	30	1	4	46	17
34	29	0	4	35	14
35	29	0	4	32	12
36	28	0	4	30	10
37	27	0	4	29	8
38	26	0	4	28	7
39	25	0	4	25	5
40	24	0	3	33	9
41	23	0	3	26	7
42	23	1	3	28	9
43	22	0	3	38	10
44	22	0	3	24	4
45	22	0	3	28	6

12. The data from Taiwo *et al.* (1998) used in Case Study 7.1 is shown in Table 8.33. Convert the data to logarithms, $y = \ln(WATER)$ and $x = \ln(STIME)$.

Table 8.33 Data for Exercise 12

STIME	0.25	0.50	0.75	1.00	2.00	3.00	4.00	5.00	6.00
WATER	4.6	5.9	6.8	8.2	9.3	10.1	10.5	10.5	10.4

(a) Fit a quadratic model, of the form $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$.
(b) Modify the independent variable to create $x^* = \ln(STIME)$ if $STIME \leq 4$, and $x^* = \ln(4)$ if $STIME > 4$. Fit the quadratic model using the new independent variable.
(c) Which model fits better?

13. An apartment complex owner is performing a study to see what improvements or changes in her complex may bring in more rental income. From a sample of 34 complexes she obtains the monthly rent on single-bedroom units and the following characteristics:

AGE: the age of the property,
SQFT: square footage of unit,
SD: amount of security deposit,

UNTS: number of units in complex,
GAR: presence of a garage (0–no, 1–yes),
CP: presence of a carport (0–no, 1–yes),
SS: security system (0–no, 1–yes),
FIT: fitness facilities (0–no, 1–yes), and
RENT: monthly rental.

The data are presented in Table 8.34.

Table 8.34 Apartment Rent Data

OBS	AGE	SQFT	SD	UNTS	GAR	CP	SS	FIT	RENT
1	7	692	150	408	0	0	1	0	508
2	7	765	100	334	0	0	1	1	553
3	8	764	150	170	0	0	1	1	488
4	13	808	100	533	0	1	1	1	558
5	7	685	100	264	0	0	0	0	471
6	7	710	100	296	0	0	0	0	481
7	5	718	100	240	0	1	1	1	577
8	6	672	100	420	0	1	0	1	556
9	4	746	100	410	1	1	1	1	636
10	4	792	100	404	1	0	1	1	737
11	8	797	150	252	0	0	1	1	546
12	7	708	100	276	0	0	1	0	445
13	8	797	150	252	0	0	0	1	533
14	6	813	100	416	0	1	0	0	617
15	7	708	100	536	0	0	1	1	475
16	16	658	100	188	1	1	1	1	525
17	8	809	150	192	0	0	1	0	461
18	7	663	100	300	0	0	0	1	495
19	1	719	100	300	1	1	1	1	601
20	1	689	100	224	0	1	1	1	567
21	1	737	175	310	1	1	1	1	633
22	1	694	150	476	1	0	1	1	616
23	7	768	150	264	0	0	1	1	507
24	6	699	150	150	0	0	0	0	454
25	6	733	100	260	0	0	1	0	502
26	7	592	100	264	0	0	1	1	431
27	6	589	150	516	0	0	1	1	418
28	8	721	75	216	0	0	1	0	538
29	5	705	75	212	1	0	1	1	506
30	6	772	150	460	0	0	1	1	543

(*Continued*)

Table 8.34 (*Continued*)

OBS	AGE	SQFT	SD	UNTS	GAR	CP	SS	FIT	RENT
31	7	758	100	260	0	0	1	0	534
32	7	764	100	269	0	0	1	0	536
33	6	722	125	216	0	0	0	1	520
34	1	703	100	248	0	0	1	0	530

(a) Perform a regression and make recommendations to the apartment complex owner.

(b) Because there is no way to change some of these characteristics, someone recommends using a model that contains only characteristics that can be modified. Comment on that recommendation.

14. **(a)** Use the data set on home prices given in Table 8.2 to do the following:

(i) Use `price` as the dependent variable and the rest of the variables as independent variables and determine the best regression using the stepwise variable selection procedure. Comment on the results.

(ii) The Modes decided to not use the data on homes whose price exceeded $200,000, because the relationship of price to size seemed to be erratic for these homes. Perform the regression using all observations, and compute the outlier detection statistics. Also compare the results of the regression with that obtained using only the under $200,000 homes. Comment on the results. Which regression would you use?

(iii) Compute and study the residuals for the home price regression. Could these be useful for someone who was considering buying one of these homes?

(b) The data originally presented in Chapter 1 (Table 1.2) also included the variables `garage` and `fp`. Perform variable selection that includes these variables as well. Explain the results.

15. In a data set with $n = 50$ observations, you try fitting two models. The first model is a simple linear model $(m_1 = 1)$ resulting in $SSE_1 = 932$. The second model is a cubic polynomial $(m_2 = 3)$ resulting in $SSE_2 = 901$. Did the second model fit significantly better than the first model? Give the formal hypothesis that corresponds to this question, and show the construction of the appropriate test statistic. Use $\alpha = 0.01$.

16. In a data set with $n = 50$ observations, you try fitting two models:

Model 1: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, giving $SSE_1 = 256$,

Model 2: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$, giving $SSE_2 = 194$ and $\hat{\beta}_3 = 2.1$.

(a) Calculate the F statistic for the null hypothesis that x_3 is not related to y, after controlling for x_1 and x_2. Interpret the result.

(b) Calculate the t statistic for the coefficient for x_3 and interpret the result.

(c) Using your result from part (b), calculate the estimated standard error for $\hat{\beta}_3 = 2.1$, then construct a 95% confidence interval for β_3.

17. A multiple regression results in the fitted equation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2 = 5 + 5x_1 + 2x_2 - 1.5x_1 x_2,$$

where x_1 represents participants' gender (0 if boy, 1 if girl) and x_2 ranges from 0 to 5.

(a) Plot the fitted regression line for y versus x_2 for boys and for girls.

(b) In simple language, how would you describe the differences between boys and girls having the same value of x_2?

(c) In terms of the true regression coefficients (the β_i), how would you represent the difference between a girl and a boy both having $x_2 = 3$?

(d) What would a reasonable point estimate be for the quantity in part (c) ?

18. A multiple regression results in the fitted equation

$$\hat{y} = 4 + 1.5x_1 - 1x_2 + 2x_1 x_2,$$

where x_1 ranges from 0 to 2 and x_2 ranges from -1 to 1.

(a) Plot the fitted regression equation of y versus x_1 using a low value of x_2 and a high value of x_2.

(b) In simple language, describe the relationship between y and x_1.

(c) Can you interpret $\hat{\beta}_1 = 1.5$ as being the expected change in y if x_1 increases by 1? Why or why not?

(d) Suppose the fitted equation had been $\hat{y} = 4 + 1.5x_1 - 1x_2 + 0x_1 x_2$. Redraw the plot. How is the description of the relationship simplified?

19. Lopez and Russell (2008) studied $y =$ Rehabilitative Orientation (RO) among a sample of n = 100 juvenile justice workers. Table 8.35 is taken from their Table 8, and summarizes the results of two of their multiple regression models. Model 2 fits two additional independent variables beyond those in Model 1.

Table 8.35 Information for Exercise 19

	Model 1		Model 2	
ind. variables	$\hat{\beta}$	s.e.($\hat{\beta}$)	$\hat{\beta}$	s.e.($\hat{\beta}$)
social support	0.53	0.20	0.53	0.19
cultural competency	0.00	0.01	−0.00	0.00
type of work[a]			−0.56	0.17
employment length			0.05	0.02
R^2	0.07		0.19	

[a] *type of work coded as* 0 = *diversion*, 1 = *nondiversion.*

(a) For this data, the total SS for y corrected for the mean was 45.778. Calculate SS due to regression model, SSE, and F for each model, and interpret each of the F statistics.

(b) In Model 2, is there significant evidence (at $\alpha = 0.05$) that type of work is associated with RO? If so, which group appears to have higher expected RO?

(c) Give a 95% confidence interval for the expected difference in RO for two workers with the same values of social support, cultural competency, and employment length, but one in diversion and the other in nondiversion work.

(d) Construct an F test of the null hypothesis that neither type of work nor employment length are associated with RO, after controlling for social support and cultural competency (use $\alpha = 5\%$).

20. Martinussen *et al.* (2007) studied burnout among Norwegian policemen. In a sample of $n = 220$, they regressed $y =$ frequency of psychosomatic complaints on demographic variables gender ($0 =$ man, $1 =$ woman) and age ($m = 2$). This regression had $R^2 = 0.05$. They then added independent variables exhaustion burnout score, cynicism burnout score, and professional efficacy burnout score ($m = 5$). This regression had $R^2 = 0.34$. Given that $TSS = 33.7$, is there significant evidence that at least one of the burnout scores is related to psychosomatic complaints, after controlling for gender and age? Use $\alpha = 0.05$.

Projects

1. Lake Data Set. The Florida Lakewatch data set (Appendix C.1) gives data on algae levels (as measured by total chlorophyll) and nutrient levels (total nitrogen and total phosphorus) separately for winter and for summer. If total chlorophyll is strongly positively related to one (or both) of these nutrients, then those nutrients are likely acting as limiting factors on the growth of algae. Build a regression model for winter data that relates chlorophyll levels to nitrogen and phosphorus, transforming the variables as necessary. Do the same using the summer data. Which, if any, of the variables appears to act as a limiting factor? Is the answer the same for winter as it is in summer?

2. State Education Data Set. This data set is described in Appendix C.2. In Project 3 in Chapter 7, it was shown that the percentage of high school seniors taking the SAT (`TakePCT`) is an important predictor of a state's mean total SAT score (`SATTotal`). After controlling for `TakePCT`, does the per capita amount a state spends on student education (`expend_pc`) have an association with `SATTotal`? What if we also control for poverty rate (`pov_rate`)? What is the practical significance of these results?

3. Cowpea Data Set. Case Study 7.1 introduces a small part of the cowpea data given by Taiwo *et al.* (1998). The complete data from their Table 2 is given in Appendix C.5. Using the data for Variety 1 = Ife-BPC, find a model that will predict the quantity of water absorbed (`WATER`) as a function of soaking time (`STIME`) and soaking temperature (`STEMP`). It may be necessary to transform some or all of these variables. It is likely that you will need a quadratic term in one of the independent variables, and it may also be necessary to include an interaction (the product of two separate independent variables).