where the shaded region is the region of observations for a multiple regression application with two predictor variables and the circled dot represents the values (X_{h1}, X_{h2}) for which a prediction is to be made. The circled dot is within the ranges of the predictor variables X_1 and X_2 individually, yet is well outside the joint region of observations. It is easy to spot this extrapolation when there are only two predictor variables, but it becomes much more difficult when the number of predictor variables is large. We discuss in Chapter 10 a procedure for identifying hidden extrapolations when there are more than two predictor variables.

6.8 Diagnostics and Remedial Measures

Diagnostics play an important role in the development and evaluation of multiple regression models. Most of the diagnostic procedures for simple linear regression that we described in Chapter 3 carry over directly to multiple regression. We review these diagnostic procedures now, as well as the remedial measures for simple linear regression that carry over directly to multiple regression.

Many specialized diagnostics and remedial procedures for multiple regression have also been developed. Some important ones will be discussed in Chapters 10 and 11.

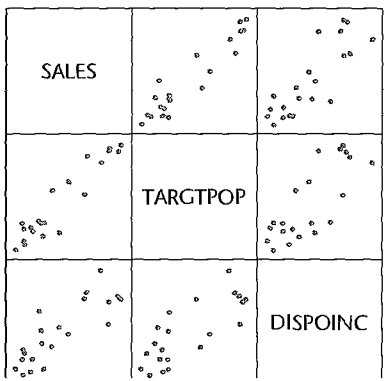
Scatter Plot Matrix

Box plots, sequence plots, stem-and-leaf plots, and dot plots for each of the predictor variables and for the response variable can provide helpful, preliminary univariate information about these variables. Scatter plots of the response variable against each predictor variable can aid in determining the nature and strength of the bivariate relationships between each of the predictor variables and the response variable and in identifying gaps in the data points as well as outlying data points. Scatter plots of each predictor variable against each of the other predictor variables are helpful for studying the bivariate relationships among the predictor variables and for finding gaps and detecting outliers.

Analysis is facilitated if these scatter plots are assembled in a *scatter plot matrix*, such as in Figure 6.4. In this figure, the Y variable for any one scatter plot is the name found in

FIGURE 6.4 SYGRAPH Scatter Plot Matrix and Correlation Matrix—Dwaine Studios Example.

(a) Scatter Plot Matrix



(b) Correlation Matrix

	SALES	TARGTPOP	DISPOINC
SALES	1.000	.945	.836
TARGTPOP		1.000	.781
DISPOINC			1.000

its row, and the X variable is the name found in its column. Thus, the scatter plot matrix in Figure 6.4 shows in the first row the plots of Y (SALES) against X_1 (TARGETPOP) and X_2 (DISPOINC), of X_1 against Y and X_2 in the second row, and of X_2 against Y and X_1 in the third row. These variables are described on page 236. Alternatively, by viewing the first column, one can compare the plots of X_1 and X_2 each against Y, and similarly for the other two columns. A scatter plot matrix facilitates the study of the relationships among the variables by comparing the scatter plots within a row or a column. Examples in this and subsequent chapters will illustrate the usefulness of scatter plot matrices.

A complement to the scatter plot matrix that may be useful at times is the correlation matrix. This matrix contains the coefficients of simple correlation $r_{Y1}, r_{Y2}, \ldots, r_{Y,p-1}$ between Y and each of the predictor variables, as well as all of the coefficients of simple correlation among the predictor variables—r_{12} between X_1 and X_2, r_{13} between X_1 and X_3, etc. The format of the correlation matrix follows that of the scatter plot matrix:

$$\begin{bmatrix} 1 & r_{Y1} & r_{Y2} & \cdots & r_{Y,p-1} \\ r_{Y1} & 1 & r_{12} & \cdots & r_{1,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ r_{Y,p-1} & r_{1,p-1} & r_{2,p-1} & \cdots & 1 \end{bmatrix} \tag{6.67}$$

Note that the correlation matrix is symmetric and that its main diagonal contains 1s because the coefficient of correlation between a variable and itself is 1. Many statistics packages provide the correlation matrix as an option. Since this matrix is symmetric, the lower (or upper) triangular block of elements is frequently omitted in the output.

Some interactive statistics packages enable the user to employ *brushing* with scatter plot matrices. When a point in a scatter plot is brushed, it is given a distinctive appearance on the computer screen in each scatter plot in the matrix. The case corresponding to the brushed point may also be identified. Brushing is helpful to see whether a case that is outlying in one scatter plot is also outlying in some or all of the other plots. Brushing may also be applied to a group of points to see, for instance, whether a group of cases that does not fit the relationship for the remaining cases in one scatter plot also follows a distinct pattern in any of the other scatter plots.

Three-Dimensional Scatter Plots

Some interactive statistics packages provide *three-dimensional scatter plots* or *point clouds*, and permit spinning of these plots to enable the viewer to see the point cloud from different perspectives. This can be very helpful for identifying patterns that are only apparent from certain perspectives. Figure 6.6 on page 238 illustrates a three-dimensional scatter plot and the use of spinning.

Residual Plots

A plot of the residuals against the fitted values is useful for assessing the appropriateness of the multiple regression function and the constancy of the variance of the error terms, as well as for providing information about outliers, just as for simple linear regression. Similarly,

a plot of the residuals against time or against some other sequence can provide diagnostic information about possible correlations between the error terms in multiple regression. Box plots and normal probability plots of the residuals are useful for examining whether the error terms are reasonably normally distributed.

In addition, residuals should be plotted against each of the predictor variables. Each of these plots can provide further information about the adequacy of the regression function with respect to that predictor variable (e.g., whether a curvature effect is required for that variable) and about possible variation in the magnitude of the error variance in relation to that predictor variable.

Residuals should also be plotted against important predictor variables that were omitted from the model, to see if the omitted variables have substantial additional effects on the response variable that have not yet been recognized in the regression model. Also, residuals should be plotted against interaction terms for potential interaction effects not included in the regression model, such as against X_1X_2, X_1X_3, and X_2X_3, to see whether some or all of these interaction terms are required in the model.

A plot of the absolute residuals or the squared residuals against the fitted values is useful for examining the constancy of the variance of the error terms. If nonconstancy is detected, a plot of the absolute residuals or the squared residuals against each of the predictor variables may identify one or several of the predictor variables to which the magnitude of the error variability is related.

Correlation Test for Normality

The correlation test for normality described in Chapter 3 carries forward directly to multiple regression. The expected values of the ordered residuals under normality are calculated according to (3.6), and the coefficient of correlation between the residuals and the expected values under normality is then obtained. Table B.6 is employed to assess whether or not the magnitude of the correlation coefficient supports the reasonableness of the normality assumption.

Brown-Forsythe Test for Constancy of Error Variance

The Brown-Forsythe test statistic (3.9) for assessing the constancy of the error variance can be used readily in multiple regression when the error variance increases or decreases with one of the predictor variables. To conduct the Brown-Forsythe test, we divide the data set into two groups, as for simple linear regression, where one group consists of cases where the level of the predictor variable is relatively low and the other group consists of cases where the level of the predictor variable is relatively high. The Brown-Forsythe test then proceeds as for simple linear regression.

Breusch-Pagan Test for Constancy of Error Variance

The Breusch-Pagan test (3.11) for constancy of the error variance in multiple regression is carried out exactly the same as for simple linear regression when the error variance increases or decreases with one of the predictor variables. The squared residuals are simply regressed against the predictor variable to obtain the regression sum of squares SSR^*, and the test proceeds as before, using the error sum of squares SSE for the full multiple regression model.

When the error variance is a function of more than one predictor variable, a multiple regression of the squared residuals against these predictor variables is conducted and the regression sum of squares SSR^* is obtained. The test statistic again uses SSE for the full multiple regression model, but now the chi-square distribution involves q degrees of freedom, where q is the number of predictor variables against which the squared residuals are regressed.

F Test for Lack of Fit

The lack of fit F test described in Chapter 3 for simple linear regression can be carried over to test whether the multiple regression response function:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}$$

is an appropriate response surface. Repeat observations in multiple regression are replicate observations on Y corresponding to levels of each of the X variables that are constant from trial to trial. Thus, with two predictor variables, repeat observations require that X_1 and X_2 each remain at given levels from trial to trial.

Once the ANOVA table, shown in Table 6.1, has been obtained, SSE is decomposed into pure error and lack of fit components. The pure error sum of squares $SSPE$ is obtained by first calculating for each replicate group the sum of squared deviations of the Y observations around the group mean, where a replicate group has the same values for each of the X variables. Let c denote the number of groups with distinct sets of levels for the X variables, and let the mean of the Y observations for the jth group be denoted by $\bar{Y}_j$. Then the sum of squares for the jth group is given by (3.17), and the pure error sum of squares is the sum of these sums of squares, as given by (3.16). The lack of fit sum of squares $SSLF$ equals the difference $SSE - SSPE$, as indicated by (3.24).

The number of degrees of freedom associated with $SSPE$ is $n - c$, and the number of degrees of freedom associated with $SSLF$ is $(n - p) - (n - c) = c - p$. Thus, for testing the alternatives:

$$\begin{aligned} H_0&: E\{Y\} = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} \\ H_a&: E\{Y\} \neq \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} \end{aligned} \tag{6.68a}$$

the appropriate test statistic is:

$$F^* = \frac{SSLF}{c-p} \div \frac{SSPE}{n-c} = \frac{MSLF}{MSPE} \tag{6.68b}$$

where $SSLF$ and $SSPE$ are given by (3.24) and (3.16), respectively, and the appropriate decision rule is:

$$\begin{aligned} &\text{If } F^* \leq F(1-\alpha; c-p, n-c), \text{ conclude } H_0 \\ &\text{If } F^* > F(1-\alpha; c-p, n-c), \text{ conclude } H_a \end{aligned} \tag{6.68c}$$

Comment

When replicate observations are not available, an approximate lack of fit test can be conducted if there are cases that have similar $\mathbf{X}_h$ vectors. These cases are grouped together and treated as pseudoreplicates, and the test for lack of fit is then carried out using these groupings of similar cases. ■

Remedial Measures

The remedial measures described in Chapter 3 are also applicable to multiple regression. When a more complex model is required to recognize curvature or interaction effects, the multiple regression model can be expanded to include these effects. For example, X_2^2 might be added as a variable to take into account a curvature effect of X_2, or X_1X_3 might be added as a variable to recognize an interaction effect between X_1 and X_3 on the response variable. Alternatively, transformations on the response and/or the predictor variables can be made, following the principles discussed in Chapter 3, to remedy model deficiencies. Transformations on the response variable Y may be helpful when the distributions of the error terms are quite skewed and the variance of the error terms is not constant. Transformations of some of the predictor variables may be helpful when the effects of these variables are curvilinear. In addition, transformations on Y and/or the predictor variables may be helpful in eliminating or substantially reducing interaction effects.

As with simple linear regression, the usefulness of potential transformations needs to be examined by means of residual plots and other diagnostic tools to determine whether the multiple regression model for the transformed data is appropriate.

Box-Cox Transformations. The Box-Cox procedure for determining an appropriate power transformation on Y for simple linear regression models described in Chapter 3 is also applicable to multiple regression models. The standardized variable W in (3.36) is again obtained for different values of the parameter λ and is now regressed against the set of X variables in the multiple regression model to find that value of λ that minimizes the error sum of squares *SSE*.

Box and Tidwell (Ref. 6.1) have also developed an iterative approach for ascertaining appropriate power transformations for each predictor variable in a multiple regression model when transformations on the predictor variables may be required.

6.9 An Example—Multiple Regression with Two Predictor Variables

In this section, we shall develop a multiple regression application with two predictor variables. We shall illustrate several diagnostic procedures and several types of inferences that might be made for this application. We shall set up the necessary calculations in matrix format but, for ease of viewing, show fewer significant digits for the elements of the matrices than are used in the actual calculations.

Setting

Dwaine Studios, Inc., operates portrait studios in 21 cities of medium size. These studios specialize in portraits of children. The company is considering an expansion into other cities of medium size and wishes to investigate whether sales (Y) in a community can be predicted from the number of persons aged 16 or younger in the community (X_1) and the per capita disposable personal income in the community (X_2). Data on these variables for the most recent year for the 21 cities in which Dwaine Studios is now operating are shown in Figure 6.5b. Sales are expressed in thousands of dollars and are labeled Y or SALES; the number of persons aged 16 or younger is expressed in thousands of persons and is

FIGURE 6.5 SYSTAT Multiple Regression Output and Basic Data—Dwaine Studios Example.

(a) Multiple Regression Output

```
DEP VAR: SALES N: 21 MULTIPLE R: 0.957 SQUARED MULTIPLE R:
                                              0.917
ADJUSTED SQUARED MULTIPLE R: .907 STANDARD ERROR OF ESTIMATE:
                                             11.0074

VARIABLE    COEFFICIENT    STD ERROR    STD COEF   TOLERANCE        T        P(2 TAIL)
CONSTANT      -68.8571       60.0170      0.0000          .      -1.1473      0.2663
TARGTPOP        1.4546        0.2118      0.7484     0.3896      6.8682      0.0000
DISPOINC        9.3655        4.0640      0.2511     0.3896      2.3045      0.0333

                        ANALYSIS OF VARIANCE

SOURCE          SUM-OF-SQUARES   DF    MEAN-SQUARE     F-RATIO      P
REGRESSION        24015.2821      2     12007.6411      99.1035    0.0000
RESIDUAL           2180.9274     18       121.1626

INVERSE (X'X)

                1            2            3
1          29.7289
2           0.0722       0.00037
3          -1.9926      -0.0056       0.1363
```

(b) Basic Data

CASE	X1	X2	Y	FITTED	RESIDUAL
1	68.5	16.7	174.4	187.184	-12.7841
2	45.2	16.8	164.4	154.229	10.1706
3	91.3	18.2	244.2	234.396	9.8037
4	47.8	16.3	154.6	153.329	1.2715
5	46.9	17.3	181.6	161.385	20.2151
6	66.1	18.2	207.5	197.741	9.7586
7	49.5	15.9	152.8	152.055	0.7449
8	52.0	17.2	163.2	167.867	-4.6666
9	48.9	16.6	145.4	157.738	-12.3382
10	38.4	16.0	137.2	136.846	0.3540
11	87.9	18.3	241.9	230.387	11.5126
12	72.8	17.1	191.1	197.185	-6.0849
13	88.4	17.4	232.0	222.686	9.3143
14	42.9	15.8	145.3	141.518	3.7816
15	52.5	17.8	161.1	174.213	-13.1132
16	85.7	18.4	209.7	228.124	-18.4239
17	41.3	16.5	146.4	145.747	0.6530
18	51.7	16.3	144.0	159.001	-15.0013
19	89.6	18.1	232.6	230.987	1.6130
20	82.7	19.1	224.1	230.316	-6.2160
21	52.3	16.0	166.5	157.064	9.4356

labeled X_1 or TARGTPOP for target population; and per capita disposable personal income is expressed in thousands of dollars and labeled X_2 or DISPOINC for disposable income.

The first-order regression model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \tag{6.69}$$

with normal error terms is expected to be appropriate, on the basis of the SYGRAPH scatter plot matrix in Figure 6.4a. Note the linear relation between target population and sales and between disposable income and sales. Also note that there is more scatter in the latter relationship. Finally note that there is also some linear relationship between the two predictor variables. The correlation matrix in Figure 6.4b bears out these visual impressions from the scatter plot matrix.

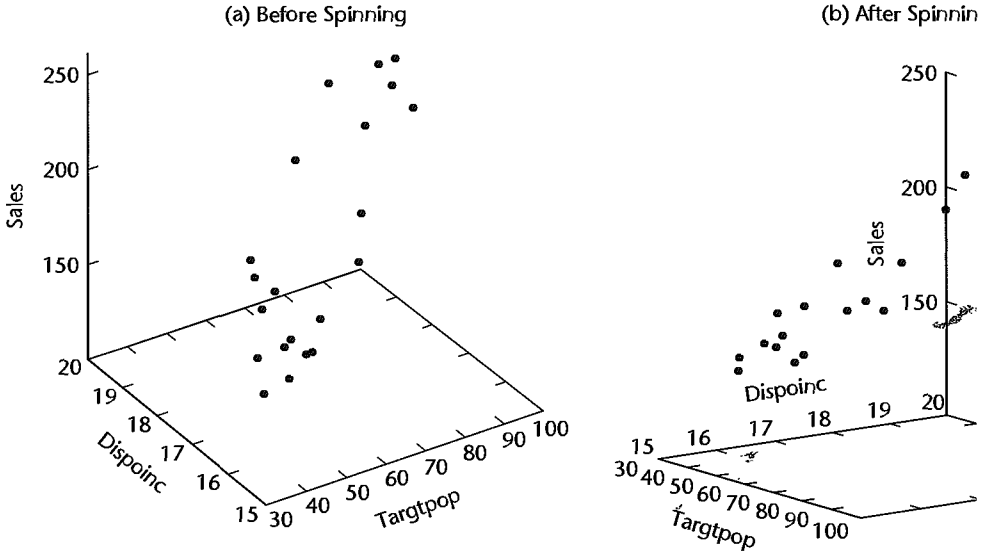
A SYGRAPH plot of the point cloud is shown in Figure 6.6a. By spinning the axes, we obtain the perspective in Figure 6.6b which supports the tentative conclusion that a response plane may be a reasonable regression function to utilize here.

Basic Calculations

The **X** and **Y** matrices for the Dwaine Studios example are as follows:

$$\mathbf{X} = \begin{bmatrix} 1 & 68.5 & 16.7 \\ 1 & 45.2 & 16.8 \\ \vdots & \vdots & \vdots \\ 1 & 52.3 & 16.0 \end{bmatrix} \qquad \mathbf{Y} = \begin{bmatrix} 174.4 \\ 164.4 \\ \vdots \\ 166.5 \end{bmatrix} \tag{6.70}$$

FIGURE 6.6 SYGRAPH Plot of Point Cloud before and after Spinning—Dwaine Studios Exa



We require:

1.

$$\mathbf{X'X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 68.5 & 45.2 & \cdots & 52.3 \\ 16.7 & 16.8 & \cdots & 16.0 \end{bmatrix} \begin{bmatrix} 1 & 68.5 & 16. \\ 1 & 45.2 & 16. \\ \vdots & \vdots & \vdots \\ 1 & 52.3 & 16. \end{bmatrix}$$

which yields:

$$\mathbf{X'X} = \begin{bmatrix} 21.0 & 1{,}302.4 & 360.0 \\ 1{,}302.4 & 87{,}707.9 & 22{,}609.2 \\ 360.0 & 22{,}609.2 & 6{,}190.3 \end{bmatrix}$$

2.

$$\mathbf{X'Y} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 68.5 & 45.2 & \cdots & 52.3 \\ 16.7 & 16.8 & \cdots & 16.0 \end{bmatrix} \begin{bmatrix} 174.4 \\ 164.4 \\ \vdots \\ 166.5 \end{bmatrix}$$

which yields:

$$\mathbf{X'Y} = \begin{bmatrix} 3{,}820 \\ 249{,}643 \\ 66{,}073 \end{bmatrix}$$

3.

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 21.0 & 1{,}302.4 & 360.0 \\ 1{,}302.4 & 87{,}707.9 & 22{,}609.2 \\ 360.0 & 22{,}609.2 & 6{,}190.3 \end{bmatrix}^{-1}$$

Using (5.23), we obtain:

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 29.7289 & .0722 & -1.9926 \\ .0722 & .00037 & -.0056 \\ -1.9926 & -.0056 & .1363 \end{bmatrix} \tag{6.73}$$

Algebraic Equivalents. Note that $\mathbf{X}'\mathbf{X}$ for the first-order regression model (6.69) with two predictor variables is:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_{11} & X_{21} & \cdots & X_{n1} \\ X_{12} & X_{22} & \cdots & X_{n2} \end{bmatrix} \begin{bmatrix} 1 & X_{11} & X_{12} \\ 1 & X_{21} & X_{22} \\ \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} \end{bmatrix}$$

or:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & \sum X_{i1} & \sum X_{i2} \\ \sum X_{i1} & \sum X_{i1}^2 & \sum X_{i1}X_{i2} \\ \sum X_{i2} & \sum X_{i2}X_{i1} & \sum X_{i2}^2 \end{bmatrix} \tag{6.74}$$

For the Dwaine Studios example, we have:

$$n = 21$$

$$\sum X_{i1} = 68.5 + 45.2 + \cdots = 1{,}302.4$$

$$\sum X_{i1}X_{i2} = 68.5(16.7) + 45.2(16.8) + \cdots = 22{,}609.2$$

etc.

These elements are found in (6.71).

Also note that $\mathbf{X}'\mathbf{Y}$ for the first-order regression model (6.69) with two predictor variables is:

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_{11} & X_{21} & \cdots & X_{n1} \\ X_{12} & X_{22} & \cdots & X_{n2} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \sum Y_i \\ \sum X_{i1}Y_i \\ \sum X_{i2}Y_i \end{bmatrix} \tag{6.75}$$

For the Dwaine Studios example, we have:

$$\sum Y_i = 174.4 + 164.4 + \cdots = 3{,}820$$

$$\sum X_{i1}Y_i = 68.5(174.4) + 45.2(164.4) + \cdots = 249{,}643$$

$$\sum X_{i2}Y_i = 16.7(174.4) + 16.8(164.4) + \cdots = 66{,}073$$

These are the elements found in (6.72).

Estimated Regression Function

The least squares estimates **b** are readily obtained by (6.25), using our basic calculations in (6.72) and (6.73):

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 29.7289 & .0722 & -1.9926 \\ .0722 & .00037 & -.0056 \\ -1.9926 & -.0056 & .1363 \end{bmatrix} \begin{bmatrix} 3{,}820 \\ 249{,}643 \\ 66{,}073 \end{bmatrix}$$

which yields:

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} -68.857 \\ 1.455 \\ 9.366 \end{bmatrix} \tag{6.76}$$

and the estimated regression function is:

$$\hat{Y} = -68.857 + 1.455X_1 + 9.366X_2$$

A three-dimensional plot of the estimated regression function, with the responses superimposed, is shown in Figure 6.7. The residuals are represented by the small vertical lines connecting the responses to the estimated regression surface.

This estimated regression function indicates that mean sales are expected to increase by 1.455 thousand dollars when the target population increases by 1 thousand persons aged 16 years or younger, holding per capita disposable personal income constant, and that mean sales are expected to increase by 9.366 thousand dollars when per capita income increases by 1 thousand dollars, holding the target population constant.

Figure 6.5a contains SYSTAT multiple regression output for the Dwaine Studios example. The estimated regression coefficients are shown in the column labeled COEFFICIENT; the output shows one more decimal place than we have given in the text.

The SYSTAT output also contains the inverse of the **X'X** matrix that we calculated earlier; only the lower portion of the symmetric matrix is shown. The results are the same as in (6.73).

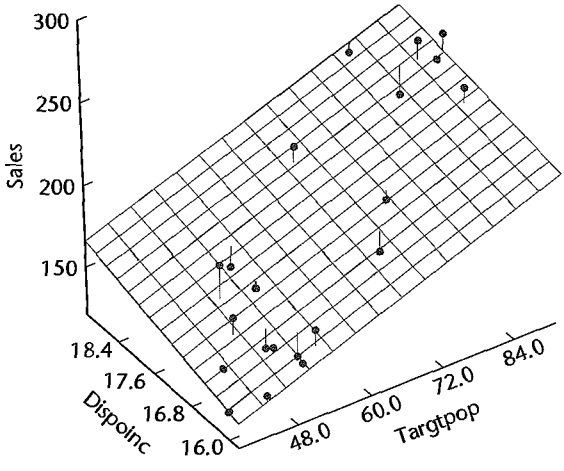


FIGURE 6.7
S-Plus Plot of Estimated Regression Surface—Dwaine Studios Example.

Algebraic Version of Normal Equations. The normal equations in algebraic form fo the case of two predictor variables can be obtained readily from (6.74) and (6.75). We have

$$(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{Y}$$

$$\begin{bmatrix} n & \sum X_{i1} & \sum X_{i2} \\ \sum X_{i1} & \sum X_{i1}^2 & \sum X_{i1}X_{i2} \\ \sum X_{i2} & \sum X_{i2}X_{i1} & \sum X_{i2}^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} \sum Y_i \\ \sum X_{i1}Y_i \\ \sum X_{i2}Y_i \end{bmatrix}$$

from which we obtain the normal equations:

$$\begin{aligned} \sum Y_i &= nb_0 + b_1 \sum X_{i1} + b_2 \sum X_{i2} \\ \sum X_{i1}Y_i &= b_0 \sum X_{i1} + b_1 \sum X_{i1}^2 + b_2 \sum X_{i1}X_{i2} \\ \sum X_{i2}Y_i &= b_0 \sum X_{i2} + b_1 \sum X_{i1}X_{i2} + b_2 \sum X_{i2}^2 \end{aligned} \qquad (6.77)$$

Fitted Values and Residuals

To examine the appropriateness of regression model (6.69) for the data at hand, we require the fitted values $\hat{Y}_i$ and the residuals $e_i = Y_i - \hat{Y}_i$. We obtain by (6.28):

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$$

$$\begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_{21} \end{bmatrix} = \begin{bmatrix} 1 & 68.5 & 16.7 \\ 1 & 45.2 & 16.8 \\ \vdots & \vdots & \vdots \\ 1 & 52.3 & 16.0 \end{bmatrix} \begin{bmatrix} -68.857 \\ 1.455 \\ 9.366 \end{bmatrix} = \begin{bmatrix} 187.2 \\ 154.2 \\ \vdots \\ 157.1 \end{bmatrix}$$

Further, by (6.29) we find:

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$$

$$\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_{21} \end{bmatrix} = \begin{bmatrix} 174.4 \\ 164.4 \\ \vdots \\ 166.5 \end{bmatrix} - \begin{bmatrix} 187.2 \\ 154.2 \\ \vdots \\ 157.1 \end{bmatrix} = \begin{bmatrix} -12.8 \\ 10.2 \\ \vdots \\ 9.4 \end{bmatrix}$$

Figure 6.5b shows the computer output for the fitted values and residuals to more decimal places than we have presented.

nalysis of Appropriateness of Model

We begin our analysis of the appropriateness of regression model (6.69) for the Dwaine Studios example by considering the plot of the residuals e against the fitted values $\hat{Y}$ in Figure 6.8a. This plot does not suggest any systematic deviations from the response plane,

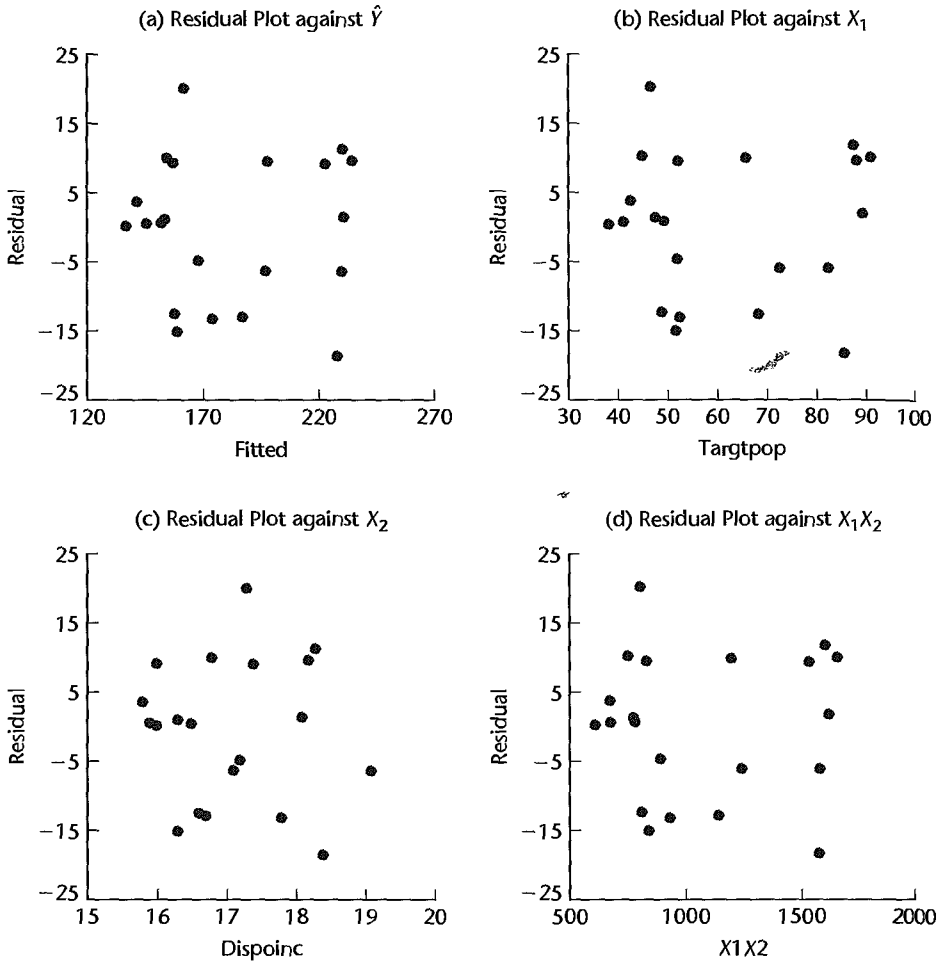


FIGURE 6.8 SYGRAPH Diagnostic Plots—Dwaine Studios Example.

nor that the variance of the error terms varies with the level of $\hat{Y}$. Plots of the residuals e against X_1 and X_2 in Figures 6.8b and 6.8c, respectively, are entirely consistent with the conclusions of good fit by the response function and constant variance of the error terms.

In multiple regression applications, there is frequently the possibility of interaction effects being present. To examine this for the Dwaine Studios example, we plotted the residuals e against the interaction term X_1X_2 in Figure 6.8d. A systematic pattern in this plot would suggest that an interaction effect may be present, so that a response function of the type:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

might be more appropriate. Figure 6.8d does not exhibit any systematic pattern; hence, no interaction effects reflected by the model term $\beta_3 X_1 X_2$ appear to be present.

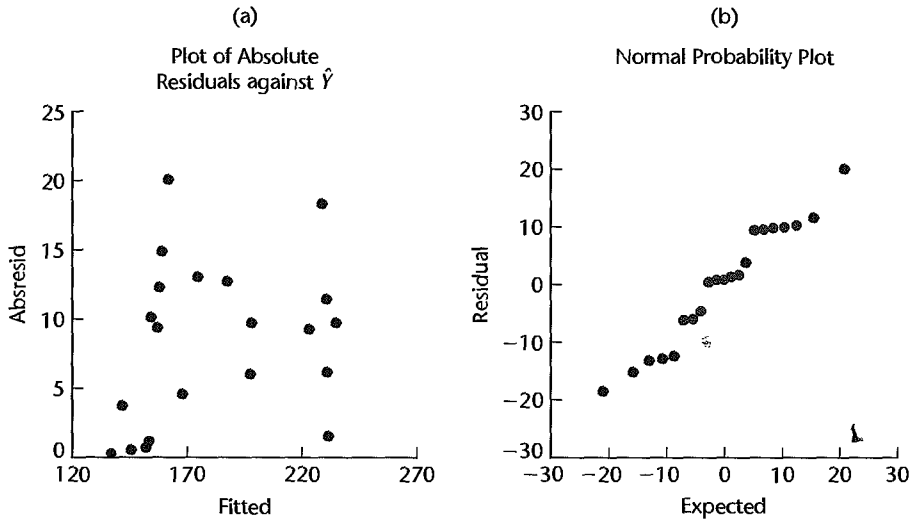


FIGURE 6.9 **Additional Diagnostic Plots—Dwaine Studios Example.**

Figure 6.9 contains two additional diagnostic plots. Figure 6.9a presents a plot of the absolute residuals against the fitted values. There is no indication of nonconstancy of the error variance. Figure 6.9b contains a normal probability plot of the residuals. The pattern is moderately linear. The coefficient of correlation between the ordered residuals and their expected values under normality is .980. This high value (the interpolated critical value in Table B.6 for $n = 21$ and $\alpha = .05$ is .9525) helps to confirm the reasonableness of the conclusion that the error terms are fairly normally distributed.

Since the Dwaine Studios data are cross-sectional and do not involve a time sequence, a time sequence plot is not relevant here. Thus, all of the diagnostics support the use of regression model (6.69) for the Dwaine Studios example.

Analysis of Variance

To test whether sales are related to target population and per capita disposable income, we require the ANOVA table. The basic quantities needed are:

$$\mathbf{Y'Y} = [174.4 \quad 164.4 \quad \cdots \quad 166.5] \begin{bmatrix} 174.4 \\ 164.4 \\ \vdots \\ 166.5 \end{bmatrix}$$

$$= 721{,}072.40$$

$$\left(\frac{1}{n}\right)\mathbf{Y'JY} = \frac{1}{21}[174.4 \quad 164.4 \quad \cdots \quad 166.5] \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} 174.4 \\ 164.4 \\ \vdots \\ 166.5 \end{bmatrix}$$

$$= \frac{(3{,}820.0)^2}{21} = 694{,}876.19$$

Thus:

$$SSTO = \mathbf{Y'Y} - \left(\frac{1}{n}\right)\mathbf{Y'JY} = 721{,}072.40 - 694{,}876.19 = 26{,}196.21$$

and, from our results in (6.72) and (6.76):

$$\begin{aligned} SSE &= \mathbf{Y'Y} - \mathbf{b'X'Y} \\ &= 721{,}072.40 - [-68.857 \quad 1.455 \quad 9.366] \begin{bmatrix} 3{,}820 \\ 249{,}643 \\ 66{,}073 \end{bmatrix} \\ &= 721{,}072.40 - 718{,}891.47 = 2{,}180.93 \end{aligned}$$

Finally, we obtain by subtraction:

$$SSR = SSTO - SSE = 26{,}196.21 - 2{,}180.93 = 24{,}015.28$$

These sums of squares are shown in the SYSTAT ANOVA table in Figure 6.5a. Also shown in the ANOVA table are degrees of freedom and mean squares. Note that three regression parameters had to be estimated; hence, $21 - 3 = 18$ degrees of freedom are associated with *SSE*. Also, the number of degrees of freedom associated with *SSR* is 2—the number of X variables in the model.

Test of Regression Relation. To test whether sales are related to target population and per capita disposable income:

$$H_0: \beta_1 = 0 \text{ and } \beta_2 = 0$$
$$H_a: \text{not both } \beta_1 \text{ and } \beta_2 \text{ equal zero}$$

we use test statistic (6.39b):

$$F^* = \frac{MSR}{MSE} = \frac{12{,}007.64}{121.1626} = 99.1$$

This test statistic is labeled F-RATIO in the SYSTAT output. For $\alpha = .05$, we require $F(.95; 2, 18) = 3.55$. Since $F^* = 99.1 > 3.55$, we conclude H_a, that sales are related to target population and per capita disposable income. The P-value for this test is .0000, as shown in the SYSTAT output labeled P.

Whether the regression relation is useful for making predictions of sales or estimates of mean sales still remains to be seen.

Coefficient of Multiple Determination. For our example, we have by (6.40):

$$R^2 = \frac{SSR}{SSTO} = \frac{24{,}015.28}{26{,}196.21} = .917$$

Thus, when the two predictor variables, target population and per capita disposable income, are considered, the variation in sales is reduced by 91.7 percent. The coefficient of multiple determination is shown in the SYSTAT output labeled SQUARED MULTIPLE R. Also shown in the output is the coefficient of multiple correlation $R = .957$ and the adjusted coefficient of multiple determination (6.42), $R_a^2 = .907$, which is labeled in the output

ADJUSTED SQUARED MULTIPLE R. Note that adjusting for the number of predictor variables in the model had only a small effect here on R^2.

Estimation of Regression Parameters

Dwaine Studios is not interested in the parameter β_0 since it falls far outside the scope of the model. It is desired to estimate β_1 and β_2 jointly with family confidence coefficient .90. We shall use the simultaneous Bonferroni confidence limits (6.52).

First, we need the estimated variance-covariance matrix $\mathbf{s}^2\{\mathbf{b}\}$:

$$\mathbf{s}^2\{\mathbf{b}\} = MSE(\mathbf{X}'\mathbf{X})^{-1}$$

MSE is given in Figure 6.5a, and $(\mathbf{X}'\mathbf{X})^{-1}$ was obtained in (6.73). Hence:

$$\mathbf{s}^2\{\mathbf{b}\} = 121.1626 \begin{bmatrix} 29.7289 & .0722 & -1.9926 \\ .0722 & .00037 & -.0056 \\ -1.9926 & -.0056 & .1363 \end{bmatrix} \tag{6.78}$$

$$= \begin{bmatrix} 3{,}602.0 & 8.748 & -241.43 \\ 8.748 & .0448 & -.679 \\ -241.43 & -.679 & 16.514 \end{bmatrix}$$

The two estimated variances we require are:

$$s^2\{b_1\} = .0448 \quad \text{or} \quad s\{b_1\} = .212$$
$$s^2\{b_2\} = 16.514 \quad \text{or} \quad s\{b_2\} = 4.06$$

These estimated standard deviations are shown in the SYSTAT output in Figure 6.5a, labeled STD ERROR, to four decimal places.

Next, we require for $g = 2$ simultaneous estimates:

$$B = t[1 - .10/2(2); 18] = t(.975; 18) = 2.101$$

The two pairs of simultaneous confidence limits therefore are $1.455 \pm 2.101(.212)$ and $9.366 \pm 2.101(4.06)$, which yield the confidence intervals:

$$1.01 \leq \beta_1 \leq 1.90$$
$$.84 \leq \beta_2 \leq 17.9$$

With family confidence coefficient .90, we conclude that β_1 falls between 1.01 and 1.90 and that β_2 falls between .84 and 17.9.

Note that the simultaneous confidence intervals suggest that both β_1 and β_2 are positive, which is in accord with theoretical expectations that sales should increase with higher target population and higher per capita disposable income, the other variable being held constant.

Estimation of Mean Response

Dwaine Studios would like to estimate expected (mean) sales in cities with target population $X_{h1} = 65.4$ thousand persons aged 16 years or younger and per capita disposable income

$X_{h2} = 17.6$ thousand dollars with a 95 percent confidence interval. We define:

$$\mathbf{X}_h = \begin{bmatrix} 1 \\ 65.4 \\ 17.6 \end{bmatrix}$$

The point estimate of mean sales is by (6.55):

$$\hat{Y}_h = \mathbf{X}_h'\mathbf{b} = [1 \quad 65.4 \quad 17.6] \begin{bmatrix} -68.857 \\ 1.455 \\ 9.366 \end{bmatrix} = 191.10$$

The estimated variance by (6.58), using the results in (6.78), is:

$$\begin{aligned} s^2\{\hat{Y}_h\} &= \mathbf{X}_h' s^2\{\mathbf{b}\}\mathbf{X}_h \\ &= [1 \quad 65.4 \quad 17.6] \begin{bmatrix} 3{,}602.0 & 8.748 & -241.43 \\ 8.748 & .0448 & -.679 \\ -241.43 & -.679 & 16.514 \end{bmatrix} \begin{bmatrix} 1 \\ 65.4 \\ 17.6 \end{bmatrix} \\ &= 7.656 \end{aligned}$$

or:

$$s\{\hat{Y}_h\} = 2.77$$

For confidence coefficient .95, we need $t(.975; 18) = 2.101$, and we obtain by (6.59) the confidence limits $191.10 \pm 2.101(2.77)$. The confidence interval for $E\{Y_h\}$ therefore is:

$$185.3 \leq E\{Y_h\} \leq 196.9$$

Thus, with confidence coefficient .95, we estimate that mean sales in cities with target population of 65.4 thousand persons aged 16 years or younger and per capita disposable income of 17.6 thousand dollars are somewhere between 185.3 and 196.9 thousand dollars. Dwaine Studios considers this confidence interval to provide information about expected (average) sales in communities of this size and income level that is precise enough for planning purposes.

Algebraic Version of Estimated Variance $s^2\{\hat{Y}_h\}$. Since by (6.58):

$$s^2\{\hat{Y}_h\} = \mathbf{X}_h' s^2\{\mathbf{b}\}\mathbf{X}_h$$

it follows for the case of two predictor variables in a first-order model:

$$\begin{aligned} s^2\{\hat{Y}_h\} = s^2\{b_0\} + X_{h1}^2 s^2\{b_1\} + X_{h2}^2 s^2\{b_2\} + 2X_{h1}s\{b_0, b_1\} \\ + 2X_{h2}s\{b_0, b_2\} + 2X_{h1}X_{h2}s\{b_1, b_2\} \end{aligned} \tag{6.79}$$

Prediction Limits for New Observations

Dwaine Studios as part of a possible expansion program would like to predict sales for two new cities, with the following characteristics:

	City A	City B
X_{h1}	65.4	53.1
X_{h2}	17.6	17.7

Prediction intervals with a 90 percent family confidence coefficient are desired. Note that the two new cities have characteristics that fall well within the pattern of the 21 cities on which the regression analysis is based.

To determine which simultaneous prediction intervals are best here, we find S as given in (6.65a) and B as given in (6.66a) for $g = 2$ and $1 - \alpha = .90$:

$$S^2 = 2F(.90; 2, 18) = 2(2.62) = 5.24 \qquad S = 2.29$$

and:

$$B = t[1 - .10/2(2); 18] = t(.975; 18) = 2.101$$

Hence, the Bonferroni limits are more efficient here.

For city A, we use the results obtained when estimating mean sales, since the levels of the predictor variables are the same here. We have from before:

$$\hat{Y}_h = 191.10 \qquad s^2\{\hat{Y}_h\} = 7.656 \qquad MSE = 121.1626$$

Hence, by (6.63a):

$$s^2\{\text{pred}\} = MSE + s^2\{\hat{Y}_h\} = 121.1626 + 7.656 = 128.82$$

or:

$$s\{\text{pred}\} = 11.35$$

In similar fashion, we obtain for city B (calculations not shown):

$$\hat{Y}_h = 174.15 \qquad s\{\text{pred}\} = 11.93$$

We previously found that the Bonferroni multiple is $B = 2.101$. Hence, by (6.66) the simultaneous Bonferroni prediction limits with family confidence coefficient .90 are $191.10 \pm 2.101(11.35)$ and $174.15 \pm 2.101(11.93)$, leading to the simultaneous prediction intervals:

$$\text{City A: } 167.3 \leq Y_{h(\text{new})} \leq 214.9$$
$$\text{City B: } 149.1 \leq Y_{h(\text{new})} \leq 199.2$$

With family confidence coefficient .90, we predict that sales in the two cities will be within the indicated limits. Dwaine Studios considers these prediction limits to be somewhat useful for planning purposes, but would prefer tighter intervals for predicting sales for a particular city. A consulting firm has been engaged to see if additional or alternative predictor variables can be found that will lead to tighter prediction intervals.

Note incidentally that even though the coefficient of multiple determination, $R^2 = .917$, is high, the prediction limits here are not fully satisfactory. This serves as another reminder that a high value of R^2 does not necessarily indicate that precise predictions can be made.

Cited Reference

6.1. Box, G. E. P., and P. W. Tidwell. "Transformations of the Independent Variables," *Technometrics* 4 (1962), pp. 531–50.

Problems

6.1. Set up the **X** matrix and $\boldsymbol{\beta}$ vector for each of the following regression models (assume $i = 1, \ldots, 4$):

a. $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1} X_{i2} + \varepsilon_i$

b. $\log Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$

6.2. Set up the **X** matrix and $\boldsymbol{\beta}$ vector for each of the following regression models (assume $i = 1, \ldots, 5$):

a. $Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}^2 + \varepsilon_i$

b. $\sqrt{Y_i} = \beta_0 + \beta_1 X_{i1} + \beta_2 \log_{10} X_{i2} + \varepsilon_i$

6.3. A student stated: "Adding predictor variables to a regression model can never reduce R^2, so we should include all available predictor variables in the model." Comment.

6.4. Why is it not meaningful to attach a sign to the coefficient of multiple correlation R, although we do so for the coefficient of simple correlation r_{12}?

6.5. **Brand preference.** In a small-scale experimental study of the relation between degree of brand liking (Y) and moisture content (X_1) and sweetness (X_2) of the product, the following results were obtained from the experiment based on a completely randomized design (data are coded):

i:	1	2	3	...	14	15	16
X_{i1}:	4	4	4	...	10	10	10
X_{i2}:	2	4	2	..	4	2	4
Y_i:	64	73	61	...	95	94	100

a. Obtain the scatter plot matrix and the correlation matrix. What information do these diagnostic aids provide here?

b. Fit regression model (6.1) to the data. State the estimated regression function. How is b_1 interpreted here?

c. Obtain the residuals and prepare a box plot of the residuals. What information does this plot provide?

d. Plot the residuals against $\hat{Y}$, X_1, X_2, and $X_1 X_2$ on separate graphs. Also prepare a normal probability plot. Interpret the plots and summarize your findings.

e. Conduct the Breusch-Pagan test for constancy of the error variance, assuming $\log \sigma_i^2 = \gamma_0 + \gamma_1 X_{i1} + \gamma_2 X_{i2}$; use $\alpha = .01$. State the alternatives, decision rule, and conclusion.

f. Conduct a formal test for lack of fit of the first-order regression function; use $\alpha = .01$. State the alternatives, decision rule, and conclusion.

6.6. Refer to **Brand preference** Problem 6.5. Assume that regression model (6.1) with independent normal error terms is appropriate.

a. Test whether there is a regression relation, using $\alpha = .01$. State the alternatives, decision rule, and conclusion. What does your test imply about β_1 and β_2?

b. What is the P-value of the test in part (a)?

c. Estimate β_1 and β_2 jointly by the Bonferroni procedure, using a 99 percent family confidence coefficient. Interpret your results.

6.7. Refer to **Brand preference** Problem 6.5.

a. Calculate the coefficient of multiple determination R^2. How is it interpreted here?

b. Calculate the coefficient of simple determination R^2 between Y_i and $\hat{Y}_i$. Does it equal the coefficient of multiple determination in part (a)?

6.8. Refer to **Brand preference** Problem 6.5. Assume that regression model (6.1) with independent normal error terms is appropriate.

a. Obtain an interval estimate of $E\{Y_h\}$ when $X_{h1} = 5$ and $X_{h2} = 4$. Use a 99 percent confidence coefficient. Interpret your interval estimate.

b. Obtain a prediction interval for a new observation $Y_{h(\text{new})}$ when $X_{h1} = 5$ and $X_{h2} = 4$. Use a 99 percent confidence coefficient.

*6.9. **Grocery retailer.** A large, national grocery retailer tracks productivity and costs of its facilities closely. Data below were obtained from a single distribution center for a one-year period. Each data point for each variable represents one week of activity. The variables included are the number of cases shipped (X_1), the indirect costs of the total labor hours as a percentage (X_2), a qualitative predictor called holiday that is coded 1 if the week has a holiday and 0 otherwise (X_3), and the total labor hours (Y).

i:	1	2	3	...	50	51	52
X_{i1}:	305,657	328,476	317,164	...	290,455	411,750	292,087
X_{i2}:	7.17	6.20	4.61	...	7.99	7.83	7.77
X_{i3}:	0	0	0	...	0	0	0
Y_i:	4264	4496	4317	...	4499	4186	4342

a. Prepare separate stem-and-leaf plots for the number of cases shipped X_{i1} and the indirect cost of the total hours X_{i2}. Are there any outlying cases present? Are there any gaps in the data?

b. The cases are given in consecutive weeks. Prepare a time plot for each predictor variable. What do the plots show?

c. Obtain the scatter plot matrix and the correlation matrix. What information do these diagnostic aids provide here?

*6.10. Refer to **Grocery retailer** Problem 6.9.

a. Fit regression model (6.5) to the data for three predictor variables. State the estimated regression function. How are b_1, b_2, and b_3 interpreted here?

b. Obtain the residuals and prepare a box plot of the residuals. What information does this plot provide?

c. Plot the residuals against $\hat{Y}$, X_1, X_2, X_3, and X_1X_2 on separate graphs. Also prepare a normal probability plot. Interpret the plots and summarize your findings.

d. Prepare a time plot of the residuals. Is there any indication that the error terms are correlated? Discuss.

e. Divide the 52 cases into two groups, placing the 26 cases with the smallest fitted values $\hat{Y}_i$ into group 1 and the other 26 cases into group 2. Conduct the Brown-Forsythe test for constancy of the error variance, using $\alpha = .01$. State the decision rule and conclusion.

*6.11. Refer to **Grocery retailer** Problem 6.9. Assume that regression model (6.5) for three predictor variables with independent normal error terms is appropriate.

a. Test whether there is a regression relation, using level of significance .05. State the alternatives, decision rule, and conclusion. What does your test result imply about β_1, β_2, and β_3? What is the P-value of the test?

b. Estimate β_1 and β_3 jointly by the Bonferroni procedure, using a 95 percent family confidence coefficient. Interpret your results.

c. Calculate the coefficient of multiple determination R^2. How is this measure interpreted here?

*6.12. Refer to **Grocery retailer** Problem 6.9. Assume that regression model (6.5) for three predictor variables with independent normal error terms is appropriate.

a. Management desires simultaneous interval estimates of the total labor hours for the following five typical weekly shipments:

	1	2	3	4	5
X_1:	302,000	245,000	280,000	350,000	295,000
X_2:	7.20	7.40	6.90	7.00	6.70
X_3:	0	0	0	0	1

Obtain the family of estimates using a 95 percent family confidence coefficient. Employ the Working-Hotelling or the Bonferroni procedure, whichever is more efficient.

b. For the data in Problem 6.9 on which the regression fit is based, would you consider a shipment of 400,000 cases with an indirect percentage of 7.20 on a nonholiday week to be within the scope of the model? What about a shipment of 400,000 cases with an indirect percentage of 9.9 on a nonholiday week? Support your answers by preparing a relevant plot.

*6.13. Refer to **Grocery retailer** Problem 6.9. Assume that regression model (6.5) for three predictor variables with independent normal error terms is appropriate. Four separate shipments with the following characteristics must be processed next month:

	1	2	3	4
X_1:	230,000	250,000	280,000	340,000
X_2:	7.50	7.30	7.10	6.90
X_3:	0	0	0	0

Management desires predictions of the handling times for these shipments so that the actual handling times can be compared with the predicted times to determine whether any are out of line. Develop the needed predictions, using the most efficient approach and a family confidence coefficient of 95 percent.

*6.14. Refer to **Grocery retailer** Problem 6.9. Assume that regression model (6.5) for three predictor variables with independent normal error terms is appropriate. Three new shipments are to be received, each with $X_{h1} = 282{,}000$, $X_{h2} = 7.10$, and $X_{h3} = 0$.

a. Obtain a 95 percent prediction interval for the mean handling time for these shipments.

b. Convert the interval obtained in part (a) into a 95 percent prediction interval for the total labor hours for the three shipments.

*6.15. **Patient satisfaction.** A hospital administrator wished to study the relation between patient satisfaction (Y) and patient's age (X_1, in years), severity of illness (X_2, an index), and anxiety

level (X_3, an index). The administrator randomly selected 46 patients and collected the data presented below, where larger values of Y, X_2, and X_3 are, respectively, associated with more satisfaction, increased severity of illness, and more anxiety.

i:	1	2	3	...	44	45	46
X_{i1}:	50	36	40	...	45	37	28
X_{i2}:	51	46	48	...	51	53	46
X_{i3}:	2.3	2.3	2.2	...	2.2	2.1	1.8
Y_i:	48	57	66	...	68	59	92

a. Prepare a stem-and-leaf plot for each of the predictor variables. Are any noteworthy features revealed by these plots?

b. Obtain the scatter plot matrix and the correlation matrix. Interpret these and state your principal findings.

c. Fit regression model (6.5) for three predictor variables to the data and state the estimated regression function. How is b_2 interpreted here?

d. Obtain the residuals and prepare a box plot of the residuals. Do there appear to be any outliers?

e. Plot the residuals against $\hat{Y}$, each of the predictor variables, and each two-factor interaction term on separate graphs. Also prepare a normal probability plot. Interpret your plots and summarize your findings.

f. Can you conduct a formal test for lack of fit here?

g. Conduct the Breusch-Pagan test for constancy of the error variance, assuming $\log \sigma_i^2 = \gamma_0 + \gamma_1 X_{i1} + \gamma_2 X_{i2} + \gamma_3 X_{i3}$; use $\alpha = .01$. State the alternatives, decision rule, and conclusion.

*6.16. Refer to **Patient satisfaction** Problem 6.15. Assume that regression model (6.5) for three predictor variables with independent normal error terms is appropriate.

a. Test whether there is a regression relation; use $\alpha = .10$. State the alternatives, decision rule, and conclusion. What does your test imply about β_1, β_2, and β_3? What is the P-value of the test?

b. Obtain joint interval estimates of β_1, β_2, and β_3, using a 90 percent family confidence coefficient. Interpret your results.

c. Calculate the coefficient of multiple determination. What does it indicate here?

*6.17. Refer to **Patient satisfaction** Problem 6.15. Assume that regression model (6.5) for three predictor variables with independent normal error terms is appropriate.

a. Obtain an interval estimate of the mean satisfaction when $X_{h1} = 35$, $X_{h2} = 45$, and $X_{h3} = 2.2$. Use a 90 percent confidence coefficient. Interpret your confidence interval.

b. Obtain a prediction interval for a new patient's satisfaction when $X_{h1} = 35$, $X_{h2} = 45$, and $X_{h3} = 2.2$. Use a 90 percent confidence coefficient. Interpret your prediction interval.

6.18. **Commercial properties.** A commercial real estate company evaluates vacancy rates, square footage, rental rates, and operating expenses for commercial properties in a large metropolitan area in order to provide clients with quantitative information upon which to make rental decisions. The data below are taken from 81 suburban commercial properties that are the newest, best located, most attractive, and expensive for five specific geographic areas. Shown here are

the age (X_1), operating expenses and taxes (X_2), vacancy rates (X_3), total square footage (X_4), and rental rates (Y).

i:	1	2	3	...	79	80	81
X_{i1}:	1	14	16	...	15	11	14
X_{i2}:	5.02	8.19	3.00	...	11.97	11.27	12.68
X_{i3}:	0.14	0.27	0	...	0.14	0.03	0.03
X_{i4}:	123,000	104,079	39,998	...	254,700	434,746	201,930
Y_i:	13.50	12.00	10.50	...	15.00	15.25	14.50

a. Prepare a stem-and-leaf plot for each predictor variable. What information do these plots provide?

b. Obtain the scatter plot matrix and the correlation matrix. Interpret these and state your principal findings.

c. Fit regression model (6.5) for four predictor variables to the data. State the estimated regression function.

d. Obtain the residuals and prepare a box plot of the residuals. Does the distribution appear to be fairly symmetrical?

e. Plot the residuals against $\hat{Y}$, each predictor variable, and each two-factor interaction term on separate graphs. Also prepare a normal probability plot. Analyze your plots and summarize your findings.

f. Can you conduct a formal test for lack of fit here?

g. Divide the 81 cases into two groups, placing the 40 cases with the smallest fitted values $\hat{Y}_i$ into group 1 and the remaining cases into group 2. Conduct the Brown-Forsythe test for constancy of the error variance, using $\alpha = .05$. State the decision rule and conclusion.

6.19. Refer to **Commercial properties** Problem 6.18. Assume that regression model (6.5) for four predictor variables with independent normal error terms is appropriate.

a. Test whether there is a regression relation; use $\alpha = .05$. State the alternatives, decision rule, and conclusion. What does your test imply about β_1, β_2, β_3, and β_4? What is the P-value of the test?

b. Estimate β_1, β_2, β_3, and β_4 jointly by the Bonferroni procedure, using a 95 percent family confidence coefficient. Interpret your results.

c. Calculate R^2 and interpret this measure.

6.20. Refer to **Commercial properties** Problem 6.18. Assume that regression model (6.5) for four predictor variables with independent normal error terms is appropriate. The researcher wishes to obtain simultaneous interval estimates of the mean rental rates for four typical properties specified as follows:

	1	2	3	4
X_1:	5.0	6.0	14.0	12.0
X_2:	8.25	8.50	11.50	10.25
X_3:	0	0.23	0.11	0
X_4:	250,000	270,000	300,000	310,000

Obtain the family of estimates using a 95 percent family confidence coefficient. Employ the most efficient procedure.

6.21. Refer to **Commercial properties** Problem 6.18. Assume that regression model (6.5) for four predictor variables with independent normal error terms is appropriate. Three properties with the following characteristics did not have any rental information available.

	1	2	3
X_1:	4.0	6.0	12.0
X_2:	10.0	11.5	12.5
X_3:	0.10	0	0.32
X_4:	80,000	120,000	340,000

Develop separate prediction intervals for the rental rates of these properties, using a 95 percent statement confidence coefficient in each case. Can the rental rates of these three properties be predicted fairly precisely? What is the family confidence level for the set of three predictions?

Exercises

6.22. For each of the following regression models, indicate whether it is a general linear regression model. If it is not, state whether it can be expressed in the form of (6.7) by a suitable transformation:

a. $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 \log_{10} X_{i2} + \beta_3 X_{i1}^2 + \varepsilon_i$

b. $Y_i = \varepsilon_i \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}^2)$

c. $Y_i = \log_{10}(\beta_1 X_{i1}) + \beta_2 X_{i2} + \varepsilon_i$

d. $Y_i = \beta_0 \exp(\beta_1 X_{i1}) + \varepsilon_i$

e. $Y_i = [1 + \exp(\beta_0 + \beta_1 X_{i1} + \varepsilon_i)]^{-1}$

6.23. (Calculus needed.) Consider the multiple regression model:

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \qquad i = 1, \ldots, n$$

where the ε_i are uncorrelated, with $E\{\varepsilon_i\} = 0$ and $\sigma^2\{\varepsilon_i\} = \sigma^2$.

a. State the least squares criterion and derive the least squares estimators of β_1 and β_2.

b. Assuming that the ε_i are independent normal random variables, state the likelihood function and obtain the maximum likelihood estimators of β_1 and β_2. Are these the same as the least squares estimators?

6.24. (Calculus needed.) Consider the multiple regression model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2 + \beta_3 X_{i2} + \varepsilon_i \qquad i = 1, \ldots, n$$

where the ε_i are independent $N(0, \sigma^2)$.

a. State the least squares criterion and derive the least squares normal equations.

b. State the likelihood function and explain why the maximum likelihood estimators will be the same as the least squares estimators.

6.25. An analyst wanted to fit the regression model $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$, $i = 1, \ldots, n$, by the method of least squares when it is known that $\beta_2 = 4$. How can the analyst obtain the desired fit by using a multiple regression computer program?

6.26. For regression model (6.1), show that the coefficient of simple determination between Y_i and $\hat{Y}_i$ equals the coefficient of multiple determination R^2.

6.27. In a small-scale regression study, the following data were obtained:

i:	1	2	3	4	5	6
X_{i1}:	7	4	16	3	21	8
X_{i2}:	33	41	7	49	5	31
Y_i:	42	33	75	28	91	55

Assume that regression model (6.1) with independent normal error terms is appropriate. Using matrix methods, obtain (a) $\mathbf{b}$; (b) $\mathbf{e}$; (c) $\mathbf{H}$; (d) SSR; (e) $\mathbf{s}^2\{\mathbf{b}\}$; (f) $\hat{Y}_h$ when $X_{h1} = 10$, $X_{h2} = 30$; (g) $s^2\{\hat{Y}_h\}$ when $X_{h1} = 10$, $X_{h2} = 30$.

Projects

6.28. Refer to the **CDI** data set in Appendix C.2. You have been asked to evaluate two alternative models for predicting the number of active physicians (Y) in a CDI. Proposed model I includes as predictor variables total population (X_1), land area (X_2), and total personal income (X_3). Proposed model II includes as predictor variables population density (X_1, total population divided by land area), percent of population greater than 64 years old (X_2), and total personal income (X_3).

a. Prepare a stem-and-leaf plot for each of the predictor variables. What noteworthy information is provided by your plots?

b. Obtain the scatter plot matrix and the correlation matrix for each proposed model. Summarize the information provided.

c. For each proposed model, fit the first-order regression model (6.5) with three predictor variables.

d. Calculate R^2 for each model. Is one model clearly preferable in terms of this measure?

e. For each model, obtain the residuals and plot them against $\hat{Y}$, each of the three predictor variables, and each of the two-factor interaction terms. Also prepare a normal probability plot for each of the two fitted models. Interpret your plots and state your findings. Is one model clearly preferable in terms of appropriateness?

6.29. Refer to the **CDI** data set in Appendix C.2.

a. For each geographic region, regress the number of serious crimes in a CDI (Y) against population density (X_1, total population divided by land area), per capita personal income (X_2), and percent high school graduates (X_3). Use first-order regression model (6.5) with three predictor variables. State the estimated regression functions.

b. Are the estimated regression functions similar for the four regions? Discuss.

c. Calculate MSE and R^2 for each region. Are these measures similar for the four regions? Discuss.

d. Obtain the residuals for each fitted model and prepare a box plot of the residuals for each fitted model. Interpret your plots and state your findings.

6.30. Refer to the **SENIC** data set in Appendix C.1. Two models have been proposed for predicting the average length of patient stay in a hospital (Y). Model I utilizes as predictor variables age (X_1), infection risk (X_2), and available facilities and services (X_3). Model II uses as predictor variables number of beds (X_1), infection risk (X_2), and available facilities and services (X_3).

a. Prepare a stem-and-leaf plot for each of the predictor variables. What information do these plots provide?

b. Obtain the scatter plot matrix and the correlation matrix for each proposed model. Interpret these and state your principal findings.

c. For each of the two proposed models, fit first-order regression model (6.5) with three predictor variables.

d. Calculate R^2 for each model. Is one model clearly preferable in terms of this measure?

e. For each model, obtain the residuals and plot them against $\hat{Y}$, each of the three predictor variables, and each of the two-factor interaction terms. Also prepare a normal probability plot of the residuals for each of the two fitted models. Interpret your plots and state your findings. Is one model clearly more appropriate than the other?

6.31. Refer to the **SENIC** data set in Appendix C.1.

a. For each geographic region, regress infection risk (Y) against the predictor variables age (X_1), routine culturing ratio (X_2), average daily census (X_3), and available facilities and services (X_4). Use first-order regression model (6.5) with four predictor variables. State the estimated regression functions.

b. Are the estimated regression functions similar for the four regions? Discuss.

c. Calculate *MSE* and R^2 for each region. Are these measures similar for the four regions? Discuss.

d. Obtain the residuals for each fitted model and prepare a box plot of the residuals for each fitted model. Interpret the plots and state your findings.