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

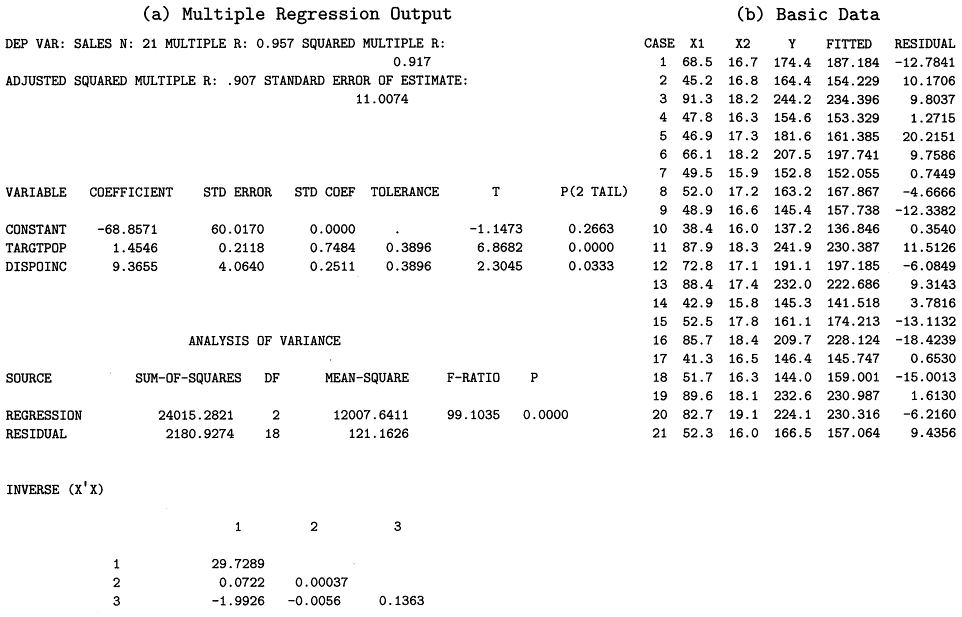


FIGURE 6.5 SYSTAT Multiple Regression Output and Basic Data—Dwaine Studios Example.

```
               (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.

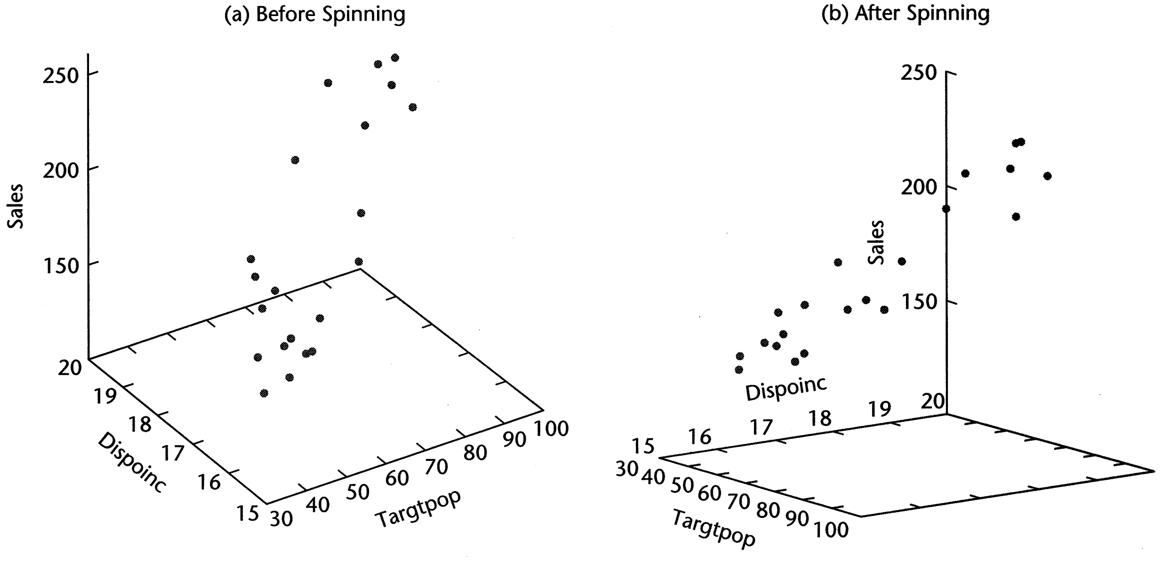
A SYGRAPH plot of the point cloud is shown in Figure 6.6a. By spinning the axes, we obtain the perspective in Figure 6.6b which supports the tentative conclusion that a response plane may be a reasonable regression function to utilize here.

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 Example.**


(a) Before Spinning

(b) After Spinning

We require:

1.

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

which yields:

$$\mathbf{X'X} = \begin{bmatrix} 21.0 & 1{,}302.4 & 360.0 \\ 1{,}302.4 & 87{,}707.9 & 22{,}609.2 \\ 360.0 & 22{,}609.2 & 6{,}190.3 \end{bmatrix} \tag{6.71}$$

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:

$$\mathbf{X'Y} = \begin{bmatrix} 3{,}820 \\ 249{,}643 \\ 66{,}073 \end{bmatrix} \tag{6.72}$$

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:

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

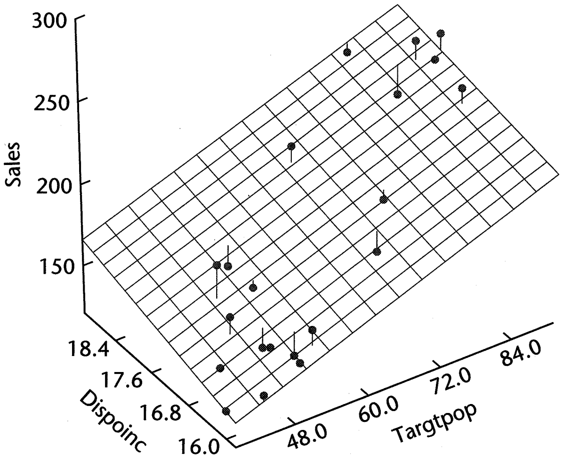
A three-dimensional plot of the estimated regression function, with the responses superimposed, is shown in Figure 6.7. The residuals are represented by the small vertical lines connecting the responses to the estimated regression surface.

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 for 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:

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

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:

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

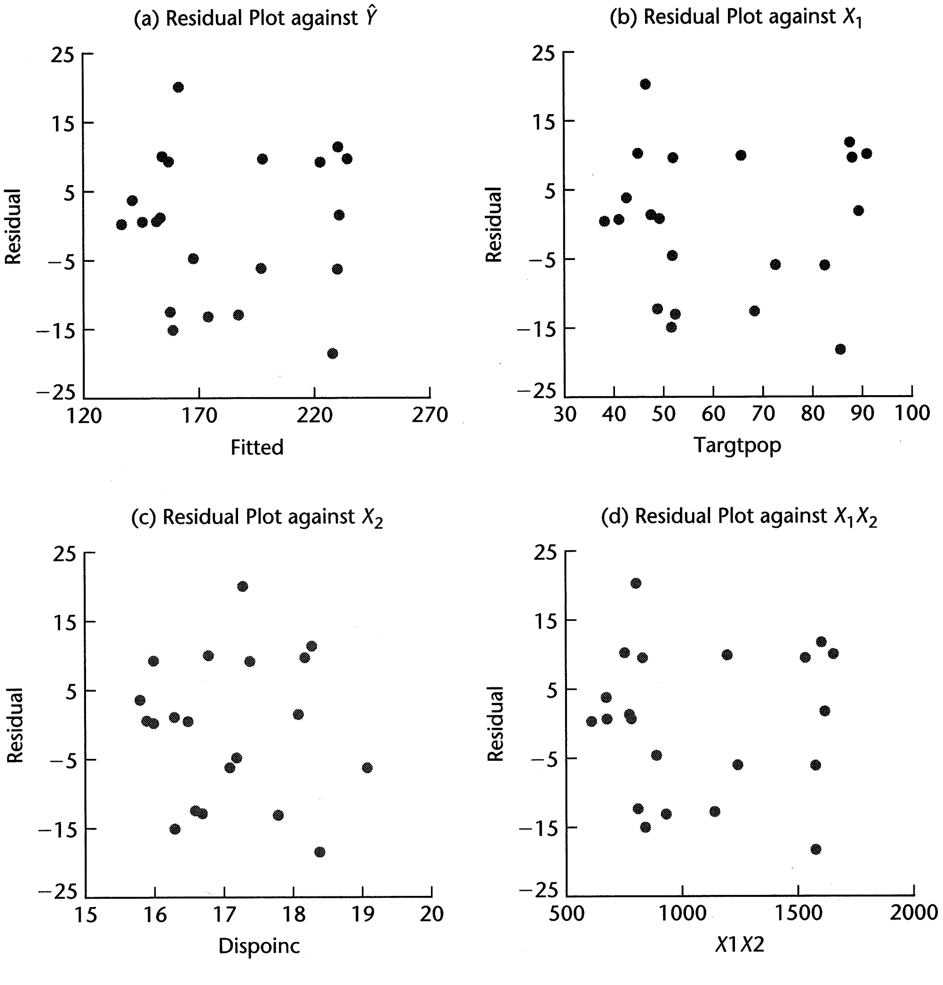
$$\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_{21} \end{bmatrix} = \begin{bmatrix} 174.4 \\ 164.4 \\ \vdots \\ 166.5 \end{bmatrix} - \begin{bmatrix} 187.2 \\ 154.2 \\ \vdots \\ 157.1 \end{bmatrix} = \begin{bmatrix} -12.8 \\ 10.2 \\ \vdots \\ 9.4 \end{bmatrix}$$

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

Analysis 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.

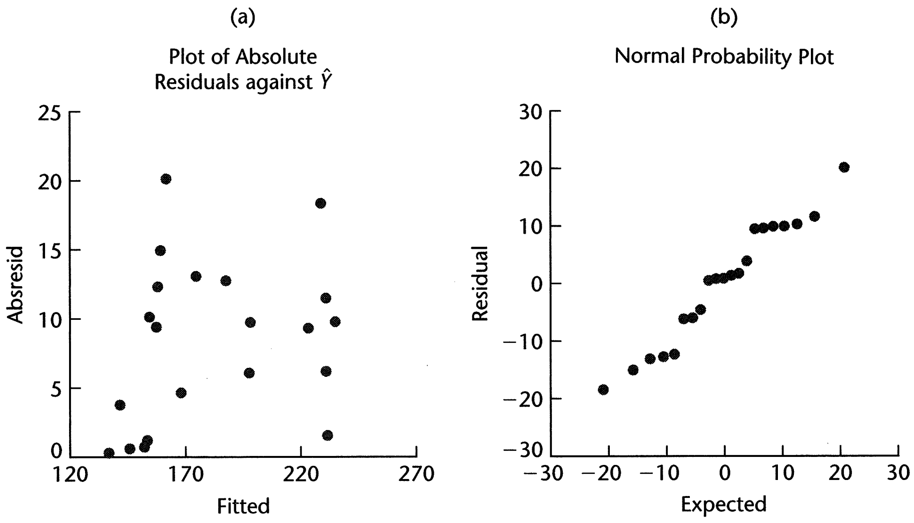


Figure 6.9 contains two additional diagnostic plots. Figure 6.9a presents a plot of the absolute residuals against the fitted values. There is no indication of nonconstancy of the error variance. Figure 6.9b contains a normal probability plot of the residuals. The pattern is moderately linear. The coefficient 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:

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

$$= 721{,}072.40$$

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

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

Thus:

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

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

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

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:

$$\begin{aligned} &H_0\colon \beta_1 = 0 \text{ and } \beta_2 = 0 \\ &H_a\colon \text{not both } \beta_1 \text{ and } \beta_2 \text{ equal zero} \end{aligned}$$

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:

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

$$= \begin{bmatrix} 3{,}602.0 & 8.748 & -241.43 \\ 8.748 & .0448 & -.679 \\ -241.43 & -.679 & 16.514 \end{bmatrix} \tag{6.78}$$

The two estimated variances we require are:

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

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:

$$\begin{aligned} 1.01 \leq &\beta_1 \leq 1.90 \\ .84 \leq &\beta_2 \leq 17.9 \end{aligned}$$

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:

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

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):

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

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.