# CHAPTER 16

# Regression Analysis: Model Building

CONTENTS

## STATISTICS *in* PRACTICE

### MONSANTO COMPANY*

*ST. LOUIS, MISSOURI*

Monsanto Company traces its roots to one entrepreneur's investment of $500 and a dusty warehouse on the Mississippi riverfront, where in 1901 John F. Queeney began manufacturing saccharin. Today, Monsanto is one of the nation's largest chemical companies, producing more than a thousand products ranging from industrial chemicals to synthetic playing surfaces used in modern sports stadiums. Monsanto is a worldwide corporation with manufacturing facilities, laboratories, technical centers, and marketing operations in 65 countries.

Monsanto's Nutrition Chemical Division manufactures and markets a methionine supplement used in poultry, swine, and cattle feed products. Because poultry growers work with high volumes and low profit margins, cost-effective poultry feed products with the best possible nutrition value are needed. Optimal feed composition will result in rapid growth and high final body weight for a given level of feed intake. The chemical industry works closely with poultry growers to optimize poultry feed products. Ultimately, success depends on keeping the cost of poultry low in comparison with the cost of beef and other meat products.

Monsanto used regression analysis to model the relationship between body weight $y$ and the amount of methionine $x$ added to the poultry feed. Initially, the following simple linear estimated regression equation was developed.

$$\hat{y} = .21 + .42x$$

This estimated regression equation proved statistically significant; however, the analysis of the residuals indicated that a curvilinear relationship would be a better model of the relationship between body weight and methionine.

Monsanto researchers used regression analysis to develop an optimal feed composition for poultry growers. © Krugloff/Shutterstock.com.

Further research conducted by Monsanto showed that although small amounts of methionine tended to increase body weight, at some point body weight leveled off and additional amounts of the methionine were of little or no benefit. In fact, when the amount of methionine increased beyond nutritional requirements, body weight tended to decline. The following estimated multiple regression equation was used to model the curvilinear relationship between body weight and methionine.

$$\hat{y} = -1.89 + 1.32x - .506x^2$$

Use of the regression results enabled Monsanto to determine the optimal level of methionine to be used in poultry feed products.

In this chapter we will extend the discussion of regression analysis by showing how curvilinear models such as the one used by Monsanto can be developed. In addition, we will describe a variety of tools that help determine which independent variables lead to the best estimated regression equation.

*The authors are indebted to James R. Ryland and Robert M. Schisla, Senior Research Specialists, Monsanto Nutrition Chemical Division, for providing this Statistics in Practice.

Model building is the process of developing an estimated regression equation that describes the relationship between a dependent variable and one or more independent variables. The major issues in model building are finding the proper functional form of the relationship and selecting the independent variables to be included in the model. In Section 16.1 we establish the framework for model building by introducing the concept of a general linear model. Section 16.2, which provides the foundation for the more sophisticated computer-based procedures, introduces a general approach for determining when to add or delete independent variables. In Section 16.3 we consider a larger regression problem involving eight independent variables and 25 observations; this problem is used to illustrate the variable selection procedures presented in Section 16.4, including stepwise regression, the forward selection procedure, the backward elimination procedure, and best-subsets regression. In Section 16.5 we show how multiple regression analysis can provide another approach to solving experimental design problems, and in Section 16.6 we show how the Durbin-Watson test can be used to detect serial or autocorrelation.

## 16.1 General Linear Model

Suppose we collected data for one dependent variable $y$ and $k$ independent variables $x_1$, $x_2, \ldots, x_k$. Our objective is to use these data to develop an estimated regression equation that provides the best relationship between the dependent and independent variables. As a general framework for developing more complex relationships among the independent variables, we introduce the concept of a **general linear model** involving $p$ independent variables.

*If you can write a regression model in the form of equation (16.1), the standard multiple regression procedures described in Chapter 15 are applicable.*

GENERAL LINEAR MODEL

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_p z_p + \epsilon \tag{16.1}$$

In equation (16.1), each of the independent variables $z_j$ (where $j = 1, 2, \ldots, p$) is a function of $x_1, x_2, \ldots, x_k$ (the variables for which data are collected). In some cases, each $z_j$ may be a function of only one $x$ variable. The simplest case is when we collect data for just one variable $x_1$ and want to predict $y$ by using a straight-line relationship. In this case $z_1 = x_1$ and equation (16.1) becomes

$$y = \beta_0 + \beta_1 x_1 + \epsilon \tag{16.2}$$

Equation (16.2) is the simple linear regression model introduced in Chapter 14 with the exception that the independent variable is labeled $x_1$ instead of $x$. In the statistical modeling literature, this model is called a *simple first-order model with one predictor variable.*

### Modeling Curvilinear Relationships

More complex types of relationships can be modeled with equation (16.1). To illustrate, let us consider the problem facing Reynolds, Inc., a manufacturer of industrial scales and laboratory equipment. Managers at Reynolds want to investigate the relationship between length of employment of their salespeople and the number of electronic laboratory scales sold. Table 16.1 gives the number of scales sold by 15 randomly selected salespeople for the most recent sales period and the number of months each salesperson has been employed by the firm. Figure 16.1 is the scatter diagram for these data. The scatter diagram indicates a possible curvilinear relationship between the length of time employed and the number of units sold. Before considering how to develop a curvilinear relationship for Reynolds, let us consider the Minitab output in Figure 16.2 corresponding to a simple first-order model; the estimated regression is

$$\text{Sales} = 111 + 2.38 \text{ Months}$$

where

$$\text{Sales} = \text{number of electronic laboratory scales sold}$$
$$\text{Months} = \text{the number of months the salesperson has been employed}$$

**TABLE 16.1**

DATA FOR THE REYNOLDS EXAMPLE

| Months Employed | Scales Sold |
|---|---|
| 41 | 275 |
| 106 | 296 |
| 76 | 317 |
| 104 | 376 |
| 22 | 162 |
| 12 | 150 |
| 85 | 367 |
| 111 | 308 |
| 40 | 189 |
| 51 | 235 |
| 9 | 83 |
| 12 | 112 |
| 6 | 67 |
| 56 | 325 |
| 19 | 189 |

WEB file

Reynolds

FIGURE 16.1 SCATTER DIAGRAM FOR THE REYNOLDS EXAMPLE

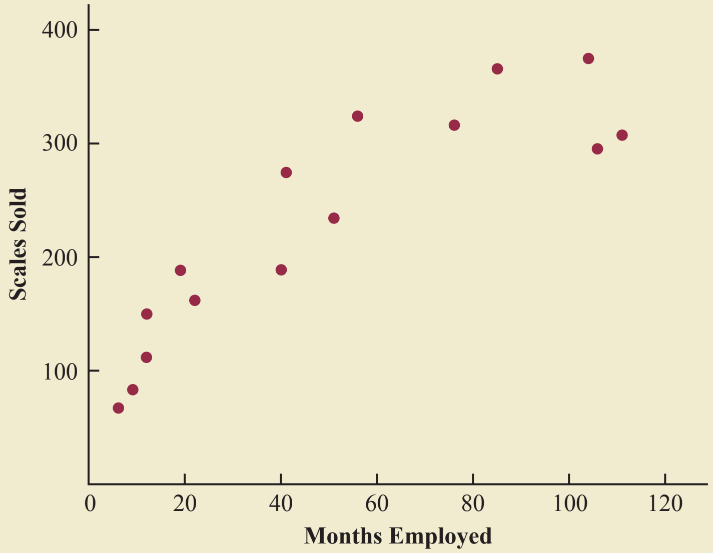


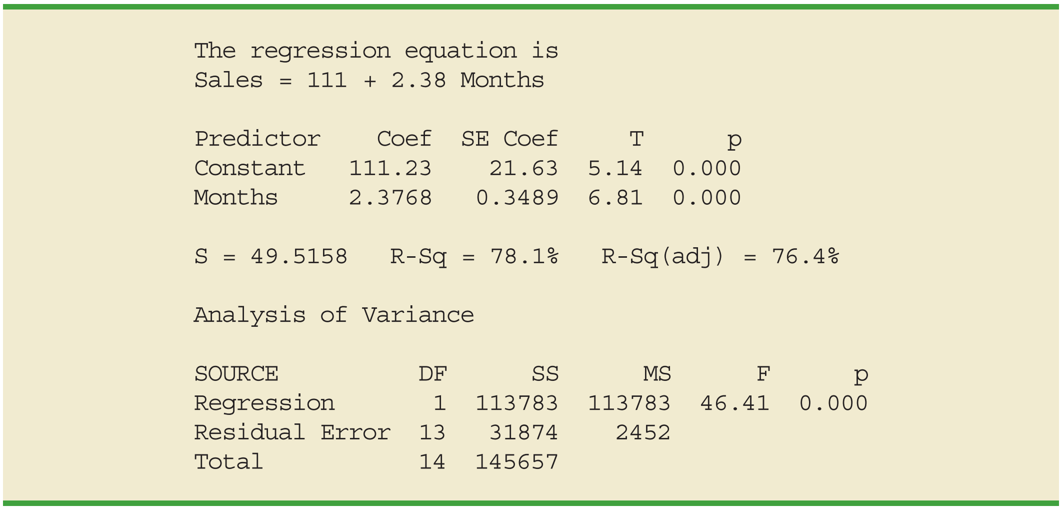
Figure 16.3 is the corresponding standardized residual plot. Although the computer output shows that the relationship is significant ($p$-value = .000) and that a linear relationship explains a high percentage of the variability in sales (R-Sq = 78.1%), the standardized residual plot suggests that a curvilinear relationship is needed.

To account for the curvilinear relationship, we set $z_1 = x_1$ and $z_2 = x_1^2$ in equation (16.1) to obtain the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \epsilon \tag{16.3}$$

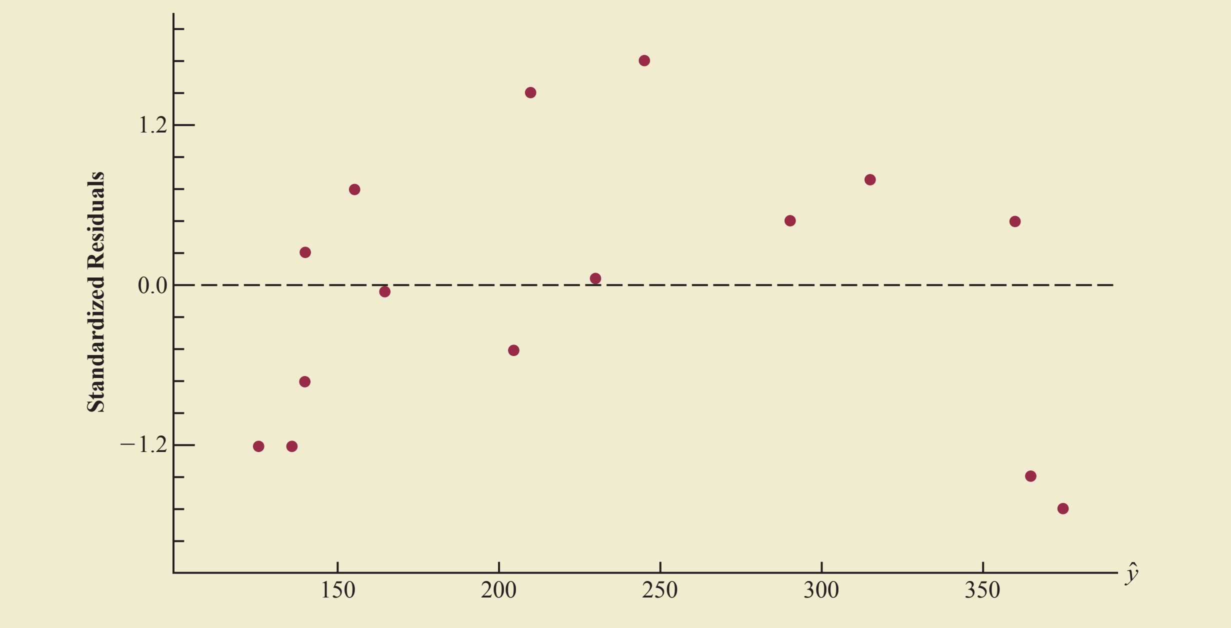
This model is called a *second-order model with one predictor variable.* To develop an estimated regression equation corresponding to this second-order model, the statistical

FIGURE 16.2 MINITAB OUTPUT FOR THE REYNOLDS EXAMPLE: FIRST-ORDER MODEL



```
The regression equation is
Sales = 111 + 2.38 Months

Predictor     Coef  SE Coef     T      p
Constant    111.23    21.63  5.14  0.000
Months      2.3768   0.3489  6.81  0.000

S = 49.5158   R-Sq = 78.1%   R-Sq(adj) = 76.4%

Analysis of Variance

SOURCE          DF      SS      MS      F      p
Regression       1  113783  113783  46.41  0.000
Residual Error  13   31874    2452
Total           14  145657
```

FIGURE 16.3 STANDARDIZED RESIDUAL PLOT FOR THE REYNOLDS EXAMPLE: FIRST-ORDER MODEL



software package we are using needs the original data in Table 16.1, as well as that data corresponding to adding a second independent variable that is the square of the number of months the employee has been with the firm. In Figure 16.4 we show the Minitab output corresponding to the second-order model; the estimated regression equation is

*The data for the* MonthsSq *independent variable is obtained by squaring the values of* Months.

$$\text{Sales} = 45.3 + 6.34 \text{ Months} - .0345 \text{ MonthsSq}$$

where

$$\text{MonthsSq} = \text{the square of the number of months the salesperson has been employed}$$

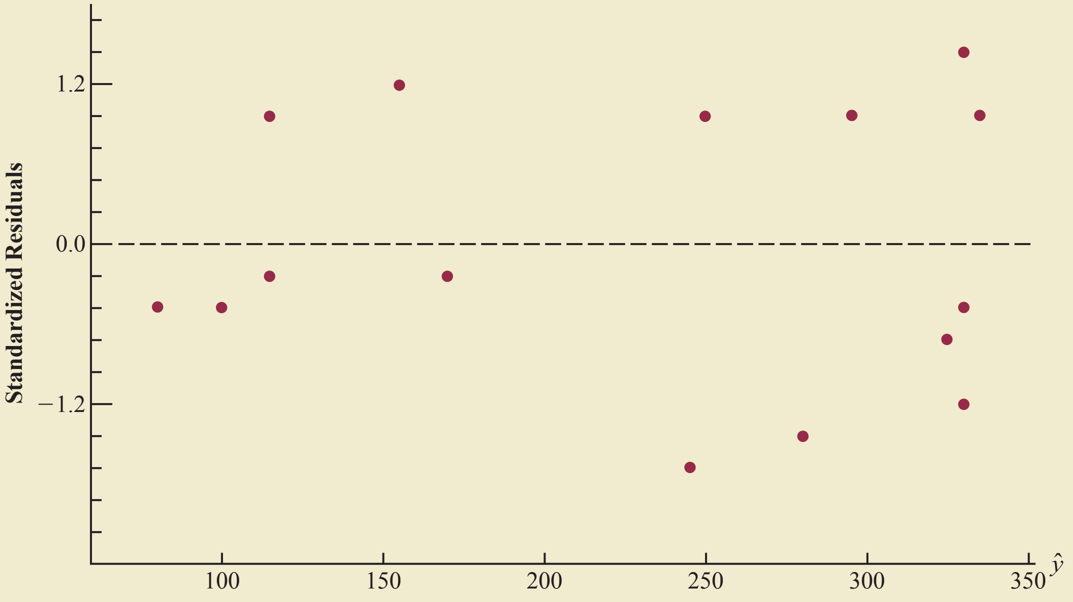
Figure 16.5 is the corresponding standardized residual plot. It shows that the previous curvilinear pattern has been removed. At the .05 level of significance, the computer output shows that the overall model is significant ($p$-value for the $F$ test is .000); note also that the $p$-value corresponding to the $t$-ratio for MonthsSq ($p$-value = .002) is less than .05, and hence we can conclude that adding MonthsSq to the model involving Months is significant. With R-Sq(adj) = 88.6%, we should be pleased with the fit provided by this estimated regression equation. More important, however, is seeing how easy it is to handle curvilinear relationships in regression analysis.

Clearly, many types of relationships can be modeled by using equation (16.1). The regression techniques with which we have been working are definitely not limited to linear, or straight-line, relationships. In multiple regression analysis the word *linear* in the term "general linear model" refers only to the fact that $\beta_0, \beta_1, \ldots, \beta_p$ all have exponents of 1; it does not imply that the relationship between $y$ and the $x_i$'s is linear. Indeed, in this section we have seen one example of how equation (16.1) can be used to model a curvilinear relationship.

FIGURE 16.4 MINITAB OUTPUT FOR THE REYNOLDS EXAMPLE: SECOND-ORDER MODEL

```
The regression equation is
Sales = 45.3 + 6.34 Months - 0.0345 MonthsSq

Predictor         Coef    SE Coef       T      p
Constant         45.35      22.77    1.99  0.070
Months           6.345      1.058    6.00  0.000
MonthsSq     -0.034486   0.008948   -3.85  0.002

S = 34.4528   R-Sq = 90.2%   R-Sq(adj) = 88.6%

Analysis of Variance

SOURCE          DF      SS     MS      F      p
Regression       2  131413  65707  55.36  0.000
Residual Error  12   14244   1187
Total           14  145657
```

FIGURE 16.5 STANDARDIZED RESIDUAL PLOT FOR THE REYNOLDS EXAMPLE: SECOND-ORDER MODEL



## Interaction

If the original data set consists of observations for $y$ and two independent variables $x_1$ and $x_2$, we can develop a second-order model with two predictor variables by setting $z_1 = x_1$, $z_2 = x_2$, $z_3 = x_1^2$, $z_4 = x_2^2$, and $z_5 = x_1x_2$ in the general linear model of equation (16.1). The model obtained is

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_1^2 + \beta_4x_2^2 + \beta_5x_1x_2 + \epsilon \qquad \textbf{(16.4)}$$

TABLE 16.2 DATA FOR THE TYLER PERSONAL CARE EXAMPLE

WEB file

Tyler

| Price | Advertising Expenditure ($1000s) | Sales (1000s) | Price | Advertising Expenditure ($1000s) | Sales (1000s) |
|---|---|---|---|---|---|
| $2.00 | 50 | 478 | $2.00 | 100 | 810 |
| $2.50 | 50 | 373 | $2.50 | 100 | 653 |
| $3.00 | 50 | 335 | $3.00 | 100 | 345 |
| $2.00 | 50 | 473 | $2.00 | 100 | 832 |
| $2.50 | 50 | 358 | $2.50 | 100 | 641 |
| $3.00 | 50 | 329 | $3.00 | 100 | 372 |
| $2.00 | 50 | 456 | $2.00 | 100 | 800 |
| $2.50 | 50 | 360 | $2.50 | 100 | 620 |
| $3.00 | 50 | 322 | $3.00 | 100 | 390 |
| $2.00 | 50 | 437 | $2.00 | 100 | 790 |
| $2.50 | 50 | 365 | $2.50 | 100 | 670 |
| $3.00 | 50 | 342 | $3.00 | 100 | 393 |

In this second-order model, the variable $z_5 = x_1x_2$ is added to account for the potential effects of the two variables acting together. This type of effect is called **interaction**.

To provide an illustration of interaction and what it means, let us review the regression study conducted by Tyler Personal Care for one of its new shampoo products. Two factors believed to have the most influence on sales are unit selling price and advertising expenditure. To investigate the effects of these two variables on sales, prices of $2.00, $2.50, and $3.00 were paired with advertising expenditures of $50,000 and $100,000 in 24 test markets. The unit sales (in thousands) that were observed are reported in Table 16.2.

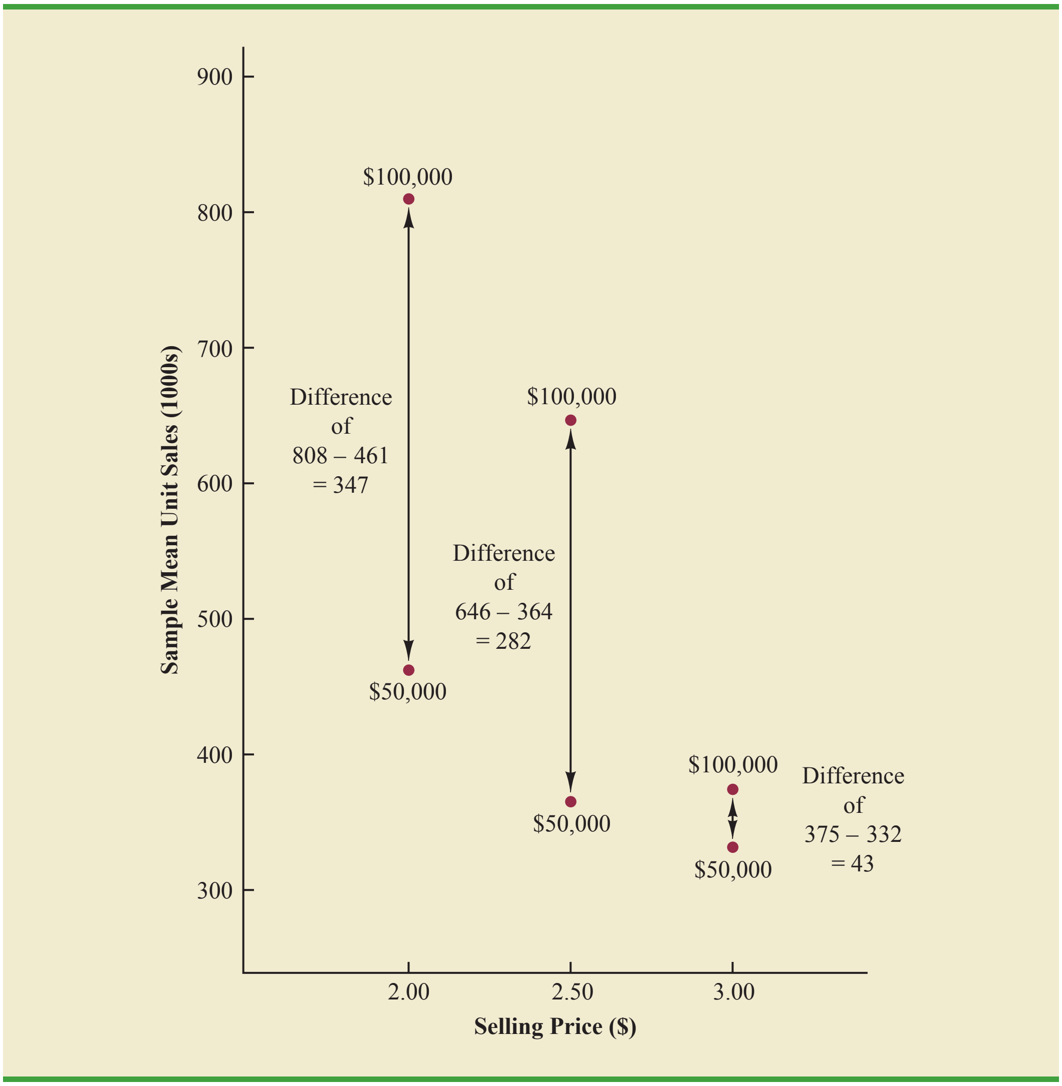
Table 16.3 is a summary of these data. Note that the sample mean sales corresponding to a price of $2.00 and an advertising expenditure of $50,000 is 461,000, and the sample mean sales corresponding to a price of $2.00 and an advertising expenditure of $100,000 is 808,000. Hence, with price held constant at $2.00, the difference in the sample mean sales between advertising expenditures of $50,000 and $100,000 is 808,000 − 461,000 = 347,000 units. When the price of the product is $2.50, the difference in the sample mean sales is 646,000 − 364,000 = 282,000 units. Finally, when the price is $3.00, the difference in the sample mean sales is 375,000 − 332,000 = 43,000 units. Clearly, the difference in the sample mean sales between advertising expenditures of $50,000 and $100,000 depends on the price of the product. In other words, at higher selling prices, the effect of increased advertising expenditure diminishes. These observations provide evidence of interaction between the price and advertising expenditure variables.

TABLE 16.3 SAMPLE MEAN UNIT SALES (1000s) FOR THE TYLER PERSONAL CARE EXAMPLE

| | | Price | | |
|---|---|---|---|---|
| | | **$2.00** | **$2.50** | **$3.00** |
| Advertising Expenditure | **$50,000** | 461 | 364 | 332 |
| | **$100,000** | 808 | 646 | 375 |

Mean sales of 808,000 units when price = $2.00 and advertising expenditure = $100,000

FIGURE 16.6 SAMPLE MEAN UNIT SALES (1000s) AS A FUNCTION OF SELLING PRICE AND ADVERTISING EXPENDITURE



To provide another perspective of interaction, Figure 16.6 shows the sample mean sales for the six price-advertising expenditure combinations. This graph also shows that the effect of advertising expenditure on the sample mean sales depends on the price of the product; we again see the effect of interaction. When interaction between two variables is present, we cannot study the effect of one variable on the response $y$ independently of the other variable. In other words, meaningful conclusions can be developed only if we consider the joint effect that both variables have on the response.

To account for the effect of interaction, we will use the following regression model.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon \qquad (16.5)$$

where

$y$ = unit sales (1000s)
$x_1$ = price ($)
$x_2$ = advertising expenditure ($1000s)

Note that equation (16.5) reflects Tyler's belief that the number of units sold depends linearly on selling price and advertising expenditure (accounted for by the $\beta_1 x_1$ and $\beta_2 x_2$ terms), and that there is interaction between the two variables (accounted for by the $\beta_3 x_1 x_2$ term).

To develop an estimated regression equation, a general linear model involving three independent variables ($z_1$, $z_2$, and $z_3$) was used.

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + \epsilon \qquad \textbf{(16.6)}$$

where

$$\begin{aligned} z_1 &= x_1 \\ z_2 &= x_2 \\ z_3 &= x_1 x_2 \end{aligned}$$

Figure 16.7 is the Minitab output corresponding to the interaction model for the Tyler Personal Care example. The resulting estimated regression equation is

$$\text{Sales} = -276 + 175 \text{ Price} + 19.7 \text{ AdvExp} - 6.08 \text{ PriceAdv}$$

where

$$\begin{aligned} \text{Sales} &= \text{unit sales (1000s)} \\ \text{Price} &= \text{price of the product (\$)} \\ \text{AdvExp} &= \text{advertising expenditure (\$1000s)} \\ \text{PriceAdv} &= \text{interaction term (Price times AdvExp)} \end{aligned}$$

*The data for the* PriceAdv *independent variable is obtained by multiplying each value of* Price *times the corresponding value of* AdvExp.

Because the model is significant ($p$-value for the $F$ test is .000) and the $p$-value corresponding to the $t$ test for PriceAdv is .000, we conclude that interaction is significant given the linear effect of the price of the product and the advertising expenditure. Thus, the regression results show that the effect of advertising expenditure on sales depends on the price.

**FIGURE 16.7** MINITAB OUTPUT FOR THE TYLER PERSONAL CARE EXAMPLE

```
The regression equation is
Sales = - 276 + 175 Price + 19.7 AdvExpen - 6.08 PriceAdv

Predictor      Coef   SE Coef       T      p
Constant     -275.8     112.8   -2.44  0.024
Price        175.00     44.55    3.93  0.001
Adver        19.680     1.427   13.79  0.000
PriceAdv    -6.0800    0.5635  -10.79  0.000

S = 28.1739   R-Sq = 97.8%   R-Sq(adj) = 97.5%

Analysis of Variance

SOURCE           DF      SS      MS       F      p
Regression        3  709316  236439  297.87  0.000
Residual Error   20   15875     794
Total            23  725191
```

TABLE 16.4 MILES-PER-GALLON RATINGS AND WEIGHTS FOR 12 AUTOMOBILES

| Weight | Miles per Gallon |
|---|---|
| 2289 | 28.7 |
| 2113 | 29.2 |
| 2180 | 34.2 |
| 2448 | 27.9 |
| 2026 | 33.3 |
| 2702 | 26.4 |
| 2657 | 23.9 |
| 2106 | 30.5 |
| 3226 | 18.1 |
| 3213 | 19.5 |
| 3607 | 14.3 |
| 2888 | 20.9 |

WEB file MPG

## Transformations Involving the Dependent Variable

In showing how the general linear model can be used to model a variety of possible relationships between the independent variables and the dependent variable, we have focused attention on transformations involving one or more of the independent variables. Often it is worthwhile to consider transformations involving the dependent variable $y$. As an illustration of when we might want to transform the dependent variable, consider the data in Table 16.4, which shows the miles-per-gallon ratings and weights for 12 automobiles. The scatter diagram in Figure 16.8 indicates a negative linear relationship between these two variables. Therefore, we use a simple first-order model to relate the two variables. The Minitab output is shown in Figure 16.9; the resulting estimated regression equation is

$$\text{MPG} = 56.1 - 0.0116 \text{ Weight}$$

where

$$\text{MPG} = \text{miles-per-gallon rating}$$
$$\text{Weight} = \text{weight of the car in pounds}$$

The model is significant ($p$-value for the $F$ test is .000) and the fit is very good (R-sq = 93.5%). However, we note in Figure 16.9 that observation 3 is identified as having a large standardized residual.

Figure 16.10 is the standardized residual plot corresponding to the first-order model. The pattern we observe does not look like the horizontal band we should expect to find if the assumptions about the error term are valid. Instead, the variability in the residuals appears to increase as the value of $\hat{y}$ increases. In other words, we see the wedge-shaped pattern referred to in Chapters 14 and 15 as being indicative of a nonconstant variance. We are not justified in reaching any conclusions about the statistical significance of the resulting estimated regression equation when the underlying assumptions for the tests of significance do not appear to be satisfied.

FIGURE 16.8 SCATTER DIAGRAM FOR THE MILES-PER-GALLON EXAMPLE

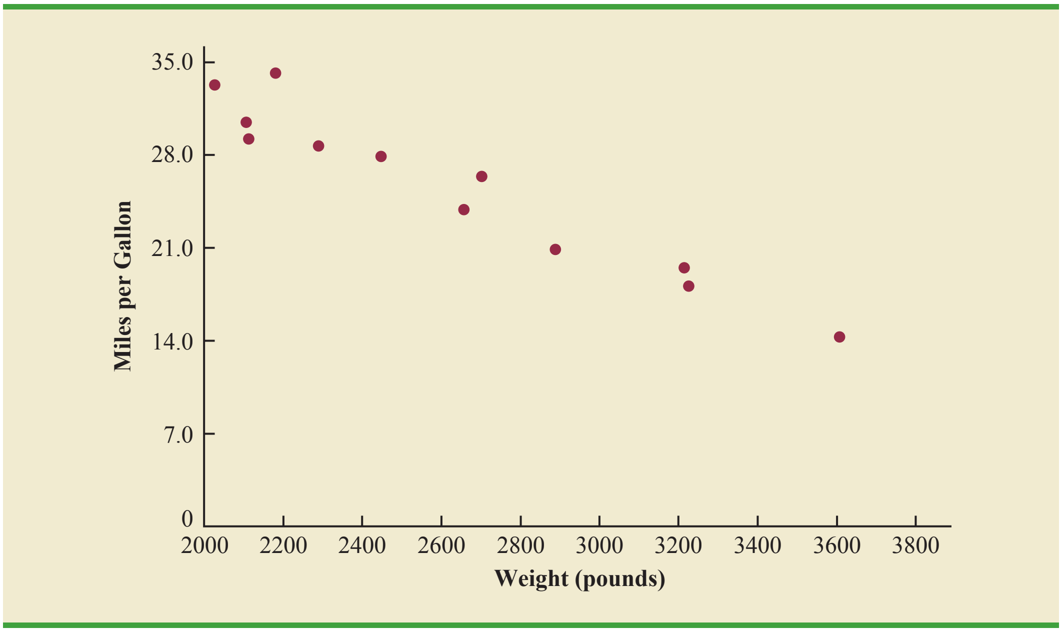
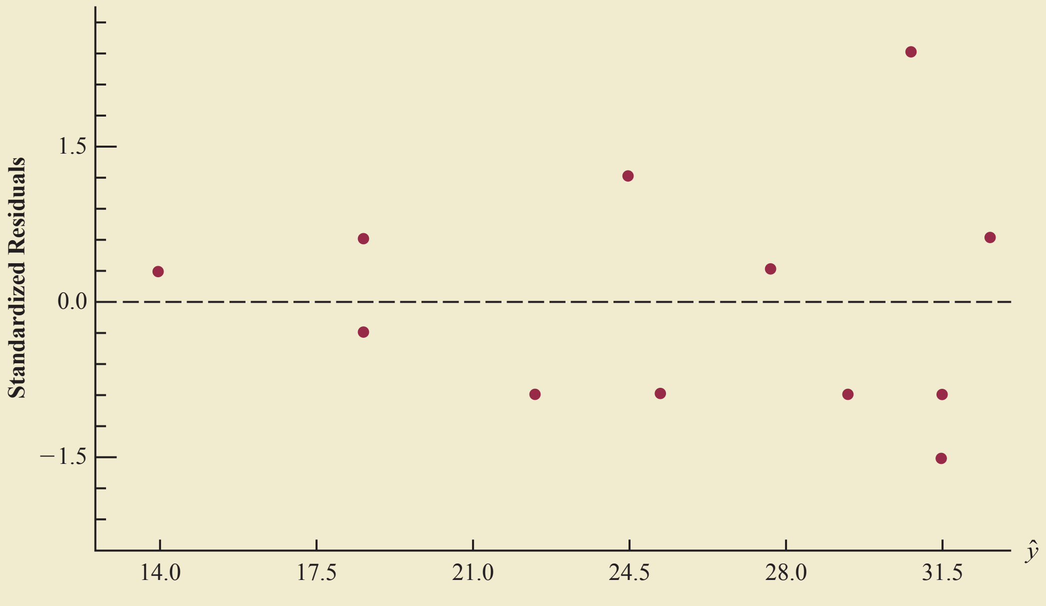


FIGURE 16.9 MINITAB OUTPUT FOR THE MILES-PER-GALLON EXAMPLE

```
The regression equation is
MPG = 56.1 - 0.0116 Weight

Predictor        Coef     SE Coef       T      p
Constant       56.096       2.582   21.72  0.000
Weight     -0.0116436  0.0009677  -12.03  0.000

S = 1.67053   R-Sq = 93.5%   R-Sq(adj) = 92.9%

Analysis of Variance

SOURCE          DF      SS      MS       F      p
Regression       1  403.98  403.98  144.76  0.000
Residual Error  10   27.91    2.79
Total           11  431.88

Unusual Observations
Obs  Weight     MPG     Fit  SE Fit  Residual  St Resid
  3    2180  34.200  30.713   0.644     3.487     2.26R

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

FIGURE 16.10 STANDARDIZED RESIDUAL PLOT FOR THE MILES-PER-GALLON EXAMPLE



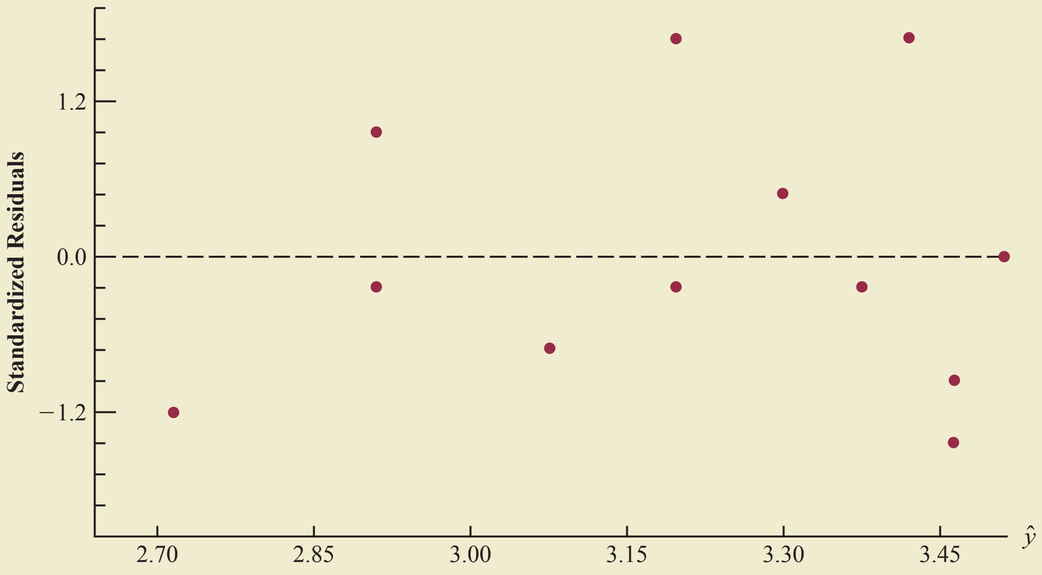
Often the problem of nonconstant variance can be corrected by transforming the dependent variable to a different scale. For instance, if we work with the logarithm of the dependent variable instead of the original dependent variable, the effect will be to compress the values of the dependent variable and thus diminish the effects of nonconstant variance.

Most statistical packages provide the ability to apply logarithmic transformations using either the base 10 (common logarithm) or the base $e = 2.71828 \ldots$ (natural logarithm). We applied a natural logarithmic transformation to the miles-per-gallon data and developed the estimated regression equation relating weight to the natural logarithm of miles-per-gallon. The regression results obtained by using the natural logarithm of miles-per-gallon as the dependent variable, labeled LogeMPG in the output, are shown in Figure 16.11; Figure 16.12 is the corresponding standardized residual plot.

FIGURE 16.11 MINITAB OUTPUT FOR THE MILES-PER-GALLON EXAMPLE: LOGARITHMIC TRANSFORMATION

```
The regression equation is
LogeMPG = 4.52 -0.000501 Weight

Predictor          Coef     SE Coef       T      p
Constant        4.52423     0.09932   45.55  0.000
Weight      -0.00050110  0.00003722  -13.46  0.000

S = 0.0642547   R-Sq = 94.8%   R-Sq(adj) = 94.2%

Analysis of Variance

SOURCE          DF       SS        MS        F      p
Regression       1  0.74822   0.74822   181.22  0.000
Residual Error  10  0.04129   0.00413
Total           11  0.78950
```

FIGURE 16.12 STANDARDIZED RESIDUAL PLOT FOR THE MILES-PER-GALLON EXAMPLE: LOGARITHMIC TRANSFORMATION



Looking at the residual plot in Figure 16.12, we see that the wedge-shaped pattern has now disappeared. Moreover, none of the observations are identified as having a large standardized residual. The model with the logarithm of miles per gallon as the dependent variable is statistically significant and provides an excellent fit to the observed data. Hence, we would recommend using the estimated regression equation

$$\text{LogeMPG} = 4.52 - .000501 \text{ Weight}$$

To predict the miles-per-gallon rating for an automobile that weighs 2500 pounds, we first develop an estimate of the logarithm of the miles-per-gallon rating.

$$\text{LogeMPG} = 4.52 - .000501(2500) = 3.2675$$

The miles-per-gallon estimate is obtained by finding the number whose natural logarithm is 3.2675. Using a calculator with an exponential function, or raising $e$ to the power 3.2675, we obtain 26.2 miles per gallon.

Another approach to problems of nonconstant variance is to use $1/y$ as the dependent variable instead of $y$. This type of transformation is called a *reciprocal transformation.* For instance, if the dependent variable is measured in miles per gallon, the reciprocal transformation would result in a new dependent variable whose units would be 1/(miles per gallon) or gallons per mile. In general, there is no way to determine whether a logarithmic transformation or a reciprocal transformation will perform best without actually trying each of them.

## Nonlinear Models That Are Intrinsically Linear

Models in which the parameters ($\beta_0, \beta_1, \ldots, \beta_p$) have exponents other than 1 are called nonlinear models. However, for the case of the exponential model, we can perform a transformation of variables that will enable us to perform regression analysis with equation (16.1), the general linear model. The exponential model involves the following regression equation.

$$E(y) = \beta_0 \beta_1^x \tag{16.7}$$

This regression equation is appropriate when the dependent variable $y$ increases or decreases by a constant percentage, instead of by a fixed amount, as $x$ increases.

As an example, suppose sales for a product $y$ are related to advertising expenditure $x$ (in thousands of dollars) according to the following regression equation.

$$E(y) = 500(1.2)^x$$

Thus, for $x = 1$, $E(y) = 500(1.2)^1 = 600$; for $x = 2$, $E(y) = 500(1.2)^2 = 720$; and for $x = 3$, $E(y) = 500(1.2)^3 = 864$. Note that $E(y)$ is not increasing by a constant amount in this case, but by a constant percentage; the percentage increase is 20%.

We can transform this nonlinear regression equation to a linear regression equation by taking the logarithm of both sides of equation (16.7).

$$\log E(y) = \log \beta_0 + x \log \beta_1 \tag{16.8}$$

Now if we let $y' = \log E(y)$, $\beta_0' = \log \beta_0$, and $\beta_1' = \log \beta_1$, we can rewrite equation (16.8) as

$$y' = \beta_0' + \beta_1' x$$

It is clear that the formulas for simple linear regression can now be used to develop estimates of $\beta_0'$ and $\beta_1'$. Denoting the estimates as $b_0'$ and $b_1'$ leads to the following estimated regression equation.

$$\hat{y}' = b_0' + b_1' x \tag{16.9}$$

To obtain predictions of the original dependent variable $y$ given a value of $x$, we would first substitute the value of $x$ into equation (16.9) and compute $\hat{y}'$. The antilog of $\hat{y}'$ would be the prediction of $y$, or the expected value of $y$.

Many nonlinear models cannot be transformed into an equivalent linear model. However, such models have had limited use in business and economic applications. Furthermore, the mathematical background needed for study of such models is beyond the scope of this text.

## Exercises

### Methods

SELF test

1. Consider the following data for two variables, $x$ and $y$.

| $x$ | 22 | 24 | 26 | 30 | 35 | 40 |
|---|---|---|---|---|---|---|
| $y$ | 12 | 21 | 33 | 35 | 40 | 36 |

a. Develop an estimated regression equation for the data of the form $\hat{y} = b_0 + b_1x$.
b. Use the results from part (a) to test for a significant relationship between $x$ and $y$. Use $\alpha = .05$.
c. Develop a scatter diagram for the data. Does the scatter diagram suggest an estimated regression equation of the form $\hat{y} = b_0 + b_1x + b_2x^2$? Explain.
d. Develop an estimated regression equation for the data of the form $\hat{y} = b_0 + b_1x + b_2x^2$.
e. Refer to part (d). Is the relationship between $x$, $x^2$, and $y$ significant? Use $\alpha = .05$.
f. Predict the value of $y$ when $x = 25$.

2. Consider the following data for two variables, $x$ and $y$.

| $x$ | 9 | 32 | 18 | 15 | 26 |
|---|---|---|---|---|---|
| $y$ | 10 | 20 | 21 | 16 | 22 |

a. Develop an estimated regression equation for the data of the form $\hat{y} = b_0 + b_1x$. Comment on the adequacy of this equation for predicting $y$.
b. Develop an estimated regression equation for the data of the form $\hat{y} = b_0 + b_1x + b_2x^2$. Comment on the adequacy of this equation for predicting $y$.
c. Predict the value of $y$ when $x = 20$.

3. Consider the following data for two variables, $x$ and $y$.

| $x$ | 2 | 3 | 4 | 5 | 7 | 7 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| $y$ | 4 | 5 | 4 | 6 | 4 | 6 | 9 | 5 | 11 |

a. Does there appear to be a linear relationship between $x$ and $y$? Explain.
b. Develop the estimated regression equation relating $x$ and $y$.
c. Plot the standardized residuals versus $\hat{y}$ for the estimated regression equation developed in part (b). Do the model assumptions appear to be satisfied? Explain.
d. Perform a logarithmic transformation on the dependent variable $y$. Develop an estimated regression equation using the transformed dependent variable. Do the model assumptions appear to be satisfied by using the transformed dependent variable? Does a reciprocal transformation work better in this case? Explain.

## Applications

4. A highway department is studying the relationship between traffic flow and speed. The following model has been hypothesized.

$$y = \beta_0 + \beta_1 x + \epsilon$$

where

$$y = \text{traffic flow in vehicles per hour}$$
$$x = \text{vehicle speed in miles per hour}$$

The following data were collected during rush hour for six highways leading out of the city.

| Traffic Flow ($y$) | Vehicle Speed ($x$) |
|---|---|
| 1256 | 35 |
| 1329 | 40 |
| 1226 | 30 |
| 1335 | 45 |
| 1349 | 50 |
| 1124 | 25 |

a. Develop an estimated regression equation for the data.
b. Use $\alpha = .01$ to test for a significant relationship.

SELF test

5. In working further with the problem of exercise 4, statisticians suggested the use of the following curvilinear estimated regression equation.

$$\hat{y} = b_0 + b_1 x + b_2 x^2$$

a. Use the data of exercise 4 to estimate the parameters of this estimated regression equation.
b. Use $\alpha = .01$ to test for a significant relationship.
c. Predict the traffic flow in vehicles per hour at a speed of 38 miles per hour.

6. A study of emergency service facilities investigated the relationship between the number of facilities and the average distance traveled to provide the emergency service. The following table gives the data collected.

| Number of Facilities | Average Distance (miles) |
|---|---|
| 9 | 1.66 |
| 11 | 1.12 |
| 16 | .83 |
| 21 | .62 |
| 27 | .51 |
| 30 | .47 |

a. Develop a scatter diagram for these data, treating average distance traveled as the dependent variable.
b. Does a simple linear regression model appear to be appropriate? Explain.
c. Develop an estimated regression equation for the data that you believe will best explain the relationship between these two variables.

7. In 2011, home prices and mortgage rates fell so far that in a number of cities the monthly cost of owning a home was less expensive than renting. The following data show the average asking rent and the monthly mortgage on the median-priced home

(including taxes and insurance) for 10 cities where the average monthly mortgage payment was less than the average asking rent (*The Wall Street Journal,* November 26–27, 2011).

WEB file

RentMortgage

| City | Rent ($) | Mortgage ($) |
|---|---|---|
| Atlanta | 840 | 539 |
| Chicago | 1062 | 1002 |
| Detroit | 823 | 626 |
| Jacksonville, Fla. | 779 | 711 |
| Las Vegas | 796 | 655 |
| Miami | 1071 | 977 |
| Minneapolis | 953 | 776 |
| Orlando, Fla. | 851 | 695 |
| Phoenix | 762 | 651 |
| St. Louis | 723 | 654 |

a. Develop a scatter diagram for these data, treating the average asking rent as the independent variable. Does a simple linear regression model appear to be appropriate?

b. Use a simple linear regression model to develop an estimated regression equation to predict the monthly mortgage on the median-priced home given the average asking rent. Construct a standardized residual plot. Based upon the standardized residual plot, does a simple linear regression model appear to be appropriate?

c. Using a second-order model, develop an estimated regression equation to predict the monthly mortgage on the median-priced home given the average asking rent.

d. Do you prefer the estimated regression equation developed in part (a) or part (c)? Explain.

8. Corvette, Ferrari, and Jaguar produced a variety of classic cars that continue to increase in value. The following data, based upon the Martin Rating System for Collectible Cars, show the rarity rating (1–20) and the high price ($1000) for 15 classic cars (*BusinessWeek* website, February 2006).

WEB file

ClassicCars

| Year | Make | Model | Rating | Price ($1000) |
|---|---|---|---|---|
| 1984 | Chevrolet | Corvette | 18 | 1600.0 |
| 1956 | Chevrolet | Corvette 265/225-hp | 19 | 4000.0 |
| 1963 | Chevrolet | Corvette coupe (340-bhp 4-speed) | 18 | 1000.0 |
| 1978 | Chevrolet | Corvette coupe Silver Anniversary | 19 | 1300.0 |
| 1960–1963 | Ferrari | 250 GTE 2+2 | 16 | 350.0 |
| 1962–1964 | Ferrari | 250 GTL Lusso | 19 | 2650.0 |
| 1962 | Ferrari | 250 GTO | 18 | 375.0 |
| 1967–1968 | Ferrari | 275 GTB/4 NART Spyder | 17 | 450.0 |
| 1968–1973 | Ferrari | 365 GTB/4 Daytona | 17 | 140.0 |
| 1962–1967 | Jaguar | E-type OTS | 15 | 77.5 |
| 1969–1971 | Jaguar | E-type Series II OTS | 14 | 62.0 |
| 1971–1974 | Jaguar | E-type Series III OTS | 16 | 125.0 |
| 1951–1954 | Jaguar | XK 120 roadster (steel) | 17 | 400.0 |
| 1950–1953 | Jaguar | XK C-type | 16 | 250.0 |
| 1956–1957 | Jaguar | XKSS | 13 | 70.0 |

a. Develop a scatter diagram of the data using the rarity rating as the independent variable and price as the independent variable. Does a simple linear regression model appear to be appropriate?

b. Develop an estimated multiple regression equation with $x$ = rarity rating and $x^2$ as the two independent variables.

c. Consider the nonlinear relationship shown by equation (16.7). Use logarithms to develop an estimated regression equation for this model.

d. Do you prefer the estimated regression equation developed in part (b) or part (c)? Explain.

WEB file

MetroAreas

9. *Kiplinger's Personal Finance Magazine* rated 359 U.S. metropolitan areas to determine the best cities to live, work, and play. The data contained in the data set named MetroAreas show the data from the Kiplinger study for the 50 metropolitan areas with a population of 1,000,000 or more (Kiplinger's website, March 2, 2009). The data set includes the following variables: Population, Income, Cost of Living Index, and Creative (%). Population is the size of the population in 1000s; Income is the median household income in $1000s; Cost of Living Index is based on 100 being the national average; and Creative (%) is the percentage of the workforce in creative fields such as science, engineering, architecture, education, art, and entertainment. Workers in creative fields are generally considered an important factor in the vitality and livability of a city and a key to future economic prosperity.
   a. Develop a scatter diagram for these data with median household income as the independent variable and the percentage of the workforce in creative fields as the dependent variable. Does a simple linear regression model appear to be appropriate?
   b. Develop a scatter diagram for these data with the cost of living index as the independent variable and the percentage of the workforce in creative fields as the dependent variable. Does a simple linear regression model appear to be appropriate?
   c. Use the data provided to develop the best estimated multiple regression equation for estimating the percentage of the workforce in creative fields.
   d. The Tucson, Arizona, metropolitan area has a population of 946,362, a median household income of $42,984, and cost of living index of 99. Develop a prediction of the percentage of the workforce in creative fields for Tucson. Are there any factors that should be considered before using this predicted value?

## 16.2 Determining When to Add or Delete Variables

In this section we will show how an $F$ test can be used to determine whether it is advantageous to add one or more independent variables to a multiple regression model. This test is based on a determination of the amount of reduction in the error sum of squares resulting from adding one or more independent variables to the model. We will first illustrate how the test can be used in the context of the Butler Trucking example.

In Chapter 15, the Butler Trucking example was introduced to illustrate the use of multiple regression analysis. Recall that the managers wanted to develop an estimated regression equation to predict total daily travel time for trucks using two independent variables: miles traveled and number of deliveries. With miles traveled $x_1$ as the only independent variable, the least squares procedure provided the following estimated regression equation.

$$\hat{y} = 1.27 + .0678x_1$$

In Chapter 15 we showed that the error sum of squares for this model was SSE = 8.029. When $x_2$, the number of deliveries, was added as a second independent variable, we obtained the following estimated regression equation.

$$\hat{y} = -.869 + .0611x_1 + .923x_2$$

The error sum of squares for this model was SSE = 2.299. Clearly, adding $x_2$ resulted in a reduction of SSE. The question we want to answer is: Does adding the variable $x_2$ lead to a *significant* reduction in SSE?

We use the notation SSE($x_1$) to denote the error sum of squares when $x_1$ is the only independent variable in the model, SSE($x_1, x_2$) to denote the error sum of squares when $x_1$ and $x_2$ are both in the model, and so on. Hence, the reduction in SSE resulting from adding $x_2$ to the model involving just $x_1$ is

$$\text{SSE}(x_1) - \text{SSE}(x_1, x_2) = 8.029 - 2.299 = 5.730$$

An $F$ test is conducted to determine whether this reduction is significant.

The numerator of the $F$ statistic is the reduction in SSE divided by the number of independent variables added to the original model. Here only one variable, $x_2$, has been added; thus, the numerator of the $F$ statistic is

$$\frac{\text{SSE}(x_1) - \text{SSE}(x_1, x_2)}{1} = 5.730$$

The result is a measure of the reduction in SSE per independent variable added to the model. The denominator of the $F$ statistic is the mean square error for the model that includes all of the independent variables. For Butler Trucking this corresponds to the model containing both $x_1$ and $x_2$; thus, $p = 2$ and

$$\text{MSE} = \frac{\text{SSE}(x_1, x_2)}{n - p - 1} = \frac{2.299}{7} = .3284$$

The following $F$ statistic provides the basis for testing whether the addition of $x_2$ is statistically significant.

$$F = \frac{\dfrac{\text{SSE}(x_1) - \text{SSE}(x_1, x_2)}{1}}{\dfrac{\text{SSE}(x_1, x_2)}{n - p - 1}} \tag{16.10}$$

The numerator degrees of freedom for this $F$ test is equal to the number of variables added to the model, and the denominator degrees of freedom is equal to $n - p - 1$.

For the Butler Trucking problem, we obtain

$$F = \frac{\dfrac{5.730}{1}}{\dfrac{2.299}{7}} = \frac{5.730}{.3284} = 17.45$$

Refer to Table 4 of Appendix B. We find that for a level of significance of $\alpha = .05$, $F_{.05} = 5.59$. Because $F = 17.45 > F_{.05} = 5.59$, we can reject the null hypothesis that $x_2$ is not statistically significant; in other words, adding $x_2$ to the model involving only $x_1$ results in a significant reduction in the error sum of squares.

When we want to test for the significance of adding only one more independent variable to a model, the result found with the $F$ test just described could also be obtained by using the $t$ test for the significance of an individual parameter (described in Section 15.4). Indeed, the $F$ statistic we just computed is the square of the $t$ statistic used to test the significance of an individual parameter.

Because the $t$ test is equivalent to the $F$ test when only one independent variable is being added to the model, we can now further clarify the proper use of the $t$ test for testing the significance of an individual parameter. If an individual parameter is not significant, the corresponding variable can be dropped from the model. However, if the $t$ test shows that two or more parameters are not significant, no more than one independent variable can ever be dropped from a model on the basis of a $t$ test; if one variable is dropped, a second variable that was not significant initially might become significant.

We now turn to a consideration of whether the addition of more than one independent variable—as a set—results in a significant reduction in the error sum of squares.

## General Case

Consider the following multiple regression model involving $q$ independent variables, where $q < p$.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q + \epsilon \tag{16.11}$$

If we add variables $x_{q+1}, x_{q+2}, \ldots, x_p$ to this model, we obtain a model involving $p$ independent variables.

$$\begin{aligned} y = \beta_0 &+ \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q \\ &+ \beta_{q+1} x_{q+1} + \beta_{q+2} x_{q+2} + \cdots + \beta_p x_p + \epsilon \end{aligned} \tag{16.12}$$

To test whether the addition of $x_{q+1}, x_{q+2}, \ldots, x_p$ is statistically significant, the null and alternative hypotheses can be stated as follows.

$$H_0\colon \beta_{q+1} = \beta_{q+2} = \cdots = \beta_p = 0$$
$$H_\text{a}\colon \text{One or more of the parameters is not equal to zero}$$

The following $F$ statistic provides the basis for testing whether the additional independent variables are statistically significant.

$$F = \frac{\dfrac{\text{SSE}(x_1, x_2, \ldots, x_q) - \text{SSE}(x_1, x_2, \ldots, x_q, x_{q+1}, \ldots, x_p)}{p - q}}{\dfrac{\text{SSE}(x_1, x_2, \ldots, x_q, x_{q+1}, \ldots, x_p)}{n - p - 1}} \tag{16.13}$$

This computed $F$ value is then compared with $F_\alpha$, the table value with $p - q$ numerator degrees of freedom and $n - p - 1$ denominator degrees of freedom. If $F > F_\alpha$, we reject $H_0$ and conclude that the set of additional independent variables is statistically significant. Note that for the special case where $q = 1$ and $p = 2$, equation (16.13) reduces to equation (16.10).

*Many computer packages, such as Minitab, provide extra sums of squares corresponding to the order in which each independent variable enters the model; in such cases, the computation of the F test for determining whether to add or delete a set of variables is simplified.*

Many students find equation (16.13) somewhat complex. To provide a simpler description of this $F$ ratio, we can refer to the model with the smaller number of independent variables as the reduced model and the model with the larger number of independent variables as the full model. If we let SSE(reduced) denote the error sum of squares for the reduced model and SSE(full) denote the error sum of squares for the full model, we can write the numerator of (16.13) as

$$\frac{\text{SSE(reduced)} - \text{SSE(full)}}{\text{number of extra terms}} \tag{16.14}$$

Note that "number of extra terms" denotes the difference between the number of independent variables in the full model and the number of independent variables in the reduced model. The denominator of equation (16.13) is the error sum of squares for the full model divided by the corresponding degrees of freedom; in other words, the denominator is the mean square error for the full model. Denoting the mean square error for the full model as MSE(full) enables us to write it as

$$F = \frac{\dfrac{\text{SSE(reduced)} - \text{SSE(full)}}{\text{number of extra terms}}}{\text{MSE(full)}} \tag{16.15}$$

To illustrate the use of this $F$ statistic, suppose we have a regression problem involving 30 observations. One model with the independent variables $x_1$, $x_2$, and $x_3$ has an error sum of squares of 150 and a second model with the independent variables $x_1$, $x_2$, $x_3$, $x_4$, and $x_5$ has an error sum of squares of 100. Did the addition of the two independent variables $x_4$ and $x_5$ result in a significant reduction in the error sum of squares?

First, note that the degrees of freedom for SST is $30 - 1 = 29$ and that the degrees of freedom for the regression sum of squares for the full model is five (the number of independent variables in the full model). Thus, the degrees of freedom for the error sum of squares for the full model is $29 - 5 = 24$, and hence MSE(full) $= 100/24 = 4.17$. Therefore the $F$ statistic is

$$F = \frac{\dfrac{150 - 100}{2}}{4.17} = 6.00$$

This computed $F$ value is compared with the table $F$ value with two numerator and 24 denominator degrees of freedom. At the .05 level of significance, Table 4 of Appendix B shows $F_{.05} = 3.40$. Because $F = 6.00$ is greater than 3.40, we conclude that the addition of variables $x_4$ and $x_5$ is statistically significant.

## Use of $p$-Values

The $p$-value criterion can also be used to determine whether it is advantageous to add one or more independent variables to a multiple regression model. In the preceding example, we showed how to perform an $F$ test to determine if the addition of two independent variables, $x_4$ and $x_5$, to a model with three independent variables, $x_1$, $x_2$, and $x_3$, was statistically significant. For this example, the computed $F$ statistic was 6.00 and we concluded (by comparing $F = 6.00$ to the critical value $F_{.05} = 3.40$) that the addition of variables $x_4$ and $x_5$ was significant. Using Minitab or Excel, the $p$-value associated with $F = 6.00$ (2 numerator and 24 denominator degrees of freedom) is .008. With a $p$-value $= .008 < \alpha = .05$, we also conclude that the addition of the two independent variables is statistically significant. It is difficult to determine the $p$-value directly from tables of the $F$ distribution, but computer software packages, such as Minitab or Excel, make computing the $p$-value easy.

## NOTES AND COMMENTS

Computation of the $F$ statistic can also be based on the difference in the regression sums of squares. To show this form of the $F$ statistic, we first note that

$$\text{SSE(reduced)} = \text{SST} - \text{SSR(reduced)}$$
$$\text{SSE(full)} = \text{SST} - \text{SSR(full)}$$

Hence

$$\begin{aligned}\text{SSE(reduced)} - \text{SSE(full)} &= [\text{SST} - \text{SSR(reduced)}] - [\text{SST} - \text{SSR(full)}]\\ &= \text{SSR(full)} - \text{SSR(reduced)}\end{aligned}$$

Thus,

$$F = \frac{\dfrac{\text{SSR(full)} - \text{SSR(reduced)}}{\text{number of extra terms}}}{\text{MSE(full)}}$$

## Exercises

### Methods

10. In a regression analysis involving 27 observations, the following estimated regression equation was developed.

$$\hat{y} = 25.2 + 5.5x_1$$

For this estimated regression equation SST = 1550 and SSE = 520.

a. At $\alpha = .05$, test whether $x_1$ is significant.

Suppose that variables $x_2$ and $x_3$ are added to the model and the following regression equation is obtained.

$$\hat{y} = 16.3 + 2.3x_1 + 12.1x_2 - 5.8x_3$$

For this estimated regression equation SST = 1550 and SSE = 100.

b. Use an $F$ test and a .05 level of significance to determine whether $x_2$ and $x_3$ contribute significantly to the model.

SELF test

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

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

For this estimated regression equation SST = 1805 and SSR = 1760.

a. At $\alpha = .05$, test the significance of the relationship among the variables.

Suppose variables $x_1$ and $x_4$ are dropped from the model and the following estimated regression equation is obtained.

$$\hat{y} = 11.1 - 3.6x_2 + 8.1x_3$$

For this model SST = 1805 and SSR = 1705.

b. Compute SSE($x_1, x_2, x_3, x_4$).

c. Compute SSE($x_2, x_3$).

d. Use an $F$ test and a .05 level of significance to determine whether $x_1$ and $x_4$ contribute significantly to the model.

### Applications

WEB file

LPGATour

12. The Ladies Professional Golfers Association (LPGA) maintains statistics on performance and earnings for members of the LPGA Tour. Year-end performance statistics for the 30 players who had the highest total earnings in LPGA Tour events for 2005 appear in the file named LPGATour (LPGA Tour website, 2006). Earnings ($1000) is the total earnings in thousands of dollars; Scoring Avg. is the average score for all events; Greens in Reg. is the percentage of time a player is able to hit the green in regulation; Putting Avg. is the average number of putts taken on greens hit in regulation; and Sand Saves is the percentage of time a player is able to get "up and down" once in a greenside sand bunker. A green is considered hit in regulation if any part of the ball is touching the putting surface and the difference between the value of par for the hole and the number of strokes taken to hit the green is at least 2.

a. Develop an estimated regression equation that can be used to predict the average score for all events given the average number of putts taken on greens hit in regulation.

b. Develop an estimated regression equation that can be used to predict the average score for all events given the percentage of time a player is able to hit the green in regulation, the average number of putts taken on greens hit in regulation, and the percentage of time a player is able to get "up and down" once in a greenside sand bunker.

c. At the .05 level of significance, test whether the two independent variables added in part (b), the percentage of time a player is able to hit the green in regulation and the percentage of time a player is able to get "up and down" once in a greenside sand bunker, contribute significantly to the estimated regression equation developed in part (a). Explain.

WEB file

LPGATour

13. Refer to exercise 12.

a. Develop an estimated regression equation that can be used to predict the total earnings for all events given the average number of putts taken on greens hit in regulation.

b. Develop an estimated regression equation that can be used to predict the total earnings for all events given the percentage of time a player is able to hit the green in regulation, the average number of putts taken on greens hit in regulation, and the percentage of time a player is able to get "up and down" once in a greenside sand bunker.

c. At the .05 level of significance, test whether the two independent variables added in part (b), the percentage of time a player is able to hit the green in regulation and the percentage of time a player is able to get "up and down" once in a greenside sand bunker, contribute significantly to the estimated regression equation developed in part (a). Explain.

d. In general, lower scores should lead to higher earnings. To investigate this option to predicting total earnings, develop an estimated regression equation that can be used to predict total earnings for all events given the average score for all events. Would you prefer to use this equation to predict total earnings or the estimated regression equation developed in part (b)? Explain.

14. A 10-year study conducted by the American Heart Association provided data on how age, blood pressure, and smoking relate to the risk of strokes. Data from a portion of this study follow. Risk is interpreted as the probability (times 100) that a person will have a stroke over the next 10-year period. For the smoker variable, 1 indicates a smoker and 0 indicates a nonsmoker.

WEB file

Stroke

| Risk | Age | Blood Pressure | Smoker |
|---|---|---|---|
| 12 | 57 | 152 | 0 |
| 24 | 67 | 163 | 0 |
| 13 | 58 | 155 | 0 |
| 56 | 86 | 177 | 1 |
| 28 | 59 | 196 | 0 |
| 51 | 76 | 189 | 1 |
| 18 | 56 | 155 | 1 |
| 31 | 78 | 120 | 0 |
| 37 | 80 | 135 | 1 |
| 15 | 78 | 98 | 0 |
| 22 | 71 | 152 | 0 |
| 36 | 70 | 173 | 1 |
| 15 | 67 | 135 | 1 |
| 48 | 77 | 209 | 1 |
| 15 | 60 | 199 | 0 |
| 36 | 82 | 119 | 1 |
| 8 | 66 | 166 | 0 |
| 34 | 80 | 125 | 1 |
| 3 | 62 | 117 | 0 |
| 37 | 59 | 207 | 1 |

a. Develop an estimated regression equation that can be used to predict the risk of stroke given the age and blood-pressure level.

b. Consider adding two independent variables to the model developed in part (a), one for the interaction between age and blood-pressure level and the other for whether the person is a smoker. Develop an estimated regression equation using these four independent variables.

c. At a .05 level of significance, test to see whether the addition of the interaction term and the smoker variable contribute significantly to the estimated regression equation developed in part (a).

15. In baseball, an earned run is any run that the opposing team scores off the pitcher except for runs scored as a result of errors. The earned run average (ERA), the statistic most often used to compare the performance of pitchers, is computed as follows:

$$\text{ERA} = \left(\frac{\text{earned runs given up}}{\text{innings pitched}}\right)9$$

WEB file

MLBPitching

Note that the average number of earned runs per inning pitched is multiplied by nine, the number of innings in a regulation game. Thus, ERA represents the average number of runs the pitcher gives up per nine innings. For instance, in 2008, Roy Halladay, a pitcher for the Toronto Blue Jays, pitched 246 innings and gave up 76 earned runs; his ERA was (76/246)9 = 2.78. To investigate the relationship between ERA and other measures of pitching performance, data for 50 Major League Baseball pitchers for the 2008 season appear in the data set named MLBPitching (MLB website, February 2009). Descriptions for variables which appear on the data set follow:

| | |
|---|---|
| W | Number of games won |
| L | Number of games lost |
| WPCT | Percentage of games won |
| H/9 | Average number of hits given up per nine innings |
| HR/9 | Average number of home runs given up per nine innings |
| BB/9 | Average number of bases on balls given up per nine innings |

a. Develop an estimated regression equation that can be used to predict the earned run average given the average number hits given up per nine innings.

b. Develop an estimated regression equation that can be used to predict the earned run average given the average number hits given up per nine innings, the average number of home runs given up per nine innings, and the average number of bases on balls given up per nine innings.

c. At the .05 level of significance, test whether the two independent variables added in part (b), the average number of home runs given up per nine innings and the average number of bases on ball given up per nine innings, contribute significantly to the estimated regression equation developed in part (a).

## 16.3 Analysis of a Larger Problem

In introducing multiple regression analysis, we used the Butler Trucking example extensively. The small size of this problem was an advantage in exploring introductory concepts but would make it difficult to illustrate some of the variable selection issues involved in model building. To provide an illustration of the variable selection procedures discussed in the next section, we introduce a data set consisting of 25 observations on eight independent variables. Permission to use these data was provided by Dr. David W. Cravens of the Department of Marketing at Texas Christian University. Consequently, we refer to the data set as the Cravens data.[1]

The Cravens data are for a company that sells products in several sales territories, each of which is assigned to a single sales representative. A regression analysis was conducted

[1] For details see David W. Cravens, Robert B. Woodruff, and Joe C. Stamper, "An Analytical Approach for Evaluating Sales Territory Performance," *Journal of Marketing,* 36 (January 1972): 31–37. Copyright © 1972 American Marketing Association.

TABLE 16.5 CRAVENS DATA

WEB file
Cravens

| Sales | Time | Poten | AdvExp | Share | Change | Accounts | Work | Rating |
|---|---|---|---|---|---|---|---|---|
| 3,669.88 | 43.10 | 74,065.1 | 4,582.9 | 2.51 | .34 | 74.86 | 15.05 | 4.9 |
| 3,473.95 | 108.13 | 58,117.3 | 5,539.8 | 5.51 | .15 | 107.32 | 19.97 | 5.1 |
| 2,295.10 | 13.82 | 21,118.5 | 2,950.4 | 10.91 | −.72 | 96.75 | 17.34 | 2.9 |
| 4,675.56 | 186.18 | 68,521.3 | 2,243.1 | 8.27 | .17 | 195.12 | 13.40 | 3.4 |
| 6,125.96 | 161.79 | 57,805.1 | 7,747.1 | 9.15 | .50 | 180.44 | 17.64 | 4.6 |
| 2,134.94 | 8.94 | 37,806.9 | 402.4 | 5.51 | .15 | 104.88 | 16.22 | 4.5 |
| 5,031.66 | 365.04 | 50,935.3 | 3,140.6 | 8.54 | .55 | 256.10 | 18.80 | 4.6 |
| 3,367.45 | 220.32 | 35,602.1 | 2,086.2 | 7.07 | −.49 | 126.83 | 19.86 | 2.3 |
| 6,519.45 | 127.64 | 46,176.8 | 8,846.2 | 12.54 | 1.24 | 203.25 | 17.42 | 4.9 |
| 4,876.37 | 105.69 | 42,053.2 | 5,673.1 | 8.85 | .31 | 119.51 | 21.41 | 2.8 |
| 2,468.27 | 57.72 | 36,829.7 | 2,761.8 | 5.38 | .37 | 116.26 | 16.32 | 3.1 |
| 2,533.31 | 23.58 | 33,612.7 | 1,991.8 | 5.43 | −.65 | 142.28 | 14.51 | 4.2 |
| 2,408.11 | 13.82 | 21,412.8 | 1,971.5 | 8.48 | .64 | 89.43 | 19.35 | 4.3 |
| 2,337.38 | 13.82 | 20,416.9 | 1,737.4 | 7.80 | 1.01 | 84.55 | 20.02 | 4.2 |
| 4,586.95 | 86.99 | 36,272.0 | 10,694.2 | 10.34 | .11 | 119.51 | 15.26 | 5.5 |
| 2,729.24 | 165.85 | 23,093.3 | 8,618.6 | 5.15 | .04 | 80.49 | 15.87 | 3.6 |
| 3,289.40 | 116.26 | 26,878.6 | 7,747.9 | 6.64 | .68 | 136.58 | 7.81 | 3.4 |
| 2,800.78 | 42.28 | 39,572.0 | 4,565.8 | 5.45 | .66 | 78.86 | 16.00 | 4.2 |
| 3,264.20 | 52.84 | 51,866.1 | 6,022.7 | 6.31 | −.10 | 136.58 | 17.44 | 3.6 |
| 3,453.62 | 165.04 | 58,749.8 | 3,721.1 | 6.35 | −.03 | 138.21 | 17.98 | 3.1 |
| 1,741.45 | 10.57 | 23,990.8 | 861.0 | 7.37 | −1.63 | 75.61 | 20.99 | 1.6 |
| 2,035.75 | 13.82 | 25,694.9 | 3,571.5 | 8.39 | −.43 | 102.44 | 21.66 | 3.4 |
| 1,578.00 | 8.13 | 23,736.3 | 2,845.5 | 5.15 | .04 | 76.42 | 21.46 | 2.7 |
| 4,167.44 | 58.44 | 34,314.3 | 5,060.1 | 12.88 | .22 | 136.58 | 24.78 | 2.8 |
| 2,799.97 | 21.14 | 22,809.5 | 3,552.0 | 9.14 | −.74 | 88.62 | 24.96 | 3.9 |

to determine whether a variety of predictor (independent) variables could explain sales in each territory. A random sample of 25 sales territories resulted in the data in Table 16.5; the variable definitions are given in Table 16.6.

As a preliminary step, let us consider the sample correlation coefficients between each pair of variables. Figure 16.13 is the correlation matrix obtained using Minitab. Note that the sample correlation coefficient between Sales and Time is .623, between Sales and Poten is .598, and so on.

TABLE 16.6 VARIABLE DEFINITIONS FOR THE CRAVENS DATA

| Variable | Definition |
|---|---|
| Sales | Total sales credited to the sales representative |
| Time | Length of time employed in months |
| Poten | Market potential; total industry sales in units for the sales territory* |
| AdvExp | Advertising expenditure in the sales territory |
| Share | Market share; weighted average for the past four years |
| Change | Change in the market share over the previous four years |
| Accounts | Number of accounts assigned to the sales representative* |
| Work | Workload; a weighted index based on annual purchases and concentrations of accounts |
| Rating | Sales representative overall rating on eight performance dimensions; an aggregate rating on a 1–7 scale |

*These data were coded to preserve confidentiality.

FIGURE 16.13 SAMPLE CORRELATION COEFFICIENTS FOR THE CRAVENS DATA

```
          Sales    Time    Poten   AdvExp    Share   Change  Accounts    Work
Time      0.623
Poten     0.598   0.454
AdvExp    0.596   0.249   0.174
Share     0.484   0.106  -0.21     0.264
Change    0.489   0.251   0.268    0.377    0.085
Accounts  0.754   0.758   0.479    0.200    0.403    0.327
Work     -0.117  -0.179  -0.259   -0.272    0.349   -0.288    -0.199
Rating    0.402   0.101   0.359    0.411   -0.024    0.549     0.229    -0.277
```

Looking at the sample correlation coefficients between the independent variables, we see that the correlation between Time and Accounts is .758; hence, if Accounts were used as an independent variable, Time would not add much more explanatory power to the model. Recall the rule-of-thumb test from the discussion of multicollinearity in Section 15.4: Multicollinearity can cause problems if the absolute value of the sample correlation coefficient exceeds .7 for any two of the independent variables. If possible, then, we should avoid including both Time and Accounts in the same regression model. The sample correlation coefficient of .549 between Change and Rating is also high and may warrant further consideration.

Looking at the sample correlation coefficients between Sales and each of the independent variables can give us a quick indication of which independent variables are, by themselves, good predictors. We see that the single best predictor of Sales is Accounts, because it has the highest sample correlation coefficient (.754). Recall that for the case of one independent variable, the square of the sample correlation coefficient is the coefficient of determination. Thus, Accounts can explain $(.754)^2(100)$, or 56.85%, of the variability in Sales. The next most important independent variables are Time, Poten, and AdvExp, each with a sample correlation coefficient of approximately .6.

Although there are potential multicollinearity problems, let us consider developing an estimated regression equation using all eight independent variables. The Minitab computer package provided the results in Figure 16.14. The eight-variable multiple regression model has an R-Sq (adj) value of 88.3%. Note, however, that the $p$-values for the $t$ tests of individual parameters show that only Poten, AdvExp, and Share are significant at the $\alpha = .05$ level, given the effect of all the other variables. Hence, we might be inclined to investigate the results that would be obtained if we used just those three variables. Figure 16.15 shows the Minitab results obtained for the estimated regression equation with those three variables. We see that the estimated regression equation has an R-Sq (adj) value of 82.7%, which, although not quite as good as that for theeight-independent-variable estimated regression equation, is high.

How can we find an estimated regression equation that will do the best job given the data available? One approach is to compute all possible regressions. That is, we could develop 8 one-variable estimated regression equations (each of which corresponds to one of the independent variables), 28 two-variable estimated regression equations (the number of combinations of eight variables taken two at a time), and so on. In all, for the Cravens data, 255 different estimated regression equations involving one or more independent variables would have to be fitted to the data.

With the excellent computer packages available today, it is possible to compute all possible regressions. But doing so involves a great amount of computation and requires the model builder to review a large volume of computer output, much of which is associated with obviously poor models. Statisticians prefer a more systematic approach to selecting the subset of independent variables that provide the best estimated regression equation. In the next section, we introduce some of the more popular approaches.

FIGURE 16.14 MINITAB OUTPUT FOR THE MODEL INVOLVING ALL EIGHT INDEPENDENT VARIABLES

```
The regression equation is
Sales = - 1508 + 2.01 Time + 0.0372 Poten + 0.151 AdvExp + 199 Share
        + 291 Change + 5.55 Accounts + 19.8 Work + 8 Rating

Predictor      Coef   SE Coef      T      p
Constant    -1507.8     778.6  -1.94  0.071
Time          2.010     1.931   1.04  0.313
Poten      0.037206  0.008202   4.54  0.000
AdvExp      0.15094   0.04711   3.21  0.006
Share        199.08     67.03   2.97  0.009
Change        290.9     186.8   1.56  0.139
Accounts      5.550     4.775   1.16  0.262
Work          19.79     33.68   0.59  0.565
Rating          8.2     128.5   0.06  0.950

S = 449.015   R-Sq = 92.2%   R-Sq(adj) = 88.3%

Analysis of Variance

SOURCE          DF        SS       MS      F      p
Regression       8  38153712  4769214  23.66  0.000
Residual Error  16   3225837   201615
Total           24  41379549
```

FIGURE 16.15 MINITAB OUTPUT FOR THE MODEL INVOLVING Poten, AdvExp, AND Share

```
The regression equation is
Sales = - 1604 + 0.0543 Poten + 0.167 AdvExp + 283 Share

Predictor      Coef   SE Coef      T      p
Constant    -1603.6     505.6  -3.17  0.005
Poten      0.054286  0.007474   7.26  0.000
AdvExp      0.16748   0.04427   3.78  0.001
Share        282.75     48.76   5.80  0.000

S = 545.515   R-Sq = 84.9%   R-Sq(adj) = 82.7%

Analysis of Variance

SOURCE          DF        SS        MS      F      p
Regression       3  35130228  11710076  39.35  0.000
Residual Error  21   6249321    297587
Total           24  41379549
```

# 16.4 Variable Selection Procedures

*Variable selection procedures are particularly useful in the early stages of building a model, but they cannot substitute for experience and judgment on the part of the analyst.*

In this section we discuss four **variable selection procedures**: stepwise regression, forward selection, backward elimination, and best-subsets regression. Given a data set with several possible independent variables, we can use these procedures to identify which independent variables provide the best model. The first three procedures are iterative; at each step of the procedure a single independent variable is added or deleted and the new model is evaluated. The process continues until a stopping criterion indicates that the procedure cannot find a better model. The last procedure (best subsets) is not a one-variable-at-a-time procedure; it evaluates regression models involving different subsets of the independent variables.

In the stepwise regression, forward selection, and backward elimination procedures, the criterion for selecting an independent variable to add or delete from the model at each step is based on the $F$ statistic introduced in Section 16.2. Suppose, for instance, that we are considering adding $x_2$ to a model involving $x_1$ or deleting $x_2$ from a model involving $x_1$ and $x_2$. To test whether the addition or deletion of $x_2$ is statistically significant, the null and alternative hypotheses can be stated as follows:

$$H_0\colon \beta_2 = 0$$
$$H_\text{a}\colon \beta_2 \neq 0$$

In Section 16.2 (see equation (16.10)) we showed that

$$F = \frac{\dfrac{\text{SSE}(x_1) - \text{SSE}(x_1, x_2)}{1}}{\dfrac{\text{SSE}(x_1, x_2)}{n - p - 1}}$$

can be used as a criterion for determining whether the presence of $x_2$ in the model causes a significant reduction in the error sum of squares. The $p$-value corresponding to this $F$ statistic is the criterion used to determine whether an independent variable should be added or deleted from the regression model. The usual rejection rule applies: Reject $H_0$ if $p$-value $\leq \alpha$.

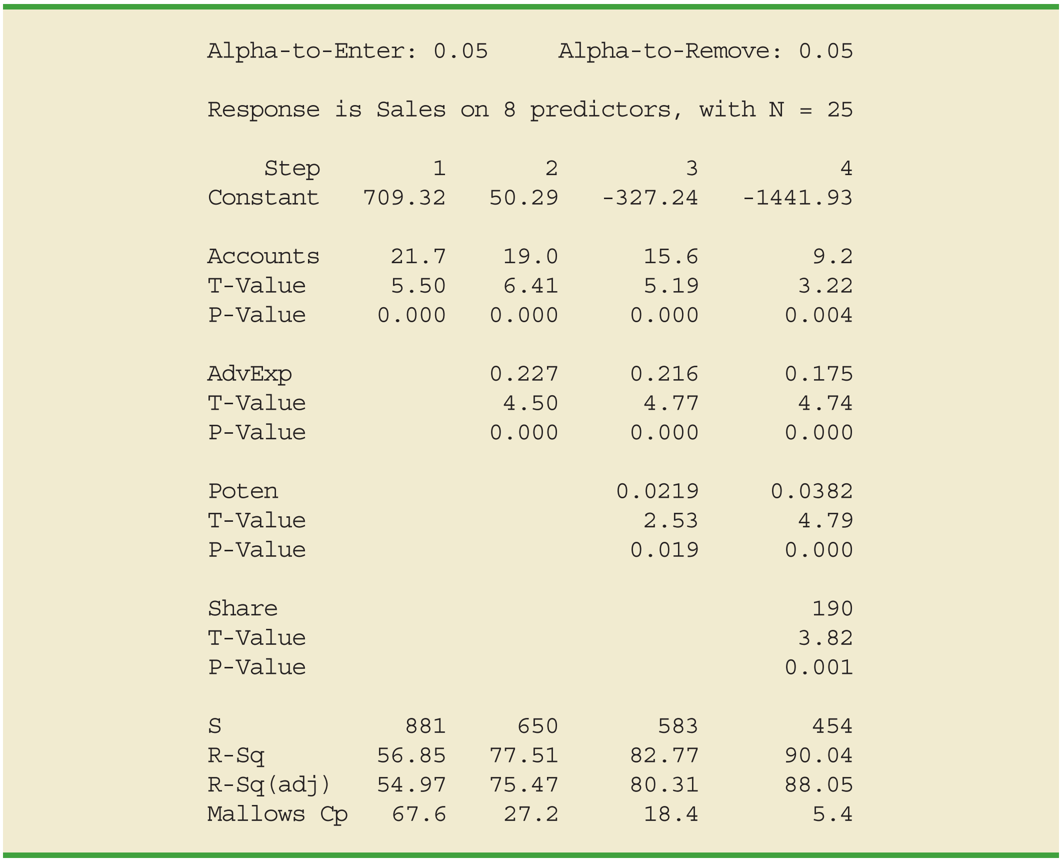
## Stepwise Regression

The stepwise regression procedure begins each step by determining whether any of the variables *already in the model* should be removed. It does so by first computing an $F$ statistic and a corresponding $p$-value for each independent variable in the model. The level of significance $\alpha$ for determining whether an independent variable should be removed from the model is referred to in Minitab as *Alpha to remove.* If the $p$-value for any independent variable is greater than *Alpha to remove,* the independent variable with the largest $p$-value is removed from the model and the stepwise regression procedure begins a new step.

If no independent variable can be removed from the model, the procedure attempts to enter another independent variable into the model. It does so by first computing an $F$ statistic and corresponding $p$-value for each independent variable that is not in the model. The level of significance $\alpha$ for determining whether an independent variable should be entered into the model is referred to in Minitab as *Alpha to enter.* The independent variable with the smallest $p$-value is entered into the model provided its $p$-value is less than or equal to *Alpha to enter.* The procedure continues in this manner until no independent variables can be deleted from or added to the model.

Figure 16.16 shows the results obtained by using the Minitab stepwise regression procedure for the Cravens data using values of .05 for *Alpha to remove* and .05 for *Alpha to enter.*

FIGURE 16.16 MINITAB STEPWISE REGRESSION OUTPUT FOR THE CRAVENS DATA



```
Alpha-to-Enter: 0.05     Alpha-to-Remove: 0.05

Response is Sales on 8 predictors, with N = 25

    Step           1        2         3          4
Constant      709.32    50.29   -327.24   -1441.93

Accounts        21.7     19.0      15.6        9.2
T-Value         5.50     6.41      5.19       3.22
P-Value        0.000    0.000     0.000      0.004

AdvExp                  0.227     0.216      0.175
T-Value                  4.50      4.77       4.74
P-Value                 0.000     0.000      0.000

Poten                            0.0219     0.0382
T-Value                            2.53       4.79
P-Value                           0.019      0.000

Share                                          190
T-Value                                       3.82
P-Value                                      0.001

S                881      650       583        454
R-Sq           56.85    77.51     82.77      90.04
R-Sq(adj)      54.97    75.47     80.31      88.05
Mallows Cp      67.6     27.2      18.4        5.4
```

The stepwise procedure terminated after four steps. The estimated regression equation identified by the Minitab stepwise regression procedure is

$$\hat{y} = -1441.93 + 9.2 \text{ Accounts} + .175 \text{ AdvExp} + .0382 \text{ Poten} + 190 \text{ Share}$$

*Because the stepwise procedure does not consider every possible subset for a given number of independent variables, it will not necessarily select the estimated regression equation with the highest R-sq value.*

Note also in Figure 16.16 that $s = \sqrt{\text{MSE}}$ has been reduced from 881 with the best one-variable model (using Accounts) to 454 after four steps. The value of R-sq has been increased from 56.85% to 90.04%, and the recommended estimated regression equation has an R-Sq(adj) value of 88.05%.

In summary, at each step of the stepwise regression procedure the first consideration is to see whether any independent variable can be removed from the current model. If none of the independent variables can be removed from the model, the procedure checks to see whether any of the independent variables that are not currently in the model can be entered. Because of the nature of the stepwise regression procedure, an independent variable can enter the model at one step, be removed at a subsequent step, and then enter the model at a later step. The procedure stops when no independent variables can be removed from or entered into the model.

## Forward Selection

The forward selection procedure starts with no independent variables. It adds variables one at a time using the same procedure as stepwise regression for determining whether an independent variable should be entered into the model. However, the forward selection procedure does not permit a variable to be removed from the model once it has been entered. The procedure stops if the *p*-value for each of the independent variables not in the model is greater than *Alpha to enter.*

The estimated regression equation obtained using Minitab's forward selection procedure is

$$\hat{y} = -1441.93 + 9.2 \text{ Accounts} + .175 \text{ AdvExp} + .0382 \text{ Poten} + 190 \text{ Share}$$

Thus, for the Cravens data, the forward selection procedure (using .05 for *Alpha to enter*) leads to the same estimated regression equation as the stepwise procedure.

## Backward Elimination

The backward elimination procedure begins with a model that includes all the independent variables. It then deletes one independent variable at a time using the same procedure as stepwise regression. However, the backward elimination procedure does not permit an independent variable to be reentered once it has been removed. The procedure stops when none of the independent variables in the model have a *p*-value greater than *Alpha to remove.*

The estimated regression equation obtained using Minitab's backward elimination procedure for the Cravens data (using .05 for *Alpha to remove*) is

$$\hat{y} = -1312 + 3.8 \text{ Time} + .0444 \text{ Poten} + .152 \text{ AdvExp} + 259 \text{ Share}$$

Comparing the estimated regression equation identified using the backward elimination procedure to the estimated regression equation identified using the forward selection procedure, we see that three independent variables—AdvExp, Poten, and Share—are common to both. However, the backward elimination procedure has included Time instead of Accounts.

*Forward selection and backward elimination may lead to different models.*

Forward selection and backward elimination are the two extremes of model building; the forward selection procedure starts with no independent variables in the model and adds independent variables one at a time, whereas the backward elimination procedure starts with all independent variables in the model and deletes variables one at a time. The two procedures may lead to the same estimated regression equation. It is possible, however, for them to lead to two different estimated regression equations, as we saw with the Cravens data. Deciding which estimated regression equation to use remains a topic for discussion. Ultimately, the analyst's judgment must be applied. The best-subsets model building procedure we discuss next provides additional model-building information to be considered before a final decision is made.

## Best-Subsets Regression

Stepwise regression, forward selection, and backward elimination are approaches to choosing the regression model by adding or deleting independent variables one at a time. None of them guarantees that the best model for a given number of variables will be found. Hence, these one-variable-at-a-time methods are properly viewed as heuristics for selecting a good regression model.

*The complete best-subsets output also includes values for the Mallows Cp statistic. More advanced texts discuss the use of this statistic.*

Some software packages use a procedure called best-subsets regression that enables the user to find, given a specified number of independent variables, the best regression model. Minitab has such a procedure. Figure 16.17 is a portion of the computer output obtained by using the best-subsets procedure for the Cravens data set.

This output identifies the two best one-variable estimated regression equations, the two best two-variable equations, the two best three-variable equations, and so on. The criterion used in determining which estimated regression equations are best for any number of

FIGURE 16.17 PORTION OF MINITAB BEST-SUBSETS REGRESSION OUTPUT

```
                                                A
                                                c
                                          A   C c   R
                                        P d S h o   a
                                      T o v h a u W t
                                      i t E a n n o I
                                      m e x r g t r n
Vars    R-Sq   R-Sq(adj)        S     e n p e e s K g

   1    56.8        55.0   881.09               X
   1    38.8        36.1   1049.3     X
   2    77.5        75.5   650.39         X     X
   2    74.6        72.3   691.11       X     X
   3    84.9        82.7   545.52       X X X
   3    82.8        80.3   582.64       X X     X
   4    90.0        88.1   453.84       X X X   X
   4    89.6        87.5   463.93     X X X X
   5    91.5        89.3   430.21     X X X X X
   5    91.2        88.9   436.75       X X X X X
   6    92.0        89.4   427.99     X X X X X X
   6    91.6        88.9   438.20       X X X X X X
   7    92.2        89.0   435.66     X X X X X X X
   7    92.0        88.8   440.29     X X X X X X   X
   8    92.2        88.3   449.02     X X X X X X X X
```

predictors is the value of the coefficient of determination (R-Sq). For instance, Accounts, with an R-Sq = 56.8%, provides the best estimated regression equation using only one independent variable; AdvExp and Accounts, with an R-Sq = 77.5%, provides the best estimated regression equation using two independent variables; and Poten, AdvExp, and Share, with an R-Sq = 84.9%, provides the best estimated regression equation with three independent variables. For the Cravens data, the adjusted coefficient of determination (R-Sq (adj) = 89.4%) is largest for the model with six independent variables: Time, Poten, AdvExp, Share, Change, and Accounts. However, the best model with four independent variables (Poten, AdvExp, Share, Accounts) has an adjusted coefficient of determination almost as high (R-Sq (adj) = 88.1%). All other things being equal, a simpler model with fewer variables is usually preferred.

## Making the Final Choice

The analysis performed on the Cravens data to this point is good preparation for choosing a final model, but more analysis should be conducted before the final choice. As we noted in Chapters 14 and 15, a careful analysis of the residuals should be done. We want the residual plot for the chosen model to resemble approximately a horizontal band. Let us assume the residuals are not a problem and that we want to use the results of the best-subsets procedure to help choose the model.

The best-subsets procedure shows us that the best four-variable model contains the independent variables Poten, AdvExp, Share, and Accounts. This result also happens to be the four-variable model identified with the stepwise regression procedure. Table 16.7 is helpful in making the final choice. It shows several possible models consisting of some or all of these four independent variables.

TABLE 16.7 SELECTED MODELS INVOLVING Accounts, AdvExp, Poten, AND Share

| Model | Independent Variables | R-Sq (adj) |
|---|---|---|
| 1 | Accounts | 55.0 |
| 2 | AdvExp, Accounts | 75.5 |
| 3 | Poten, Share | 72.3 |
| 4 | Poten, AdvExp, Accounts | 80.3 |
| 5 | Poten, AdvExp, Share | 82.7 |
| 6 | Poten, AdvExp, Share, Accounts | 88.1 |

From Table 16.7, we see that the model with just AdvExp and Accounts is good. The adjusted coefficient of determination is R-Sq (adj) = 75.5%, and the model with all four variables provides only a 12.6-percentage-point improvement. The simpler two-variable model might be preferred, for instance, if it is difficult to measure market potential (Poten). However, if the data are readily available and highly accurate predictions of sales are needed, the model builder would clearly prefer the model with all four variables.

## NOTES AND COMMENTS

**1.** The stepwise procedure requires that *Alpha to remove* be greater than or equal to *Alpha to enter.* This requirement prevents the same variable from being removed and then reentered at the same step.

**2.** Functions of the independent variables can be used to create new independent variables for use with any of the procedures in this section. For instance, if we wanted $x_1x_2$ in the model to account for interaction, we would use the data for $x_1$ and $x_2$ to create the data for $z = x_1x_2$.

**3.** None of the procedures that add or delete variables one at a time can be guaranteed to identify the best regression model. But they are excellent approaches to finding good models—especially when little multicollinearity is present.

## Exercises

### Applications

16. A study provided data on variables that may be related to the number of weeks a manufacturing worker has been jobless. The dependent variable in the study (Weeks) was defined as the number of weeks a worker has been jobless due to a layoff. The following independent variables were used in the study.

WEB file Layoffs

| | |
|---|---|
| Age | The age of the worker |
| Educ | The number of years of education |
| Married | A dummy variable; 1 if married, 0 otherwise |
| Head | A dummy variable; 1 if the head of household, 0 otherwise |
| Tenure | The number of years on the previous job |
| Manager | A dummy variable; 1 if management occupation, 0 otherwise |
| Sales | A dummy variable; 1 if sales occupation, 0 otherwise |

The data are available in the file named Layoffs.

a. Develop the best one-variable estimated regression equation.

b. Use the stepwise procedure to develop the best estimated regression equation. Use values of .05 for *Alpha to enter* and *Alpha to remove*.

c. Use the forward selection procedure to develop the best estimated regression equation. Use a value of .05 for *Alpha to enter*.

d. Use the backward elimination procedure to develop the best estimated regression equation. Use a value of .05 for *Alpha to remove*.

e. Use the best-subsets regression procedure to develop the best estimated regression equation.

WEB file

LPGATour2

17. The Ladies Professional Golfers Association (LPGA) maintains statistics on performance and earnings for members of the LPGA Tour. Year-end performance statistics for the 30 players who had the highest total earnings in LPGA Tour events for 2005 appear in the file named LPGATour2 (LPGA Tour website, 2006). Earnings ($1000) is the total earnings in thousands of dollars; Scoring Avg. is the average score for all events; Drive Average is the average length of a players drive in yards; Greens in Reg. is the percentage of time a player is able to hit the green in regulation; Putting Avg. is the average number of putts taken on greens hit in regulation; and Sand Saves is the percentage of time a player is able to get "up and down" once in a greenside sand bunker. A green is considered hit in regulation if any part of the ball is touching the putting surface and the difference between the value of par for the hole and the number of strokes taken to hit the green is at least 2. Let DriveGreens denote a new independent variable that represents the interaction between the average length of a player's drive and the percentage of time a player is able to hit the green in regulation. Use the methods in this section to develop the best estimated multiple regression equation for predicting a player's average score for all events.

18. Jeff Sagarin has been providing sports ratings for *USA Today* since 1985. In baseball his predicted RPG (runs/game) statistic takes into account the entire player's offensive statistics, and is claimed to be the best measure of a player's true offensive value. The following data show the RPG and a variety of offensive statistics for the 2005 Major League Baseball (MLB) season for 20 members of the New York Yankees (*USA Today* website, March 3, 2006). The labels on columns are defined as follows: RPG, predicted runs per game statistic; H, hits; 2B, doubles; 3B, triples; HR, home runs; RBI, runs batted in; BB, bases on balls (walks); SO, strikeouts; SB, stolen bases; CS, caught stealing; OBP, on-base percentage; SLG, slugging percentage; and AVG, batting average.

WEB file

Yankees

| **Player** | **RPG** | **H** | **2B** | **3B** | **HR** | **RBI** | **BB** | **SO** | **SB** | **CS** | **OBP** | **SLG** | **AVG** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| D Jeter | 6.51 | 202 | 25 | 5 | 19 | 70 | 77 | 117 | 14 | 5 | .389 | .450 | .309 |
| H Matsui | 6.32 | 192 | 45 | 3 | 23 | 116 | 63 | 78 | 2 | 2 | .367 | .496 | .305 |
| A Rodriguez | 9.06 | 194 | 29 | 1 | 48 | 130 | 91 | 139 | 21 | 6 | .421 | .610 | .321 |
| G Sheffield | 6.93 | 170 | 27 | 0 | 34 | 123 | 78 | 76 | 10 | 2 | .379 | .512 | .291 |
| R Cano | 5.01 | 155 | 34 | 4 | 14 | 62 | 16 | 68 | 1 | 3 | .320 | .458 | .297 |
| B Williams | 4.14 | 121 | 19 | 1 | 12 | 64 | 53 | 75 | 1 | 2 | .321 | .367 | .249 |
| J Posada | 5.36 | 124 | 23 | 0 | 19 | 71 | 66 | 94 | 1 | 0 | .352 | .430 | .262 |
| J Giambi | 9.11 | 113 | 14 | 0 | 32 | 87 | 108 | 109 | 0 | 0 | .440 | .535 | .271 |
| T Womack | 2.91 | 82 | 8 | 1 | 0 | 15 | 12 | 49 | 27 | 5 | .276 | .280 | .249 |
| T Martinez | 5.08 | 73 | 9 | 0 | 17 | 49 | 38 | 54 | 2 | 0 | .328 | .439 | .241 |
| M Bellhorn | 4.07 | 63 | 20 | 0 | 8 | 30 | 52 | 112 | 3 | 0 | .324 | .357 | .210 |
| R Sierra | 3.27 | 39 | 12 | 0 | 4 | 29 | 9 | 41 | 0 | 0 | .265 | .371 | .229 |
| J Flaherty | 1.83 | 21 | 5 | 0 | 2 | 11 | 6 | 26 | 0 | 0 | .206 | .252 | .165 |
| B Crosby | 3.48 | 27 | 0 | 1 | 1 | 6 | 4 | 14 | 4 | 1 | .304 | .327 | .276 |
| M Lawton | 5.15 | 6 | 0 | 0 | 2 | 4 | 7 | 8 | 1 | 0 | .263 | .250 | .125 |
| R Sanchez | 3.36 | 12 | 1 | 0 | 0 | 2 | 2 | 3 | 0 | 1 | .326 | .302 | .279 |
| A Phillips | 2.13 | 6 | 4 | 0 | 1 | 4 | 1 | 13 | 0 | 0 | .171 | .325 | .150 |
| M Cabrera | 1.19 | 4 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | .211 | .211 | .211 |
| R Johnson | 3.44 | 4 | 2 | 0 | 0 | 0 | 1 | 4 | 0 | 0 | .300 | .333 | .222 |
| F Escalona | 5.31 | 4 | 1 | 0 | 0 | 2 | 1 | 4 | 0 | 0 | .375 | .357 | .286 |

Let the dependent variable be the RPG statistic.

a. Develop the best one-variable estimated regression equation.

b. Use the methods in this section to develop the best estimated multiple regression equation for predicting a player's RPG.

WEB file

Stroke

19. Refer to exercise 14. Using age, blood pressure, whether a person is a smoker, and any interaction involving those variables, develop an estimated regression equation that can be used to predict risk. Briefly describe the process you used to develop an estimated regression equation for these data.

## 16.5 Multiple Regression Approach to Experimental Design

In Section 15.7 we discussed the use of dummy variables in multiple regression analysis. In this section we show how the use of dummy variables in a multiple regression equation can provide another approach to solving experimental design problems. We will demonstrate the multiple regression approach to experimental design by applying it to the Chemitech, Inc., completely randomized design introduced in Chapter 13.

Recall that Chemitech developed a new filtration system for municipal water supplies. The components for the new filtration system will be purchased from several suppliers, and Chemitech will assemble the components at its plant in Columbia, South Carolina. Three different assembly methods, referred to as methods A, B, and C, have been proposed. Managers at Chemitech want to determine which assembly method can produce the greatest number of filtration systems per week.

A random sample of 15 employees was selected, and each of the three assembly methods was randomly assigned to 5 employees. The number of units assembled by each employee is shown in Table 16.8. The sample mean number of units produced with each of the three assembly methods is as follows:

| Assembly Method | Mean Number Produced |
|---|---|
| A | 62 |
| B | 66 |
| C | 52 |

Although method B appears to result in higher production rates than either of the other methods, the issue is whether the three sample means observed are different enough for us to conclude that the means of the populations corresponding to the three methods of assembly are different.

We begin the regression approach to this problem by defining dummy variables that will be used to indicate which assembly method was used. Because the Chemitech problem has

TABLE 16.8 NUMBER OF UNITS PRODUCED BY 15 WORKERS

| | Method | |
|---|---|---|
| A | B | C |
| 58 | 58 | 48 |
| 64 | 69 | 57 |
| 55 | 71 | 59 |
| 66 | 64 | 47 |
| 67 | 68 | 49 |

**TABLE 16.9** DUMMY VARIABLES FOR THE CHEMITECH EXPERIMENT

| **A** | **B** | |
|---|---|---|
| 1 | 0 | Observation is associated with assembly method A |
| 0 | 1 | Observation is associated with assembly method B |
| 0 | 0 | Observation is associated with assembly method C |

three assembly methods or treatments, we need two dummy variables. In general, if the factor being investigated involves $k$ distinct levels or treatments, we need to define $k - 1$ dummy variables. For the Chemitech experiment we define dummy variables A and B as shown in Table 16.9.

We can use the dummy variables to relate the number of units produced per week, $y$, to the method of assembly the employee uses.

$$\begin{aligned} E(y) &= \text{Expected value of the number of units produced per week} \\ &= \beta_0 + \beta_1\text{A} + \beta_2\text{B} \end{aligned}$$

Thus, if we are interested in the expected value of the number of units assembled per week for an employee who uses method C, our procedure for assigning numerical values to the dummy variables would result in setting A = B = 0. The multiple regression equation then reduces to

$$E(y) = \beta_0 + \beta_1(0) + \beta_2(0) = \beta_0$$

We can interpret $\beta_0$ as the expected value of the number of units assembled per week for an employee who uses method C. In other words, $\beta_0$ is the mean number of units assembled per week using method C.

Next let us consider the forms of the multiple regression equation for each of the other methods. For method A the values of the dummy variables are A = 1 and B = 0, and

$$E(y) = \beta_0 + \beta_1(1) + \beta_2(0) = \beta_0 + \beta_1$$

For method B we set A = 0 and B = 1, and

$$E(y) = \beta_0 + \beta_1(0) + \beta_2(1) = \beta_0 + \beta_2$$

We see that $\beta_0 + \beta_1$ represents the mean number of units assembled per week using method A, and $\beta_0 + \beta_2$ represents the mean number of units assembled per week using method B.

We now want to estimate the coefficients $\beta_0$, $\beta_1$, and $\beta_2$ and hence develop an estimate of the mean number of units assembled per week for each method. Table 16.10 shows the sample data, consisting of 15 observations of A, B, and $y$. Figure 16.18 shows the corresponding Minitab multiple regression output. We see that the estimates of $\beta_0$, $\beta_1$, and $\beta_2$ are $b_0 = 52$, $b_1 = 10$, and $b_2 = 14$. Thus, the best estimate of the mean number of units assembled per week for each assembly method is as follows:

| **Assembly Method** | **Prediction of *E*(*y*)** |
|---|---|
| A | $b_0 + b_1 = 52 + 10 = 62$ |
| B | $b_0 = 52 + 14 = 66$ |
| C | $b_0 = 52$ |

TABLE 16.10 INPUT DATA FOR THE CHEMITECH COMPLETELY RANDOMIZED DESIGN

| A | B | $y$ |
|---|---|---|
| 1 | 0 | 58 |
| 1 | 0 | 64 |
| 1 | 0 | 55 |
| 1 | 0 | 66 |
| 1 | 0 | 67 |
| 0 | 1 | 58 |
| 0 | 1 | 69 |
| 0 | 1 | 71 |
| 0 | 1 | 64 |
| 0 | 1 | 68 |
| 0 | 0 | 48 |
| 0 | 0 | 57 |
| 0 | 0 | 59 |
| 0 | 0 | 47 |
| 0 | 0 | 49 |

WEB file

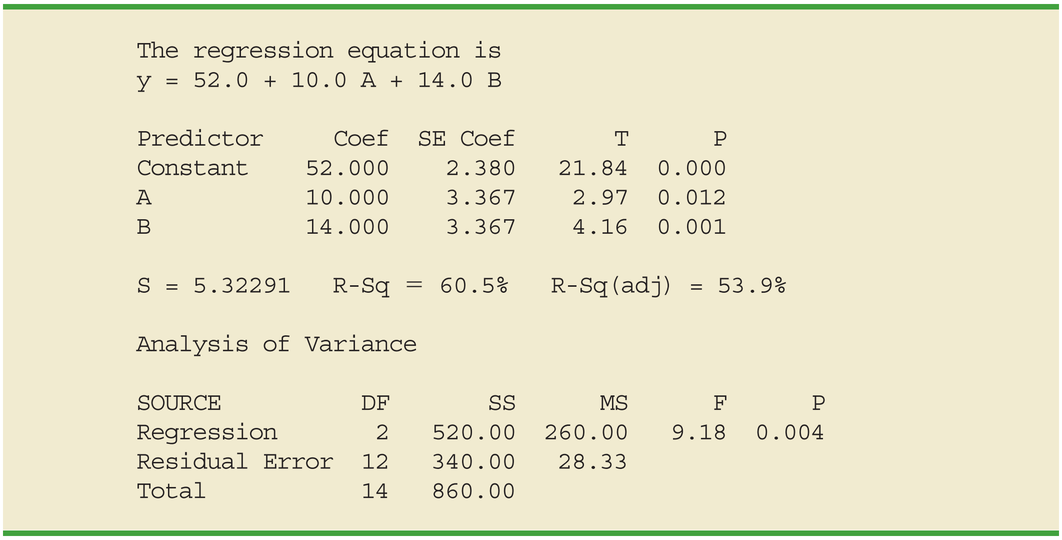
Chemitech

Note that the estimate of the mean number of units produced with each of the three assembly methods obtained from the regression analysis is the same as the sample mean shown previously.

Now let us see how we can use the output from the multiple regression analysis to perform the ANOVA test on the difference among the means for the three plants. First, we observe that if the means do not differ

$$E(y) \text{ for method A} - E(y) \text{ for method C} = 0$$
$$E(y) \text{ for method B} - E(y) \text{ for method C} = 0$$

FIGURE 16.18 MULTIPLE REGRESSION OUTPUT FOR THE CHEMITECH COMPLETELY RANDOMIZED DESIGN



```
The regression equation is
y = 52.0 + 10.0 A + 14.0 B

Predictor     Coef  SE Coef      T      P
Constant    52.000    2.380  21.84  0.000
A           10.000    3.367   2.97  0.012
B           14.000    3.367   4.16  0.001

S = 5.32291   R-Sq = 60.5%   R-Sq(adj) = 53.9%

Analysis of Variance

SOURCE          DF      SS      MS      F      P
Regression       2  520.00  260.00   9.18  0.004
Residual Error  12  340.00   28.33
Total           14  860.00
```

Because $\beta_0$ equals $E(y)$ for method C and $\beta_0 + \beta_1$ equals $E(y)$ for method A, the first difference is equal to $(\beta_0 + \beta_1) - \beta_0 = \beta_1$. Moreover, because $\beta_0 + \beta_2$ equals $E(y)$ for method B, the second difference is equal to $(\beta_0 + \beta_2) - \beta_0 = \beta_2$. We would conclude that the three methods do not differ if $\beta_1 = 0$ and $\beta_2 = 0$. Hence, the null hypothesis for a test for difference of means can be stated as

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

Suppose the level of significance is $\alpha = .05$. Recall that to test this type of null hypothesis about the significance of the regression relationship we use the $F$ test for overall significance. The Minitab output in Figure 16.18 shows that the $p$-value corresponding to $F = 9.18$ is .004. Because the $p$-value $= .004 < \alpha = .05$, we reject $H_0 : \beta_1 = \beta_2 = 0$ and conclude that the means for the three assembly methods are not the same. Because the $F$ test shows that the multiple regression relationship is significant, a $t$ test can be conducted to determine the significance of the individual parameters, $\beta_1$ and $\beta_2$. Using $\alpha = .05$, the $p$-values of .012 and .001 on the Minitab output indicate that we can reject $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$. Hence, both parameters are statistically significant. Thus, we can also conclude that the means for methods A and C are different and that the means for methods B and C are different.

## Exercises

### Methods

SELF test

20. Consider a completely randomized design involving four treatments: A, B, C, and D. Write a multiple regression equation that can be used to analyze these data. Define all variables.
21. Write a multiple regression equation that can be used to analyze the data for a randomized block design involving three treatments and two blocks. Define all variables.
22. Write a multiple regression equation that can be used to analyze the data for a two-factorial design with two levels for factor A and three levels for factor B. Define all variables.

### Applications

SELF test

23. The Jacobs Chemical Company wants to estimate the mean time (minutes) required to mix a batch of material on machines produced by three different manufacturers. To limit the cost of testing, four batches of material were mixed on machines produced by each of the three manufacturers. The times needed to mix the material follow.

| Manufacturer 1 | Manufacturer 2 | Manufacturer 3 |
|---|---|---|
| 20 | 28 | 20 |
| 26 | 26 | 19 |
| 24 | 31 | 23 |
| 22 | 27 | 22 |

a. Write a multiple regression equation that can be used to analyze the data.
b. What are the best estimates of the coefficients in your regression equation?

c. In terms of the regression equation coefficients, what hypotheses must we test to see whether the mean time to mix a batch of material is the same for all three manufacturers?
d. For an $\alpha = .05$ level of significance, what conclusion should be drawn?

24. Four different paints are advertised as having the same drying time. To check the manufacturers' claims, five samples were tested for each of the paints. The time in minutes until the paint was dry enough for a second coat to be applied was recorded for each sample. The data obtained follow.

| Paint 1 | Paint 2 | Paint 3 | Paint 4 |
|---|---|---|---|
| 128 | 144 | 133 | 150 |
| 137 | 133 | 143 | 142 |
| 135 | 142 | 137 | 135 |
| 124 | 146 | 136 | 140 |
| 141 | 130 | 131 | 153 |

a. Use $\alpha = .05$ to test for any significant differences in mean drying time among the paints.
b. What is your estimate of the mean drying time for paint 2? How is it obtained from the computer output?

25. An automobile dealer conducted a test to determine whether the time needed to complete a minor engine tune-up depends on whether a computerized engine analyzer or an electronic analyzer is used. Because tune-up time varies among compact, intermediate, and full-sized cars, the three types of cars were used as blocks in the experiment. The data (time in minutes) obtained follow.

| | | | Car | |
|---|---|---|---|---|
| | | Compact | Intermediate | Full Size |
| Analyzer | Computerized | 50 | 55 | 63 |
| | Electronic | 42 | 44 | 46 |

Use $\alpha = .05$ to test for any significant differences.

26. A mail-order catalog firm designed a factorial experiment to test the effect of the size of a magazine advertisement and the advertisement design on the number (in thousands) of catalog requests received. Three advertising designs and two sizes of advertisements were considered. The following data were obtained. Test for any significant effects due to type of design, size of advertisement, or interaction. Use $\alpha = .05$.

| | | Size of Advertisement | |
|---|---|---|---|
| | | Small | Large |
| Design | A | 8 | 12 |
| | | 12 | 8 |
| | B | 22 | 26 |
| | | 14 | 30 |
| | C | 10 | 18 |
| | | 18 | 14 |

## 16.6 Autocorrelation and the Durbin-Watson Test

Often, the data used for regression studies in business and economics are collected over time. It is not uncommon for the value of $y$ at time $t$, denoted by $y_t$, to be related to the value of $y$ at previous time periods. In such cases, we say **autocorrelation** (also called **serial correlation**) is present in the data. If the value of $y$ in time period $t$ is related to its value in time period $t - 1$, first-order autocorrelation is present. If the value of $y$ in time period $t$ is related to the value of $y$ in time period $t - 2$, second-order autocorrelation is present, and so on.

One of the assumptions of the regression model is the error terms are independent. However, when autocorrelation is present, this assumption is violated. In the case of first-order autocorrelation, the error at time $t$, denoted $\epsilon_t$, will be related to the error at time period $t - 1$, denoted $\epsilon_{t-1}$. Two cases of first-order autocorrelation are illustrated in Figure 16.19. Panel A is the case of positive autocorrelation; panel B is the case of negative autocorrelation. With positive autocorrelation we expect a positive residual in one period to be followed by a positive residual in the next period, a negative residual in one period to be followed by a negative residual in the next period, and so on. With negative autocorrelation, we expect a positive residual in one period to be followed by a negative residual in the next period, then a positive residual, and so on.

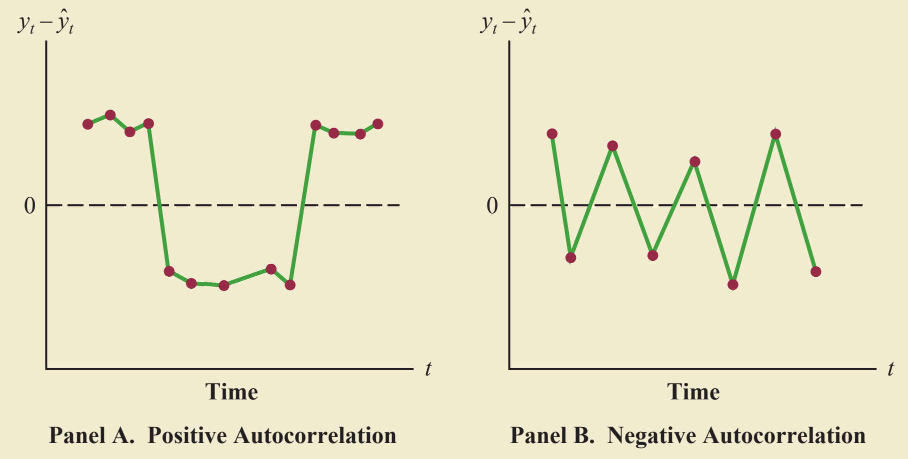
When autocorrelation is present, serious errors can be made in performing tests of statistical significance based upon the assumed regression model. It is therefore important to be able to detect autocorrelation and take corrective action. We will show how the Durbin-Watson statistic can be used to detect first-order autocorrelation.

Suppose the values of $\epsilon$ are not independent but are related in the following manner:

$$\epsilon_t = \rho\epsilon_{t-1} + z_t \qquad \textbf{(16.16)}$$

where $\rho$ is a parameter with an absolute value less than one and $z_t$ is a normally and independently distributed random variable with a mean of zero and a variance of $\sigma^2$. From equation (16.16) we see that if $\rho = 0$, the error terms are not related, and each has a mean of zero and a variance of $\sigma^2$. In this case, there is no autocorrelation and the regression assumptions

FIGURE 16.19 TWO DATA SETS WITH FIRST-ORDER AUTOCORRELATION



are satisfied. If $\rho > 0$, we have positive autocorrelation; if $\rho < 0$, we have negative autocorrelation. In either of these cases, the regression assumptions about the error term are violated.

The **Durbin-Watson test** for autocorrelation uses the residuals to determine whether $\rho = 0$. To simplify the notation for the Durbin-Watson statistic, we denote the $i$th residual by $e_i = y_i - \hat{y}_i$. The Durbin-Watson test statistic is computed as follows.

DURBIN-WATSON TEST STATISTIC

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \tag{16.17}$$

If successive values of the residuals are close together (positive autocorrelation), the value of the Durbin-Watson test statistic will be small. If successive values of the residuals are far apart (negative autocorrelation), the value of the Durbin-Watson statistic will be large.

The Durbin-Watson test statistic ranges in value from zero to four, with a value of two indicating no autocorrelation is present. Durbin and Watson developed tables that can be used to determine when their test statistic indicates the presence of autocorrelation. Table 16.11 shows lower and upper bounds ($d_L$ and $d_U$) for hypothesis tests using $\alpha = .05$; $n$ denotes the number of observations. The null hypothesis to be tested is always that there is no autocorrelation.

$$H_0\colon \rho = 0$$

The alternative hypothesis to test for positive autocorrelation is

$$H_a\colon \rho > 0$$

**TABLE 16.11** CRITICAL VALUES FOR THE DURBIN-WATSON TEST FOR AUTOCORRELATION

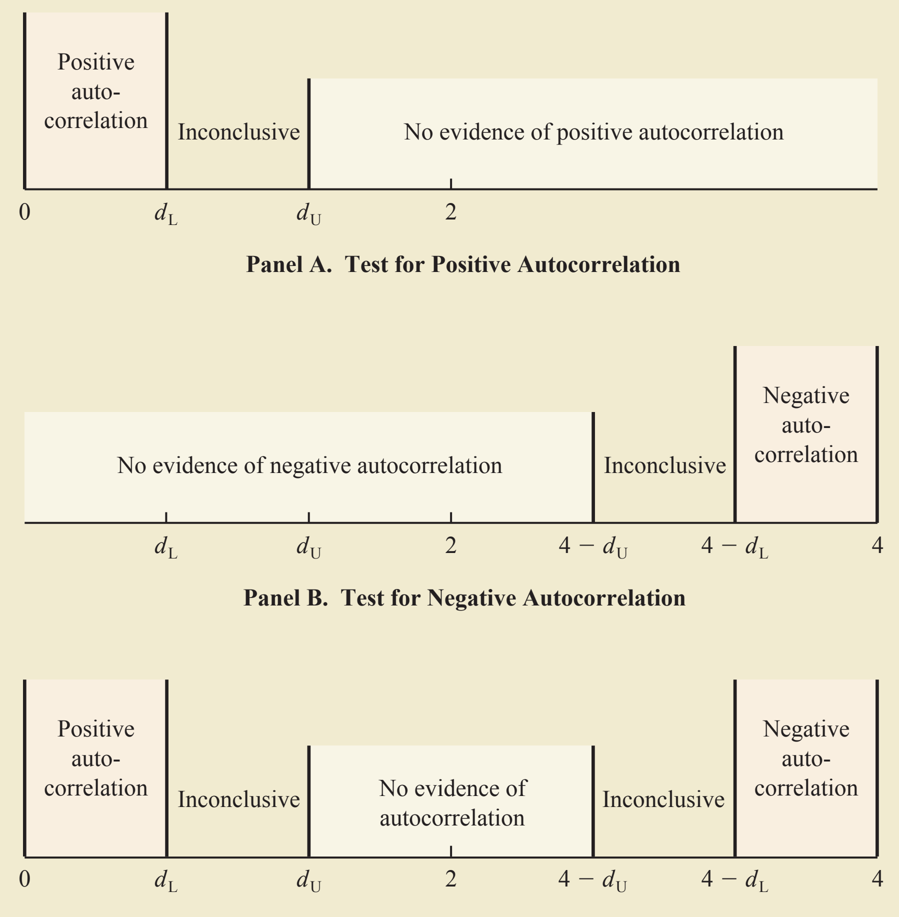
*Note:* Entries in the table are the critical values for a one-tailed Durbin-Watson test for autocorrelation. For a two-tailed test, the level of significance is doubled.

**Significance Points of $d_L$ and $d_U$: $\alpha = .05$**
**Number of Independent Variables**

| | **1** | | **2** | | **3** | | **4** | | **5** | |
|---|---|---|---|---|---|---|---|---|---|---|
| ***n**** | $d_L$ | $d_U$ | $d_L$ | $d_U$ | $d_L$ | $d_U$ | $d_L$ | $d_U$ | $d_L$ | $d_U$ |
| 15 | 1.08 | 1.36 | .95 | 1.54 | .82 | 1.75 | .69 | 1.97 | .56 | 2.21 |
| 20 | 1.20 | 1.41 | 1.10 | 1.54 | 1.00 | 1.68 | .90 | 1.83 | .79 | 1.99 |
| 25 | 1.29 | 1.45 | 1.21 | 1.55 | 1.12 | 1.66 | 1.04 | 1.77 | .95 | 1.89 |
| 30 | 1.35 | 1.49 | 1.28 | 1.57 | 1.21 | 1.65 | 1.14 | 1.74 | 1.07 | 1.83 |
| 40 | 1.44 | 1.54 | 1.39 | 1.60 | 1.34 | 1.66 | 1.29 | 1.72 | 1.23 | 1.79 |
| 50 | 1.50 | 1.59 | 1.46 | 1.63 | 1.42 | 1.67 | 1.38 | 1.72 | 1.34 | 1.77 |
| 70 | 1.58 | 1.64 | 1.55 | 1.67 | 1.52 | 1.70 | 1.49 | 1.74 | 1.46 | 1.77 |
| 100 | 1.65 | 1.69 | 1.63 | 1.72 | 1.61 | 1.74 | 1.59 | 1.76 | 1.57 | 1.78 |

*Interpolate linearly for intermediate *n* values.

FIGURE 16.20 HYPOTHESIS TEST FOR AUTOCORRELATION USING THE DURBIN-WATSON TEST


Positive auto-correlation
Inconclusive
No evidence of positive autocorrelation
0
$d_L$
$d_U$
2

Panel A. Test for Positive Autocorrelation

No evidence of negative autocorrelation
Inconclusive
Negative auto-correlation
$d_L$
$d_U$
2
$4 - d_U$
$4 - d_L$
4

Panel B. Test for Negative Autocorrelation

Positive auto-correlation
Inconclusive
No evidence of autocorrelation
Inconclusive
Negative auto-correlation
0
$d_L$
$d_U$
2
$4 - d_U$
$4 - d_L$
4

Panel C. Two-Sided Test for Autocorrelation

The alternative hypothesis to test for negative autocorrelation is

$$H_a: \rho < 0$$

A two-sided test is also possible. In this case the alternative hypothesis is

$$H_a: \rho \neq 0$$

Figure 16.20 shows how the values of $d_L$ and $d_U$ in Table 16.11 are used to test for autocorrelation. Panel A illustrates the test for positive autocorrelation. If $d < d_L$, we conclude that positive autocorrelation is present. If $d_L \leq d \leq d_U$, we say the test is inconclusive. If $d > d_U$, we conclude that there is no evidence of positive autocorrelation.

Panel B illustrates the test for negative autocorrelation. If $d > 4 - d_L$, we conclude that negative autocorrelation is present. If $4 - d_U \leq d \leq 4 - d_L$, we say the test is inconclusive. If $d < 4 - d_U$, we conclude that there is no evidence of negative autocorrelation.

Panel C illustrates the two-sided test. If $d < d_L$ or $d > 4 - d_L$, we reject $H_0$ and conclude that autocorrelation is present. If $d_L \leq d \leq d_U$ or $4 - d_U \leq d \leq 4 - d_L$, we say the test is inconclusive. If $d_U < d < 4 - d_U$, we conclude that there is no evidence of autocorrelation.

If significant autocorrelation is identified, we should investigate whether we omitted one or more key independent variables that have time-ordered effects on the dependent variable. If no such variables can be identified, including an independent variable that measures the time of the observation (for instance, the value of this variable could be one for the first observation, two for the second observation, and so on) will sometimes eliminate or reduce the autocorrelation. When these attempts to reduce or remove autocorrelation do not work, transformations on the dependent or independent variables can prove helpful; a discussion of such transformations can be found in more advanced texts on regression analysis.

Note that the Durbin-Watson tables list the smallest sample size as 15. The reason is that the test is generally inconclusive for smaller sample sizes; in fact, many statisticians believe the sample size should be at least 50 for the test to produce worthwhile results.

## Exercises

### Applications

27. The following data show the daily closing prices (in dollars per share) for a stock.

WEB file
ClosingPrice

| Date | Price ($) |
|---|---|
| Nov. 3 | 82.87 |
| Nov. 4 | 83.00 |
| Nov. 7 | 83.61 |
| Nov. 8 | 83.15 |
| Nov. 9 | 82.84 |
| Nov. 10 | 83.99 |
| Nov. 11 | 84.55 |
| Nov. 14 | 84.36 |
| Nov. 15 | 85.53 |
| Nov. 16 | 86.54 |
| Nov. 17 | 86.89 |
| Nov. 18 | 87.77 |
| Nov. 21 | 87.29 |
| Nov. 22 | 87.99 |
| Nov. 23 | 88.80 |
| Nov. 25 | 88.80 |
| Nov. 28 | 89.11 |
| Nov. 29 | 89.10 |
| Nov. 30 | 88.90 |
| Dec. 1 | 89.21 |

a. Define the independent variable Period, where Period = 1 corresponds to the data for November 3, Period = 2 corresponds to the data for November 4, and so on. Develop the estimated regression equation that can be used to predict the closing price given the value of Period.

b. At the .05 level of significance, test for any positive autocorrelation in the data.

28. Refer to the Cravens data set in Table 16.5. In Section 16.3 we showed that the estimated regression equation involving Accounts, AdvExp, Poten, and Share had an adjusted coefficient