CHAPTER 14

Building Multiple Regression Models

LEARNING OBJECTIVES

This chapter presents several advanced topics in multiple regression analysis, enabling you to:

1. Generalize linear regression models as polynomial regression models using model transformation and Tukey's ladder of transformation, accounting for possible interaction among the independent variables
2. Examine the role of indicator, or dummy, variables as predictors or independent variables in multiple regression analysis
3. Use all possible regressions, stepwise regression, forward selection, and backward elimination search procedures to develop regression models that account for the most variation in the dependent variable and are parsimonious
4. Recognize when multicollinearity is present, understanding general techniques for preventing and controlling it

Panoramic Images/Getty Images

Determining Compensation for CEOs

Decision Dilemma

Chief executive officers for large companies receive widely varying amounts of compensation for their work. Why is the range so wide? What are some of the variables that seem to contribute to the diversity of CEO compensation packages?

As a starting place, one might examine the role of company size as measured by sales volume, number of employees, number of plants, and so on in driving CEO compensation. It could be argued that CEOs of larger companies carry larger responsibilities and hence should receive higher compensation. Some researchers believe CEO compensation is related to such things as industry performance of the firm, percentage of stock that has outside ownership, and proportion of insiders on the board. At least a significant proportion of CEOs are likely to be compensated according to the performance of their companies during the fiscal period preceding compensation. Company performance can be measured by such variables as earnings per share, percentage change in profit, sales, and profit. In addition, some theorize that companies with outside ownership are more oriented toward declaring dividends to stockholders than toward large CEO compensation packages.

Do CEOs' individual and family characteristics play a role in their compensation? Do such things as CEO age, degrees obtained, marital status, military experience, and number of children matter in compensation? Do type of industry and geographic location of the company matter? What are the significant factors in determining CEO compensation?

What follow are CEO compensation data generated by using management compensation models published by Wyatt Data Services. In the first column on the left are cash compensation figures (in $1,000) for 20 CEOs. Those figures represent salary, bonuses, and any other cash remuneration given to the CEO as part of compensation. The four columns to the right contain data on four variables associated with each CEO's company: sales, number of employees, capital investment, and whether the company is in manufacturing. Sales figures and capital investment figures are given in $ millions.

Cash Compensation	Sales	Number of Employees	Capital Investment	Manufacturing
212	35.0	248.00	10.5	1
226	27.2	156.00	3.8	0
237	49.5	348.00	14.9	1
239	34.0	196.00	5.0	0
242	52.8	371.00	15.9	1
245	37.6	216.00	5.7	0
253	60.7	425.00	18.3	1
262	49.2	285.00	8.0	0
271	75.1	524.00	22.6	1
285	69.0	401.00	12.3	0
329	137.2	947.00	41.4	1
340	140.1	825.00	30.3	0
353	162.9	961.00	36.7	0
384	221.7	1517.00	67.1	1
405	261.6	1784.00	79.2	1
411	300.1	1788.00	79.8	0
456	455.5	2733.00	135.7	0
478	437.6	2957.00	132.7	1
525	802.1	4857.00	278.4	0
564	731.5	4896.00	222.2	1

Managerial and Statistical Questions

1. Can a model be developed to predict CEO compensation?
2. If a model is developed, how can the model be evaluated to determine whether it is valid?
3. Is it possible to sort out variables that appear to be related to CEO compensation and determine which variables are more significant predictors?
4. Are some of the variables related to CEO compensation but in a nonlinear manner?
5. Are some variables highly interrelated and redundant in their potential for determining CEO compensation?

Sources: Adapted from Jeffrey L. Kerr and Leslie Kren,"Effect of Relative Decision Monitoring on Chief Executive Compensation," *Academy of Management Journal*, vol. 35, no. 2 (June 1992). Used with permission. Robin L. Bartlett, James H. Grant, and Timothy I. Miller,"The Earnings of Top Executives: Compensating Differentials for Risky Business," *Quarterly Reviews of Economics and Finance*, vol. 32, no. 1 (Spring 1992). Used with permission. Database derived using models published in *1993/1994 Top Management Compensation Regression Analysis Report*, 44th ed. Fort Lee, NJ: Wyatt Data Services/ECS, December 1994.

14.1 NONLINEAR MODELS: MATHEMATICAL TRANSFORMATION

The regression models presented thus far are based on the general linear regression model, which has the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \in, \tag{14.1}$$

where

β_0 = the regression constant
$\beta_1, \beta_2 \ldots, \beta_k$ are the partial regression coefficients for the k independent variables
$x_1, \ldots, x_k$ are the independent variables
k = the number of independent variables

In this general linear model, the parameters, β_i, are linear. It does not mean, however, that the dependent variable, y, is necessarily linearly related to the predictor variables. Scatter plots sometimes reveal a curvilinear relationship between x and y. Multiple regression response surfaces are not restricted to linear surfaces and may be curvilinear.

To this point, the variables, x_i, have represented different predictors. For example, in the real estate example presented in Chapter 13, the variables—x_1, x_2—represented two predictors: number of square feet in the house and the age of the house, respectively. Certainly, regression models can be developed for more than two predictors. For example, a marketing site location model could be developed in which sales, as the response variable, is predicted by population density, number of competitors, size of the store, and number of salespeople. Such a model could take the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \in$$

This regression model has four x_i variables, each of which represents a different predictor.

The general linear model also applies to situations in which some x_i represent recoded data from a predictor variable already represented in the model by another independent variable. In some models, x_i represents variables that have undergone a mathematical transformation to allow the model to follow the form of the general linear model.

In this section of this chapter, we explore some of these other models, including polynomial regression models, regression models with interaction, and models with transformed variables.

Polynomial Regression

Regression models in which the highest power of any predictor variable is 1 and in which there are no interaction terms—cross products ($x_i \cdot x_j$)—are referred to as *first-order models*. Simple regression models like those presented in Chapter 12 are *first-order models with one independent variable*. The general model for simple regression is

$$y = \beta_0 + \beta_1 x_1 + \in$$

If a second independent variable is added, the model is referred to as a first-order model with two independent variables and appears as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \in$$

Polynomial regression models are regression models that are second- or higher-order models. They contain squared, cubed, or higher powers of the predictor variable(s) and contain response surfaces that are curvilinear. Yet, they are still special cases of the general linear model given in formula 14.1.

Consider a regression model with one independent variable where the model includes a second predictor, which is the independent variable squared. Such a model is referred to as a second-order model with one independent variable because the highest power among the predictors is 2, but there is still only one independent variable. This model takes the

TABLE 14.1
Sales Data for 13 Manufacturing Companies

Manufacturer	Sales ($ millions)	Number of Manufacturing Representatives
1	2.1	2
2	3.6	1
3	6.2	2
4	10.4	3
5	22.8	4
6	35.6	4
7	57.1	5
8	83.5	5
9	109.4	6
10	128.6	7
11	196.8	8
12	280.0	10
13	462.3	11

following form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \in$$

This model can be used to explore the possible fit of a quadratic model in predicting a dependent variable. A **quadratic model** is *a multiple regression model in which the predictors are a variable and the square of the variable.* How can this be a special case of the general linear model? Let x_2 of the general linear model be equal to x_1^2 then $y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \in$, becomes $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \in$. Through what process does a researcher go to develop the regression constant and coefficients for a curvilinear model such as this one?

Multiple regression analysis assumes a linear fit of the regression coefficients and regression constant, but not necessarily a linear relationship of the independent variable values (x). Hence, a researcher can often accomplish curvilinear regression by recoding the data before the multiple regression analysis is attempted.

As an example, consider the data given in Table 14.1. This table contains sales volumes (in $ millions) for 13 manufacturing companies along with the number of manufacturer's representatives associated with each firm. A simple regression analysis to predict sales by the number of manufacturer's representatives results in the Excel output in Figure 14.1.

FIGURE 14.1
Excel Simple Regression Output for Manufacturing Example

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.933
R Square	0.870
Adjusted R Square	0.858
Standard Error	51.098
Observations	13

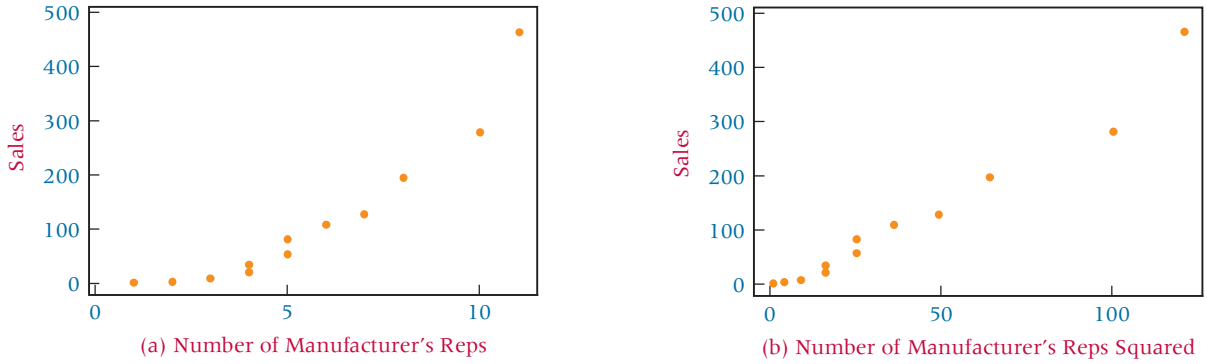
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	192395.416	192395.416	73.69	0.0000033
Residual	11	28721.452	2611.041		
Total	12	221116.868			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	–107.029	28.737	–3.72	0.0033561
Reps	41.026	4.779	8.58	0.0000033

FIGURE 14.2

Minitab Scatter Plots of Manufacturing Data



This regression output shows a regression model with an r^2 of 87.0%, a standard error of the estimate equal to 51.10, a significant overall F test for the model, and a significant t ratio for the predictor number of manufacturer's representatives.

Figure 14.2(a) is a scatter plot for the data in Table 14.1. Notice that the plot of number of representatives and sales is not a straight line and is an indication that the relationship between the two variables may be curvilinear. To explore the possibility that a quadratic relationship may exist between sales and number of representatives, the business researcher creates a second predictor variable, (number of manufacturer's representatives)^2, to use in the regression analysis to predict sales along with number of manufacturer's representatives, as shown in Table 14.2. Thus, a variable can be created to explore second-order parabolic relationships by squaring the data from the independent variable of the linear model and entering it into the analysis. Figure 14.2(b) is a scatter plot of sales with (number of manufacturer's reps)^2. Note that this graph, with the squared term, more closely approaches a straight line than does the graph in Figure 14.2(a). By recoding the predictor variable, the researcher creates a potentially better regression fit.

With these data, a multiple regression model can be developed. Figure 14.3 shows the Excel output for the regression analysis to predict sales by number of manufacturer's representatives and (number of manufacturer's representatives)^2.

Examine the output in Figure 14.3 and compare it with the output in Figure 14.1 for the simple regression model. The R^2 for this model is 97.3%, which is an increase from the r^2 of 87.0% for the single linear predictor model. The standard error of the estimate for this model is 24.59, which is considerably lower than the 51.10 value obtained from the simple

TABLE 14.2

Display of Manufacturing Data with Newly Created Variable

Manufacturer	Sales ($ millions) y	Number of Mgfr. Reps x_1	Number of (Mgfr. Reps)^2 $x_2 = (x_1)^2$
1	2.1	2	4
2	3.6	1	1
3	6.2	2	4
4	10.4	3	9
5	22.8	4	16
6	35.6	4	16
7	57.1	5	25
8	83.5	5	25
9	109.4	6	36
10	128.6	7	49
11	196.8	8	64
12	280.0	10	100
13	462.3	11	121

FIGURE 14.3

Excel Output for Quadratic Model of Manufacturing Example

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.986
R Square	0.973
Adjusted R Square	0.967
Standard Error	24.593
Observations	13

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	215068.6001	107534.3000	177.79	0.000000015
Residual	10	6048.2676	604.8268		
Total	12	221116.8677			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	18.067	24.673	0.73	0.4808194
Reps	–15.723	9.550	–1.65	0.1307046
Reps Squared	4.750	0.776	6.12	0.0001123

regression model. Remember, the sales figures were $ millions. The quadratic model reduced the standard error of the estimate by 26.51($1,000,000), or $26,510,000. It appears that the quadratic model is a better model for predicting sales.

An examination of the *t* statistic for the squared term and its associated probability in Figure 14.3 shows that it is statistically significant at $\alpha = .001$ ($t = 6.12$ with a probability of .0001). If this *t* statistic were not significant, the researcher would most likely drop the squared term and revert to the first-order model (simple regression model).

In theory, third- and higher-order models can be explored. Generally, business researchers tend to utilize first- and second-order regression models more than higher-order models. Remember that most regression analysis is used in business to aid decision making. Higher-power models (third, fourth, etc.) become difficult to interpret and difficult to explain to decision makers. In addition, the business researcher is usually looking for trends and general directions. The higher the order in regression modeling, the more the model tends to follow irregular fluctuations rather than meaningful directions.

Tukey's Ladder of Transformations

As just shown with the manufacturing example, recoding data can be a useful tool in improving the regression model fit. Many other ways of recoding data can be explored in this process. John W. Tukey* presented a "ladder of expressions" that can be explored to straighten out a plot of *x* and *y*, thereby offering potential improvement in the predictability of the regression model. **Tukey's ladder of transformations** gives the following expressions for both *x* and *y*.

Ladder for *x*

← Up Ladder ↓Neutral Down Ladder →

$$\ldots, x^4, x^3, x^2, x, \sqrt{x}, x, \log x, -\frac{1}{\sqrt{x}}, -\frac{1}{x}, -\frac{1}{x^2}, -\frac{1}{x^3}, -\frac{1}{x^4}, \ldots$$

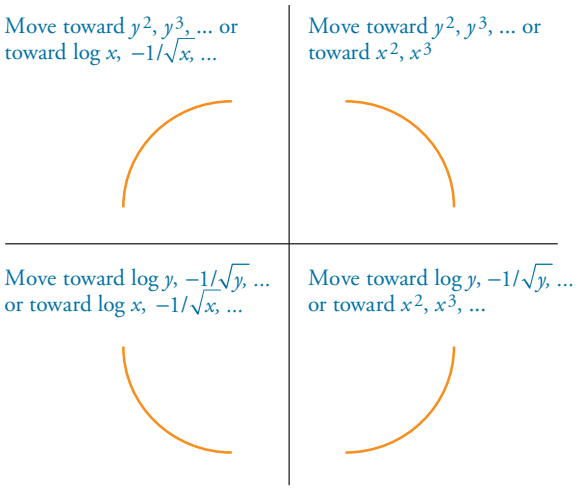
Ladder for *y*

← Up Ladder ↓Neutral Down Ladder →

$$\ldots, y^4, y^3, y^2, y, \sqrt{y}, y, \log y, -\frac{1}{\sqrt{y}}, -\frac{1}{y}, -\frac{1}{y^2}, -\frac{1}{y^3}, -\frac{1}{y^4}, \ldots$$

*John W. Tukey, *Exploratory Data Analysis*. Reading, MA: Addison-Wesley, 1977.

FIGURE 14.4
Tukey's Four-Quadrant Approach



These ladders suggest to the user potential ways to recode the data. Tukey published a **four-quadrant approach** to determining which expressions on the ladder are more appropriate for a given situation. This approach is based on the shape of the scatter plot of x and y. Figure 14.4 shows the four quadrants and the associated recoding expressions. For example, if the scatter plot of x and y indicates a shape like that shown in the upper left quadrant, recoding should move "down the ladder" for the x variable toward

$$\log x, -\frac{1}{\sqrt{x}}, -\frac{1}{x}, -\frac{1}{x^2}, -\frac{1}{x^3}, -\frac{1}{x^4}, \ldots$$

or "up the ladder" for the y variable toward

$$y^2, y^3, y^4, \ldots$$

Or, if the scatter plot of x and y indicates a shape like that of the lower right quadrant, the recoding should move "up the ladder" for the x variable toward

$$x^2, x^3, x^4, \ldots$$

or "down the ladder" for the y variable toward

$$\log y, -\frac{1}{\sqrt{y}}, -\frac{1}{y}, -\frac{1}{y^2}, -\frac{1}{y^3}, -\frac{1}{y^4}, \ldots$$

In the manufacturing example, the graph in Figure 14.2(a) is shaped like the curve in the lower right quadrant of Tukey's four-quadrant approach. His approach suggests that the business researcher move "up the ladder" on x as was done by using the squared term. The researcher could have explored other options such as continuing on up the ladder of x or going down the ladder of y. Tukey's ladder is a continuum and leaves open other recoding possibilities between the expressions. For example, between x^2 and x^3 are many possible powers of x that can be explored, such as $x^{2.1}$, $x^{2.5}$, or $x^{2.86}$.

Regression Models with Interaction

Often when two different independent variables are used in a regression analysis, an *interaction* occurs between the two variables. This interaction was discussed in Chapter 11 in two-way analysis of variance, where one variable will act differently over a given range of values for the second variable than it does over another range of values for the second variable. For example, in a manufacturing plant, temperature and humidity might interact in such a way as to have an effect on the hardness of the raw material. The air humidity may affect the raw material differently at different temperatures.

In regression analysis, interaction can be examined as a separate independent variable. An interaction predictor variable can be designed by multiplying the data values of one

TABLE 14.3

Prices of Three Stocks over a 15 Month Period

Stock 1	Stock 2	Stock 3
41	36	35
39	36	35
38	38	32
45	51	41
41	52	39
43	55	55
47	57	52
49	58	54
41	62	65
35	70	77
36	72	75
39	74	74
33	83	81
28	101	92
31	107	91

variable by the values of another variable, thereby creating a new variable. A model that includes an interaction variable is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$$

The x_1x_2 term is the interaction term. Even though this model has 1 as the highest power of any one variable, it is considered to be a second-order equation because of the x_1x_2 term.

Suppose the data in Table 14.3 represent the closing stock prices for three corporations over a period of 15 months. An investment firm wants to use the prices for stocks 2 and 3 to develop a regression model to predict the price of stock 1. The form of the general linear regression equation for this model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

where

y = price of stock 1
x_1 = price of stock 2
x_2 = price of stock 3

Using Minitab to develop this regression model, the firm's researcher obtains the first output displayed in Figure 14.5. This regression model is a first-order model with two predictors, x_1 and x_2. This model produced a modest R^2 of .472. Both of the t ratios are small and statistically nonsignificant ($t = -.62$ with a p-value of .549 and $t = -.36$ with a p-value of .728). Although the overall model is statistically significant, $F = 5.37$ with probability of .022, neither predictor is significant.

Sometimes the effects of two variables are not additive because of the interacting effects between the two variables. In such a case, the researcher can use multiple regression

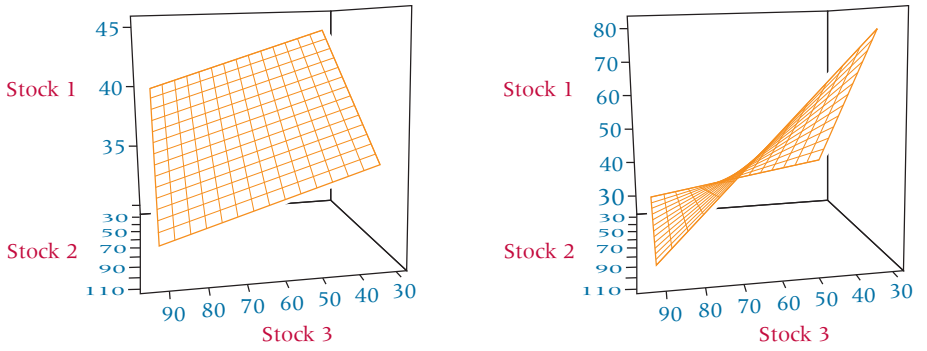
FIGURE 14.5

Two Minitab Regression Outputs—without and with Interaction

```
Regression Analysis: Stock 1 versus Stock 2, Stock 3

The regression equation is
Stock 1 = 50.9 - 0.119 Stock 2 - 0.071 Stock 3

Predictor      Coef  SE Coef       T      P
Constant     50.855    3.791   13.41  0.000
Stock 2     -0.1190   0.1931   -0.62  0.549
Stock 3     -0.0708   0.1990   -0.36  0.728

S = 4.57020   R-Sq = 47.2%    R-Sq(adj) = 38.4%

Analysis of Variance

Source           DF       SS      MS      F      P
Regression        2   224.29  112.15   5.37  0.022
Residual Error   12   250.64   20.89
Total            14   474.93

Regression Analysis: Stock 1 versus Stock 2, Stock 3, Interaction

The regression equation is
Stock 1 = 12.0 + 0.879 Stock 2 + 0.220 Stock 3 - 0.00998 Interaction

Predictor        Coef   SE Coef      T      P
Constant       12.046     9.312   1.29  0.222
Stock 2        0.8788    0.2619   3.36  0.006
Stock 3        0.2205    0.1435   1.54  0.153
Interaction -0.009985  0.002314  -4.31  0.001

S = 2.90902   R-Sq = 80.4% R-Sq(adj) = 75.1%

Analysis of Variance

Source           DF       SS      MS      F      P
Regression        3   381.85  127.28  15.04  0.000
Residual Error   11    93.09    8.46
Total            14   474.93
```

FIGURE 14.6

Response Surfaces for the Stock Example—Without and with Interaction



(a) Without Interaction **(b)** With Interaction

analysis to explore the interaction effects by including an interaction term in the equation.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$$

The equation fits the form of the general linear model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

where $x_3 = x_1 x_2$. Each individual observation of x_3 is obtained through a recoding process by multiplying the associated observations of x_1 and x_2.

Applying this procedure to the stock example, the researcher uses the interaction term and Minitab to obtain the second regression output shown in Figure 14.5. This output contains x_1, x_2, and the interaction term, $x_1 x_2$. Observe the R^2, which equals .804 for this model. The introduction of the interaction term caused the R^2 to increase from 47.2% to 80.4%. In addition, the standard error of the estimate decreased from 4.570 in the first model to 2.909 in the second model. The t ratios for both the x_1 term and the interaction term are statistically significant in the second model ($t = 3.36$ with a p-value of .006 for x_1 and $t = -4.31$ with a probability of.001 for $x_1 x_2$). The inclusion of the interaction term helped the regression model account for a substantially greater amount of the dependent variable and is a significant contributor to the model.

Figure 14.6(a) is the response surface for the first regression model presented in Figure 14.5 (the model without interaction). As you observe the response plane with stock 3 as the point of reference, you see the plane moving upward with increasing values of stock 1 as the plane moves away from you toward smaller values of stock 2. Now examine Figure 14.6(b), the response surface for the second regression model presented in Figure 14.5 (the model with interaction). Note how the response plane is twisted, with its slope changing as it moves along stock 2. This pattern is caused by the interaction effects of stock 2 prices and stock 3 prices. A cross-section of the plane taken from left to right at any given stock 2 price produces a line that attempts to predict the price of stock 3 from the price of stock 1. As you move back through different prices of stock 2, the slope of that line changes, indicating that the relationship between stock 1 and stock 3 varies according to stock 2.

A researcher also could develop a model using two independent variables with their squares and interaction. Such a model would be a second-order model with two independent variables. The model would look like this.

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

Model Transformation

To this point in examining polynomial and interaction models, the focus has been on recoding values of x variables. Some multiple regression situations require that the dependent variable, y, be recoded. To examine different relationships between x and y, Tukey's four-quadrant analysis and ladder of transformations can be used to explore ways to recode x or y in attempting to construct regression models with more predictability. Included on the ladder are such y transformations as log y and $1/y$.

Suppose the following data represent the annual sales and annual advertising expenditures for seven companies. Can a regression model be developed from these figures that can be used to predict annual sales by annual advertising expenditures?

Company	Sales ($ million/year)	Advertising ($ million/year)
1	2,580	1.2
2	11,942	2.6
3	9,845	2.2
4	27,800	3.2
5	18,926	2.9
6	4,800	1.5
7	14,550	2.7

One mathematical model that is a good candidate for fitting these data is an exponential model of the form

$$y = \beta_0 \beta_1^x \in$$

This model can be transformed (by taking the log of each side) so that it is in the form of the general linear equation.

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

This transformed model requires a recoding of the y data through the use of logarithms. Notice that x is not recoded but that the regression constant and coefficient are in logarithmic scale. If we let $y' = \log y$, $\beta_0' = \log \beta_0$, and $\beta_1' = \log \beta_1$, the exponential model is in the form of the general linear model.

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

The process begins by taking the log of the y values. The data used to build the regression model and the Excel regression output for these data follow.

Log Sales (y)	Advertising (x)
3.4116	1.2
4.0771	2.6
3.9932	2.2
4.4440	3.2
4.2771	2.9
3.6812	1.5
4.1629	2.7

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.990
R Square	0.980
Adjusted R Square	0.977
Standard Error	0.0543
Observations	7

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.739215	0.739215	250.36	0.000018
Residual	5	0.014763	0.002953		
Total	6	0.753979			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	2.9003	0.0729	39.80	0.00000019
Advertising	0.4751	0.0300	15.82	0.00001834

A simple regression model (without the log recoding of the y variable) yields an R^2 of 87%, whereas the exponential model R^2 is 98%. The t statistic for advertising is 15.82 with a p-value of 0.00001834 in the exponential model and 5.77 with a p-value of 0.00219 in the simple regression model. Thus the exponential model gives a better fit than does the simple regression model. An examination of (x^2, y) and (x^3, y) models reveals R^2 of .930 and .969, respectively, which are quite high but still not as good as the R^2 yielded by the exponential model (the output for these models is not shown here).

The resulting equation of the exponential regression model is

$$y = 2.9003 + .4751x$$

In using this regression equation to determine predicted values of y for x, remember that the resulting predicted y value is in logarithmic form and the antilog of the predicted y must be taken to get the predicted y value in raw units. For example, to get the predicted y value (sales) for an advertising figure of 2.0 ($ million), substitute $x = 2.0$ into the regression equation.

$$y = 2.9003 + .4751x = 2.9003 + .4751(2.0) = 3.8505$$

The log of sales is 3.8505. Taking the antilog of 3.8505 results in the predicted sales in raw units.

$$\text{antilog}(3.8505) = 7087.61(\$ \text{ million})$$

Thus, the exponential regression model predicts that $2.0 million of advertising will result in $7,087.61 million of sales.

Other ways can be used to transform mathematical models so that they can be treated like the general linear model. One example is an inverse model such as

$$y = \frac{1}{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon}$$

Such a model can be manipulated algebraically into the form

$$\frac{1}{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

Substituting $y' = 1/y$ into this equation results in an equation that is in the form of the general linear model.

$$y' = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

To use this "inverse" model, recode the data values for y by using $1/y$. The regression analysis is done on the $1/y$, x_1, and x_2 data. To get predicted values of y from this model, enter the raw values of x_1 and x_2. The resulting predicted value of y from the regression equation will be the inverse of the actual predicted y value.

DEMONSTRATION PROBLEM 14.1

In the aerospace and defense industry, some cost estimators predict the cost of new space projects by using mathematical models that take the form

$$y = \beta_0 x^{\beta_1} \epsilon$$

These cost estimators often use the weight of the object being sent into space as the predictor (x) and the cost of the object as the dependent variable (y). Quite often β_1 turns out to be a value between 0 and 1, resulting in the predicted value of y equaling some root of x.

Use the sample cost data given here to develop a cost regression model in the form just shown to determine the equation for the predicted value of y. Use this regression equation to predict the value of y for $x = 3{,}000$.

y (cost in billions)	*x* (weight in tons)
1.2	450
9.0	20,200
4.5	9,060
3.2	3,500
13.0	75,600
0.6	175
1.8	800
2.7	2,100

Solution

The equation

$$y = \beta_0 x^{\beta_1} \in$$

is not in the form of the general linear model, but it can be transformed by using logarithms:

$$\log y = \log \beta_0 + \beta_1 \log x + \in$$

which takes on the general linear form

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

where

$$y' = \log y$$
$$\beta_0' = \log \beta_0$$
$$x' = \log x$$

This equation requires that both x and y be recoded by taking the logarithm of each.

log *y*	log *x*
.0792	2.6532
.9542	4.3054
.6532	3.9571
.5051	3.5441
1.1139	4.8785
−.2218	2.2430
.2553	2.9031
.4314	3.3222

Using these data, the computer produces the following regression constant and coefficient:

$$b_0' = -1.25292 \quad b_1 = .49606$$

From these values, the equation of the predicted y value is determined to be

$$\log \hat{y} = -1.25292 + .49606 \log x$$

If $x = 3{,}000$, $\log x = 3.47712$, and

$$\log \hat{y} = -1.25292 + .49606(3.47712) = .47194$$

then

$$\hat{y} = \text{antilog}(\log \hat{y}) = \text{antilog}(.47194) = 2.9644$$

The predicted value of y is \$2.9644 billion for $x = 3000$ tons of weight.

Taking the antilog of $b_0' = -1.25292$ yields .055857. From this and $b_1 = .49606$, the

model can be written in the original form:

$$y = (.055857)x^{.49606}$$

Substituting $x = 3000$ into this formula also yields $2.9645 billion for the predicted value of y.

14.1 PROBLEMS

14.1 Use the following data to develop a quadratic model to predict y from x. Develop a simple regression model from the data and compare the results of the two models. Does the quadratic model seem to provide any better predictability? Why or why not?

x	y	x	y
14	200	15	247
9	74	8	82
6	29	5	21
21	456	10	94
17	320		

14.2 Develop a multiple regression model of the form

$$y = b_0 b_1^x \in$$

using the following data to predict y from x. From a scatter plot and Tukey's ladder of transformation, explore ways to recode the data and develop an alternative regression model. Compare the results.

y	x	y	x
2485	3.87	740	2.83
1790	3.22	4010	3.62
874	2.91	3629	3.52
2190	3.42	8010	3.92
3610	3.55	7047	3.86
2847	3.61	5680	3.75
1350	3.13	1740	3.19

14.3 The Publishers Information Bureau in New York City released magazine advertising expenditure data compiled by leading national advertisers. The data were organized by product type over several years. Shown here are data on total magazine advertising expenditures and household equipment and supplies advertising expenditures. Using these data, develop a regression model to predict total magazine advertising expenditures by household equipment and supplies advertising expenditures and by (household equipment and supplies advertising expenditures)^2. Compare this model to a regression model to predict total magazine advertising expenditures by only household equipment and supplies advertising expenditures. Construct a scatter plot of the data. Does the shape of the plot suggest some alternative models in light of Tukey's four-quadrant approach? If so, develop at least one other model and compare the model to the other two previously developed.

Total Magazine Advertising Expenditures ($ millions)	Household Equipment and Supplies Expenditures ($ millions)
1193	34
2846	65
4668	98
5120	93
5943	102
6644	103

14.4 Dun & Bradstreet reports, among other things, information about new business incorporations and number of business failures over several years. Shown here are data

on business failures and current liabilities of the failing companies over several years. Use these data and the following model to predict current liabilities of the failing companies by the number of business failures. Discuss the strength of the model.

$$y = b_0 b_1^x \in$$

Now develop a different regression model by recoding x. Use Tukey's four-quadrant approach as a resource. Compare your models.

Rate of Business Failures (10,000)	Current Liabilities of Failing Companies ($ millions)
44	1,888
43	4,380
42	4,635
61	6,955
88	15,611
110	16,073
107	29,269
115	36,937
120	44,724
102	34,724
98	39,126
65	44,261

14.5 Use the following data to develop a curvilinear model to predict y. Include both x_1 and x_2 in the model in addition to x_1^2 and x_2^2, and the interaction term x_1x_2. Comment on the overall strength of the model and the significance of each predictor. Develop a regression model with the same independent variables as the first model but without the interaction variable. Compare this model to the model with interaction.

y	x_1	x_2
47.8	6	7.1
29.1	1	4.2
81.8	11	10.0
54.3	5	8.0
29.7	3	5.7
64.0	9	8.8
37.4	3	7.1
44.5	4	5.4
42.1	4	6.5
31.6	2	4.9
78.4	11	9.1
71.9	9	8.5
17.4	2	4.2
28.8	1	5.8
34.7	2	5.9
57.6	6	7.8
84.2	12	10.2
63.2	8	9.4
39.0	3	5.7
47.3	5	7.0

14.6 What follows is Excel output from a regression model to predict y using x_1, x_2, x_1^2, x_2^2, and the interaction term, x_1x_2. Comment on the overall strength of the model and the significance of each predictor. The data follow the Excel output. Develop a regression model with the same independent variables as the first model but without the interaction variable. Compare this model to the model with interaction.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.954
R Square	0.910
Adjusted R Square	0.878
Standard Error	7.544
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	8089.274577	1617.855	28.43	0.00000073
Residual	14	796.725	56.909		
Total	19	8886			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	464.4433	503.0955	0.92	0.3716
X_1	–10.5101	6.0074	–1.75	0.1021
X_2	–1.2212	1.9791	–0.62	0.5471
X_1Sq	0.0357	0.0195	1.84	0.0876
X_2Sq	–0.0002	0.0021	–0.08	0.9394
X_1*X_2	0.0243	0.0107	2.28	0.0390

y	x_1	x_2	y	x_1	x_2
34	120	190	45	96	245
56	105	240	34	79	288
78	108	238	23	66	312
90	110	250	89	88	315
23	78	255	76	80	320
34	98	230	56	73	335
45	89	266	43	69	335
67	92	270	23	75	250
78	95	272	45	63	372
65	85	288	56	74	360

14.2 INDICATOR (DUMMY) VARIABLES

Some variables are referred to as **qualitative variables** (as opposed to *quantitative* variables) because qualitative variables do not yield quantifiable outcomes. Instead, *qualitative variables yield nominal- or ordinal-level information*, which is used more to categorize items. These variables have a role in multiple regression and are referred to as **indicator**, or **dummy variables**. In this section, we will examine the role of indicator, or dummy, variables as predictors or independent variables in multiple regression analysis.

Indicator variables arise in many ways in business research. Mail questionnaire or personal interview demographic questions are prime candidates because they tend to generate qualitative measures on such items as sex, geographic region, occupation, marital status, level of education, economic class, political affiliation, religion, management/nonmanagement status, buying/leasing a home, method of transportation, or type of broker. In one business study, business researchers were attempting to develop a multiple regression model to predict the distances shoppers drive to malls in the greater Cleveland area. One independent variable was whether the mall was located on the shore of Lake Erie. In a second study, a site location model for pizza restaurants included indicator variables for (1) whether the restaurant served beer and (2) whether the restaurant had a salad bar.

These indicator variables are qualitative in that no interval or ratio level measurement is assigned to a response. For example, if a mall is located on the shore of Lake Erie,

awarding it a score of 20 or 30 or 75 because of its location makes no sense. In terms of sex, what value would you assign to a man or a woman in a regression study? Yet these types of indicator, or dummy, variables are often useful in multiple regression studies and can be included if they are coded in the proper format.

Most researchers code indicator variables by using 0 or 1. For example, in the shopping mall study, malls located on the shore of Lake Erie could be assigned a 1, and all other malls would then be assigned a 0. The assignment of 0 or 1 is arbitrary, with the number merely holding a place for the category. For this reason, the coding is referred to as "dummy" coding; the number represents a category by holding a place and is not a measurement.

Many indicator, or dummy, variables are dichotomous, such as male/female, salad bar/no salad bar, employed/not employed, and rent/own. For these variables, a value of 1 is arbitrarily assigned to one category and a value of 0 is assigned to the other category. Some qualitative variables contain several categories, such as the variable "type of job," which might have the categories assembler, painter, and inspector. In this case, using a coding of 1, 2, and 3, respectively, is tempting. However, that type of coding creates problems for multiple regression analysis. For one thing, the category "inspector" would receive a value that is three times that of "painter." In addition, the values of 1, 2, and 3 indicate a hierarchy of job types: assembler $<$ painter $<$ inspector. The proper way to code such indicator variables is with the 0, 1 coding. Two separate independent variables should be used to code the three categories of type of job. The first variable is assembler, where a 1 is recorded if the person's job is assembler and a 0 is recorded if it is not. The second variable is painter, where a 1 is recorded if the person's job is painter and a 0 is recorded if it is not. A variable should not be assigned to inspector, because all workers in the study for whom a 1 was not recorded either for the assembler variable or the painter variable must be inspectors. Thus, coding the inspector variable would result in redundant information and is not necessary. This reasoning holds for all indicator variables with more than two categories. If an indicator variable has c categories, then $c - 1$ dummy variables must be created and inserted into the regression analysis in order to include the indicator variable in the multiple regression.[†]

An example of an indicator variable with more than two categories is the result of the following question taken from a typical questionnaire.

Your office is located in which region of the country?

_____ Northeast _____ Midwest _____ South _____ West

Suppose a researcher is using a multiple regression analysis to predict the cost of doing business and believes geographic location of the office is a potential predictor. How does the researcher insert this qualitative variable into the analysis? Because $c = 4$ for this question, three dummy variables are inserted into the analysis. Table 14.4 shows one possible way this process works with 13 respondents. Note that rows 2, 7, and 11 contain all zeros, which indicate that those respondents have offices in the West. Thus, a fourth dummy variable for the West region is not necessary and, indeed, should not be included because the information contained in such a fourth variable is contained in the other three variables.

TABLE 14.4

Coding for the Indicator Variable of Geographic Location for Regression Analysis

Northeast x_1	Midwest x_2	South x_3
1	0	0
0	0	0
1	0	0
0	0	1
0	1	0
0	1	0
0	0	0
0	0	1
1	0	0
1	0	0
0	0	0
0	1	0
0	0	1

A word of caution is in order. Because of degrees of freedom and interpretation considerations, it is important that a multiple regression analysis have enough observations to handle adequately the number of independent variables entered. Some researchers recommend as a rule of thumb at least three observations per independent variable. If a qualitative variable has multiple categories, resulting in several dummy independent variables, and if several qualitative variables are being included in an analysis, the number of predictors can rather quickly exceed the limit of recommended number of variables per number of observations. Nevertheless, dummy variables can be useful and are a way in which nominal or ordinal information can be recoded and incorporated into a multiple regression model.

As an example, consider the issue of sex discrimination in the salary earnings of workers in some industries. In examining this issue, suppose a random sample of 15 workers is

[†]If c indicator variables are included in the analysis, no unique estimator of the regression coefficients can be found. [J. Neter, M. H. Kuter, W. Wasserman, and C. Nachtsheim, *Applied Linear Regression Models,* 3rd ed. Chicago: Richard D. Irwin, 1996.]

TABLE 14.5
Data for the Monthly Salary Example

Monthly Salary ($1,000)	Age (10 years)	Sex (1 = male, 0 = female)
1.548	3.2	1
1.629	3.8	1
1.011	2.7	0
1.229	3.4	0
1.746	3.6	1
1.528	4.1	1
1.018	3.8	0
1.190	3.4	0
1.551	3.3	1
0.985	3.2	0
1.610	3.5	1
1.432	2.9	1
1.215	3.3	0
.990	2.8	0
1.585	3.5	1

drawn from a pool of employed laborers in a particular industry and the workers' average monthly salaries are determined, along with their age and gender. The data are shown in Table 14.5. As sex can be only male or female, this variable is a dummy variable requiring 0, 1 coding. Suppose we arbitrarily let 1 denote male and 0 denote female. Figure 14.7 is the multiple regression model developed from the data of Table 14.5 by using Minitab to predict the dependent variable, monthly salary, by two independent variables, age and sex.

The computer output in Figure 14.7 contains the regression equation for this model.

$$\text{Salary} = 0.732 + 0.111\ \text{Age} + 0.459\ \text{Sex}$$

An examination of the t ratios reveals that the dummy variable "sex" has a regression coefficient that is significant at $\alpha = .001$ ($t = 8.58$, $p = .000$). The overall model is significant at $\alpha = .001$ ($F = 48.54$, $p = .000$). The standard error of the estimate, $s_e = .09679$, indicates that approximately 68% of the errors of prediction are within ± $96.79 (.09679 · $1,000). The R^2 is relatively high at 89.0%, and the adjusted R^2 is 87.2%.

The t value for sex indicates that it is a significant predictor of monthly salary in this model. This significance is apparent when one looks at the effects of this dummy variable another way. Figure 14.8 shows the graph of the regression equation when sex = 1 (male) and the graph of the regression equation when sex = 0 (female). When sex = 1 (male), the regression equation becomes

$$.732 + .111(\text{Age}) + .459(1) = 1.191 + .111(\text{Age})$$

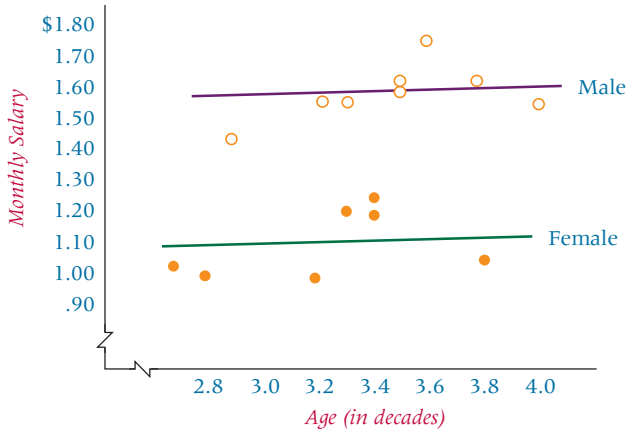
FIGURE 14.7
Minitab Regression Output for the Monthly Salary Example

Regression Analysis: Salary versus Age, Sex

```
The regression equation is
Salary = 0.732 + 0.111 Age + 0.459  Sex

Predictor     Coef  SE Coef     T      P
Constant    0.7321   0.2356  3.11  0.009
Age        0.11122  0.07208  1.54  0.149
Sex        0.45868  0.05346  8.58  0.000

S = 0.0967916  R-Sq = 89.0%    R-Sq(adj) = 87.2%

Analysis of Variance

Source           DF       SS       MS      F      P
Regression        2  0.90949  0.45474  48.54  0.000
Residual Error   12  0.11242  0.00937
Total            14  1.02191
```

FIGURE 14.8

Regression Model for Male and Female Sex



When sex = 0 (female), the regression equation becomes

$$.732 + .111(\text{Age}) + .459(0) = .732 + .111(\text{Age}).$$

The full regression model (with both predictors) has a response surface that is a plane in a three-dimensional space. However, if a value of 1 is entered for sex into the full regression model, as just shown, the regression model is reduced to a line passing through the plane formed by monthly salary and age. If a value of 0 is entered for sex, as shown, the full regression model also reduces to a line passing through the plane formed by monthly salary and age. Figure 14.8 displays these two lines. Notice that the only difference in the two lines is the y-intercept. Observe the monthly salary with male sex, as depicted by ○, versus the monthly salary with female sex, depicted by ●. The difference in the y-intercepts of these two lines is .459, which is the value of the regression coefficient for sex. This intercept figure signifies that, on average, men earn $459 per month more than women for this population.

STATISTICS IN BUSINESS TODAY

Predicting Export Intensity of Chinese Manufacturing Firms Using Multiple Regression Analysis

According to business researchers Hongxin Zhao and Shaoming Zou, little research has been done on the impact of external or uncontrollable variables on the export performance of a company. These two researchers conducted a study of Chinese manufacturing firms and used multiple regression to determine whether both domestic market concentration and firm location are good predictors of a firm's export intensity. The study included 999 Chinese manufacturing firms that exported. The dependent variable was "export intensity," which was defined to be the proportion of production output that is exported and was computed by dividing the firm's export value by its production output value. The higher the proportion was, the higher the export intensity. Zhao and Zou used covariate techniques (beyond the scope of this text) to control for the fact that companies in the study varied by size, capital intensity, innovativeness, and industry. The independent variables were industry concentration and location. Industry concentration was computed as a ratio, with higher values indicating more concentration in the industry. The location variable was a composite index taking into account total freight volume, available modes of transportation, number of telephones, and size of geographic area.

The multiple regression model produced an R^2 of approximately 52%. Industry concentration was a statistically significant predictor at $\alpha = .01$, and the sign on the regression coefficient indicated that a negative relationship may exist between industry concentration and export intensity. It means export intensity is lower in highly concentrated industries and higher in lower concentrated industries. The researchers believe that in a more highly concentrated industry, the handful of firms dominating the industry will stifle the export competitiveness of firms. In the absence of dominating firms in a more fragmented setting, more competition and an increasing tendency to export are noted. The location variable was also a significant predictor at $\alpha = .01$. Firms located in coastal areas had higher export intensities than did those located in inland areas.

Source: Hongxin Zhao and Shaoming Zou, "The Impact of Industry Concentration and Firm Location on Export Propensity and Intensity: An Empirical Analysis of Chinese Manufacturing Firms," *Journal of International Marketing*, vol. 10, no. 1 (2002), pp. 52–71.

14.2 PROBLEMS

14.7 Analyze the following data by using a multiple regression computer software package to predict y using x_1 and x_2. Notice that x_2 is a dummy variable. Discuss the output from the regression analysis; in particular, comment on the predictability of the dummy variable.

y	x_1	x_2
16.8	27	1
13.2	16	0
14.7	13	0
15.4	11	1
11.1	17	0
16.2	19	1
14.9	24	1
13.3	21	0
17.8	16	1
17.1	23	1
14.3	18	0
13.9	16	0

14.8 Given here are the data from a dependent variable and two independent variables. The second independent variable is an indicator variable with several categories. Hence, this variable is represented by x_2, x_3, and x_4. How many categories are needed in total for this independent variable? Use a computer to perform a multiple regression analysis on this data to predict y from the x values. Discuss the output and pay particular attention to the dummy variables.

y	x_1	x_2	x_3	x_4
11	1.9	1	0	0
3	1.6	0	1	0
2	2.3	0	1	0
5	2.0	0	0	1
9	1.8	0	0	0
14	1.9	1	0	0
10	2.4	1	0	0
8	2.6	0	0	0
4	2.0	0	1	0
9	1.4	0	0	0
11	1.7	1	0	0
4	2.5	0	0	1
6	1.0	1	0	0
10	1.4	0	0	0
3	1.9	0	1	0
4	2.3	0	1	0
9	2.2	0	0	0
6	1.7	0	0	1

14.9 The Minitab output displayed here is the result of a multiple regression analysis with three independent variables. Variable x_1 is a dummy variable. Discuss the computer output and the role x_1 plays in this regression model.

The regression equation is
$Y = 121 + 13.4\ X_1 - 0.632\ X_2 + 1.42\ X_3$

Predictor	Coef	Stdev	T	p
Constant	121.31	11.56	10.50	.000
X_1	13.355	4.714	2.83	.014
X_2	−0.6322	0.2270	−2.79	.015
X_3	1.421	3.342	0.43	.678

S = 7.041 R-sq = 79.5% R-sq(adj) = 74.7%

Analysis of Variance

Source	df	SS	MS	F	p
Regression	3	2491.98	830.66	16.76	.000
Error	13	644.49	49.58		
Total	16	3136.47			

14.10 Given here is Excel output for a multiple regression model that was developed to predict y from two independent variables, x_1 and x_2. Variable x_2 is a dummy variable. Discuss the strength of the multiple regression model on the basis of the output. Focus on the contribution of the dummy variable. Plot x_1 and y with x_2 as 0, and then plot x_1 and y with x_2 as 1. Compare the two lines and discuss the differences.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.623
R Square	0.388
Adjusted R Square	0.341
Standard Error	11.744
Observations	29

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	2270.11	1135.05	8.23	0.0017
Residual	26	3585.75	137.91		
Total	28	5855.86			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	41.225	6.3800	6.46	0.00000076
X_1	1.081	1.3530	0.80	0.4316
X_2	–18.404	4.5470	–4.05	0.0004

14.11 Falvey, Fried, and Richards developed a multiple regression model to predict the average price of a meal at New Orleans restaurants. The variables explored included such indicator variables as the following: Accepts reservations, Accepts credit cards, Has its own parking lot, Has a separate bar or lounge, Has a maitre d', Has a dress code, Is candlelit, Has live entertainment, Serves alcoholic beverages, Is a steakhouse, Is in the French Quarter. Suppose a relatively simple model is developed to predict the average price of a meal at a restaurant in New Orleans from the number of hours the restaurant is open per week, the probability of being seated upon arrival, and whether the restaurant is located in the French Quarter. Use the following data and a computer to develop such a model. Comment on the output.

Price	Hours	Probability of Being Seated	French Quarter
$ 8.52	65	.62	0
21.45	45	.43	1
16.18	52	.58	1
6.21	66	.74	0
12.19	53	.19	1
25.62	55	.49	1
13.90	60	.80	0
18.66	72	.75	1
5.25	70	.37	0
7.98	55	.64	0
12.57	48	.51	1
14.85	60	.32	1
8.80	52	.62	0
6.27	64	.83	0

14.12 A researcher gathered 155 observations on four variables: job satisfaction, occupation, industry, and marital status. She wants to develop a multiple regression model to predict job satisfaction by the other three variables. All three predictor variables are qualitative variables with the following categories.

1. Occupation: accounting, management, marketing, finance
2. Industry: manufacturing, healthcare, transportation
3. Marital status: married, single

How many variables will be in the regression model? Delineate the number of predictors needed in each category and discuss the total number of predictors.

14.3 MODEL-BUILDING: SEARCH PROCEDURES

To this point in the chapter, we have explored various types of multiple regression models. We evaluated the strengths of regression models and learned how to understand more about the output from multiple regression computer packages. In this section we examine procedures for developing several multiple regression model options to aid in the decision-making process.

Suppose a researcher wants to develop a multiple regression model to predict the world production of crude oil. The researcher realizes that much of the world crude oil market is driven by variables related to usage and production in the United States. The researcher decides to use as predictors the following five independent variables.

1. U.S. energy consumption
2. Gross U.S. nuclear electricity generation
3. U.S. coal production
4. Total U.S. dry gas (natural gas) production
5. Fuel rate of U.S.-owned automobiles

The researcher measured data for each of these variables for the year preceding each data point of world crude oil production, figuring that the world production is driven by the previous year's activities in the United States. It would seem that as the energy consumption of the United States increases, so would world production of crude oil. In addition, it makes sense that as nuclear electricity generation, coal production, dry gas production, and fuel rates increase, world crude oil production would decrease if energy consumption stays approximately constant.

Table 14.6 shows data for the five independent variables along with the dependent variable, world crude oil production. Using the data presented in Table 14.6, the researcher attempted to develop a multiple regression model using five different independent variables. The result of this process was the Minitab output in Figure 14.9. Examining the output, the researcher can reach some conclusions about that particular model and its variables.

The output contains an R^2 value of 92.1%, a standard error of the estimate of 1.215, and an overall significant F value of 46.62. Notice from Figure 14.9 that the t ratios indicate that the regression coefficients of four of the predictor variables, nuclear, coal, dry gas, and fuel rate, are not significant at $\alpha = .05$. If the researcher were to drop these four variables out of the regression analysis and rerun the model with the other predictor only, what would happen to the model? What if the researcher ran a regression model with only three predictors? How would these models compare to the full model with all five predictors? Are all the predictors necessary?

Developing regression models for business decision making involves at least two considerations. The first is to develop a regression model that accounts for the most variation of the dependent variable—that is, develop models that maximize the explained proportion of the deviation of the y values. At the same time, the regression model should be as parsimonious (simple and economical) as possible. The more complicated a quantitative model becomes, the harder it is for managers to understand and implement the model. In

TABLE 14.6
Data for Multiple Regression Model to Predict Crude Oil Production

World Crude Oil Production (million barrels per day)	U.S. Energy Consumption (quadrillion BTUs generation per year)	U.S. Nuclear Electricity (billion kilowatt-hours)	U.S. Coal Gross Production (million short-tons)	U.S. Total Dry Gas Production (trillion cubic feet)	U.S. Fuel Rate for Automobiles (miles per gallon)
55.7	74.3	83.5	598.6	21.7	13.4
55.7	72.5	114.0	610.0	20.7	13.6
52.8	70.5	172.5	654.6	19.2	14.0
57.3	74.4	191.1	684.9	19.1	13.8
59.7	76.3	250.9	697.2	19.2	14.1
60.2	78.1	276.4	670.2	19.1	14.3
62.7	78.9	255.2	781.1	19.7	14.6
59.6	76.0	251.1	829.7	19.4	16.0
56.1	74.0	272.7	823.8	19.2	16.5
53.5	70.8	282.8	838.1	17.8	16.9
53.3	70.5	293.7	782.1	16.1	17.1
54.5	74.1	327.6	895.9	17.5	17.4
54.0	74.0	383.7	883.6	16.5	17.5
56.2	74.3	414.0	890.3	16.1	17.4
56.7	76.9	455.3	918.8	16.6	18.0
58.7	80.2	527.0	950.3	17.1	18.8
59.9	81.4	529.4	980.7	17.3	19.0
60.6	81.3	576.9	1029.1	17.8	20.3
60.2	81.1	612.6	996.0	17.7	21.2
60.2	82.2	618.8	997.5	17.8	21.0
60.2	83.9	610.3	945.4	18.1	20.6
61.0	85.6	640.4	1033.5	18.8	20.8
62.3	87.2	673.4	1033.0	18.6	21.1
64.1	90.0	674.7	1063.9	18.8	21.2
66.3	90.6	628.6	1089.9	18.9	21.5
67.0	89.7	666.8	1109.8	18.9	21.6

addition, as more variables are included in a model, it becomes more expensive to gather historical data or update present data for the model. These two considerations (dependent variable explanation and parsimony of the model) are quite often in opposition to each other. Hence the business researcher, as the model builder, often needs to explore many model options.

FIGURE 14.9
Minitab Output of Regression for Crude Oil Production Example

```
Regression Analysis: CrOilPrd Versus USEnCons, USNucGen, ...

The regression equation is
CrOilPrd = 2.71 + 0.836 USEnCons - 0.00654 USNucGen + 0.00983
           USCoalPr - 0.143 USDryGas - 0.734 FuelRate

Predictor       Coef   SE Coef      T      P
Constant       2.708     8.909   0.30  0.764
USEnCons      0.8357    0.1802   4.64  0.000
USNucGen   -0.006544  0.009854  -0.66  0.514
USCoalPr    0.009825  0.007286   1.35  0.193
USDryGas     -0.1432    0.4484  -0.32  0.753
FuelRate     -0.7341    0.5488  -1.34  0.196

S = 1.21470   R-Sq = 92.1%   R-Sq(adj) = 90.1%

Analysis of Variance
Source           DF       SS      MS      F      P
Regression        5  343.916  68.783  46.62  0.000
Residual Error   20   29.510   1.476
Total            25  373.427
```

TABLE 14.7
Predictors for All Possible Regressions with Five Independent Variables

Single Predictor	Two Predictors	Three Predictors	Four Predictors	Five Predictors
x_1	x_1, x_2	x_1, x_2, x_3	x_1, x_2, x_3, x_4	x_1, x_2, x_3, x_4, x_5
x_2	x_1, x_3	x_1, x_2, x_4	x_1, x_2, x_3, x_5	
x_3	x_1, x_4	x_1, x_2, x_5	x_1, x_2, x_4, x_5	
x_4	x_1, x_5	x_1, x_3, x_4	x_1, x_3, x_4, x_5	
x_5	x_2, x_3	x_1, x_3, x_5	x_2, x_3, x_4, x_5	
	x_2, x_4	x_1, x_4, x_5		
	x_2, x_5	x_2, x_3, x_4		
	x_3, x_4	x_2, x_3, x_5		
	x_3, x_5	x_2, x_4, x_5		
	x_4, x_5	x_3, x_4, x_5		

In the world crude oil production regression model, if three variables explain the deviation of world crude oil production nearly as well as five variables, the simpler model is more attractive. How might researchers conduct regression analysis so that they can examine several models and then choose the most attractive one? The answer is to use search procedures.

Search Procedures

Search procedures are processes *whereby more than one multiple regression model is developed for a given database, and the models are compared and sorted by different criteria,* depending on the given procedure. Virtually all search procedures are done on a computer. Several search procedures are discussed in this section, including all possible regressions, stepwise regression, forward selection, and backward elimination.

All Possible Regressions

The all **possible regressions** search procedure *computes all possible linear multiple regression models from the data using all variables.* If a data set contains k independent variables, all possible regressions will determine $2^k - 1$ different models.

For the crude oil production example, the procedure of all possible regressions would produce $2^5 - 1 = 31$ different models from the $k = 5$ independent variables. With $k = 5$ predictors, the procedure produces all single-predictor models, all models with two predictors, all models with three predictors, all models with four predictors, and all models with five predictors, as shown in Table 14.7.

The all possible regressions procedure enables the business researcher to examine every model. In theory, this method eliminates the chance that the business researcher will never consider some models, as can be the case with other search procedures. On the other hand, the search through all possible models can be tedious, time-consuming, inefficient, and perhaps overwhelming.

Stepwise Regression

Perhaps the most widely known and used of the search procedures is stepwise regression. **Stepwise regression** is *a step-by-step process that begins by developing a regression model with a single predictor variable and adds and deletes predictors one step at a time,* examining the fit of the model at each step until no more significant predictors remain outside the model.

STEP 1. In step 1 of a stepwise regression procedure, the k independent variables are examined one at a time by developing a simple regression model for each independent variable to predict the dependent variable. The model containing the largest absolute value of t for an independent variable is selected, and the

independent variable associated with the model is selected as the "best" single predictor of y at the first step. Some computer software packages use an F value instead of a t value to make this determination. Most of these computer programs allow the researcher to predetermine critical values for t or F, but also contain a default value as an option. If the first independent variable selected at step 1 is denoted x_1, the model appears in the form

$$\hat{y} = b_0 + b_1 x_1$$

If, after examining all possible single-predictor models, it is concluded that none of the independent variables produces a t value that is significant at α, then the search procedure stops at step 1 and recommends no model.

STEP 2. In step 2, the stepwise procedure examines all possible two-predictor regression models with x_1 as one of the independent variables in the model and determines which of the other $k-1$ independent variables in conjunction with x_1 produces the highest absolute t value in the model. If this other variable selected from the remaining independent variables is denoted x_2 and is included in the model selected at step 2 along with x_1, the model appears in the form

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

At this point, stepwise regression pauses and examines the t value of the regression coefficient for x_1. Occasionally, the regression coefficient for x_1 will become statistically nonsignificant when x_2 is entered into the model. In that case, stepwise regression will drop x_1 out of the model and go back and examine which of the other $k-2$ independent variables, if any, will produce the largest significant absolute t value when that variable is included in the model along with x_2. If no other variables show significant t values, the procedure halts. It is worth noting that the regression coefficients are likely to change from step to step to account for the new predictor being added in the process. Thus, if x_1 stays in the model at step 2, the value of b_1 at step 1 will probably be different from the value of b_1 at step 2.

STEP 3. Step 3 begins with independent variables, x_1 and x_2 (the variables that were finally selected at step 2), in the model. At this step, a search is made to determine which of the $k-2$ remaining independent variables in conjunction with x_1 and x_2 produces the largest significant absolute t value in the regression model. Let us denote the one that is selected as x_3. If no significant t values are acknowledged at this step, the process stops here and the model determined in step 2 is the final model. At step 3, the model appears in the form

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$$

In a manner similar to step 2, stepwise regression now goes back and examines the t values of the regression coefficients of x_1 and x_2 in this step 3 model. If either or both of the t values are now nonsignificant, the variables are dropped out of the model and the process calls for a search through the remaining $k-3$ independent variables to determine which, if any, in conjunction with x_3 produce the largest significant t values in this model. The stepwise regression process continues step by step until no significant independent variables remain that are not in the model.

In the crude oil production example, recall that Table 14.6 contained data that can be used to develop a regression model to predict world crude oil production from as many as five different independent variables. Figure 14.9 displayed the results of a multiple regression analysis to produce a model using all five predictors. Suppose the researcher were to use a stepwise regression search procedure on these data to find a regression model. Recall that the following independent variables were being considered.

1. U.S. energy consumption
2. U.S. nuclear generation
3. U.S. coal production

TABLE 14.8
Step 1: Results of Simple Regression Using Each Independent Variable to Predict Oil Production

Dependent Variable	Independent Variable	t Ratio	R^2
Oil production	Energy consumption	11.77	85.2%
Oil production	Nuclear	4.43	45.0
Oil production	Coal	3.91	38.9
Oil production	Dry gas	1.08	4.6
Oil production	Fuel rate	3.54	34.2

→ Variable selected to serve as x_1

4. U.S. dry gas production
5. U.S. fuel rate

STEP 1. Each of the independent variables is examined one at a time to determine the strength of each predictor in a simple regression model. The results are reported in Table 14.8.

Note that the independent variable "energy consumption" was selected as the predictor variable, x_1, in step 1. An examination of Table 14.8 reveals that energy consumption produced the largest absolute t value (11.77) of the single predictors. By itself, energy consumption accounted for 85.2% of the variation of the y values (world crude oil production). The regression equation taken from the computer output for this model is

$$y = 13.075 + .580x_1$$

where

y = world crude oil production
x_1 = U.S. energy consumption

STEP 2. In step 2, x_1 was retained initially in the model and a search was conducted among the four remaining independent variables to determine which of those variables in conjunction with x_1 produced the largest significant t value. Table 14.9 reports the results of this search.

The information in Table 14.9 shows that the model selected in step 2 includes the independent variables "energy consumption" and "fuel rate." Fuel rate has the largest absolute t value (−3.75), and it is significant at $\alpha = .05$. Other variables produced varying sizes of t values. The model produced at step 2 has an R^2 of 90.8%. These two variables taken together account for almost 91% of the variation of world crude oil production in this sample.

From other computer information, it is ascertained that the t value for the x_1 variable in this model is 11.91, which is even higher than in step 1. Therefore, x_1 will not be dropped from the model by the stepwise regression procedure. The step 2 regression model from the computer output is

$$y = 7.14 + 0.772x_1 - 0.517x_2$$

TABLE 14.9
Step 2: Regression Results with Two Predictors

Dependent Variable y	Independent Variable x_1	Independent Variable x_2	t Ratio of x_2	R^2
Oil production	Energy consumption	Nuclear	−3.60	90.6%
Oil production	Energy consumption	Coal	−2.44	88.3
Oil production	Energy consumption	Dry gas	2.23	87.9
Oil production	Energy consumption	Fuel rate	−3.75	90.8

→ Variables selected at step 2

TABLE 14.10
Step 3: Regression Results with Three Predictors

Dependent Variable y	Independent Variable x_1	Independent Variable x_2	Independent Variable x_3	t Ratio of x_3	R^2
Oil Production	Energy consumption	Fuel rate	Nuclear	−0.43	90.9%
Oil Production	Energy consumption	Fuel rate	Coal	1.71	91.9
Oil Production	Energy consumption	Fuel rate	Dry gas	−0.46	90.9
No t ratio is significant at $\alpha = .05$. No new variables are added to the model.					

where
y = world crude oil production
x_1 = U.S. energy consumption
x_2 = U.S. fuel rate

Note that the regression coefficient for x_1 changed from .580 at step 1 in the model to .772 at step 2.

The R^2 for the model in step 1 was 85.2%. Notice that none of the R^2 values produced from step 2 models is less than 85.2%. The reason is that x_1 is still in the model, so the R^2 at this step must be at least as high as it was in step 1, when only x_1 was in the model. In addition, by examining the R^2 values in Table 14.9, you can get a feel for how much the prospective new predictor adds to the model by seeing how much R^2 increases from 85.2%. For example, with x_2 (fuel rate) added to the model, the R^2 goes up to 90.8%. However, adding the variable "dry gas" to x_1 increases R^2 very little (it goes up 87.9%).

STEP 3. In step 3, the search procedure continues to look for an additional predictor variable from the three independent variables remaining out of the solution. Variables x_1 and x_2 are retained in the model. Table 14.10 reports the result of this search.

In this step, regression models are explored that contain x_1 (energy consumption) and x_2 (fuel rate) in addition to one of the three remaining variables. None of the three models produces t ratios that are significant at $\alpha = .05$. No new variables are added to the model produced in step 2. The stepwise regression process ends.

Figure 14.10 shows the Minitab stepwise regression output for the world crude oil production example. The results printed in the table are virtually identical to the step-by-step results discussed in this section but are in a different format.

FIGURE 14.10
Minitab Stepwise Regression Output for the Crude Oil Production Example

```
Stepwise Regression: CrOilPrd versus USEnCons, USNucGen, ...
Alpha-to-Enter: 0.1 Alpha-to-Remove: 0.1
Response is CrOilPrd on 5 predictors, with N = 26
Step            1       2
Constant   13.075   7.140

USEnCons    0.580   0.772
T-Value     11.77   11.91
P-Value     0.000   0.000

FuelRate            -0.52
T-Value             -3.75
P-Value             0.001

S            1.52    1.22
R-Sq        85.24   90.83
R-Sq(adj)   84.62   90.03
```

Each column in Figure 14.10 contains information about the regression model at each step. Thus, column 1 contains data on the regression model for step 1. In each column at each step you can see the variables in the model. As an example, at step 2, energy consumption and fuel rate are in the model. The numbers above the t ratios are the regression coefficients. The coefficients and the constant in column 2, for example, yield the regression model equation values for step 2.

$$\hat{y} = 7.140 + 0.772x_1 - 0.52x_2$$

The values of R^2 (R-Sq) and the standard error of the estimate (S) are displayed on the bottom row of the output along with the adjusted value of R^2.

Forward Selection

Another search procedure is forward selection. **Forward selection** is essentially the same as stepwise regression, but once a variable is entered into the process, it is never dropped out. Forward selection begins by finding the independent variable that will produce the largest absolute value of t (and largest R^2) in predicting y. The selected variable is denoted here as x_1 and is part of the model

$$\hat{y} = b_0 + b_1x_1$$

Forward selection proceeds to step 2. While retaining x_1, it examines the other $k - 1$ independent variables and determines which variable in the model with x_1 produces the highest absolute value of t that is significant. To this point, forward selection is the same as stepwise regression. If this second variable is designated x_2, the model is

$$\hat{y} = b_0 + b_1x_1 + b_2x_2$$

At this point, forward selection does not reexamine the t value of x_1. Both x_1 and x_2 remain in the model as other variables are examined and included. When independent variables are correlated in forward selection, the overlapping of information can limit the potential predictability of two or more variables in combination. Stepwise regression takes this into account, in part, when it goes back to reexamine the t values of predictors already in the model to determine whether they are still significant predictors of y given the variables that have now entered the process. In other words, stepwise regression acknowledges that the strongest single predictor of y that is selected at step 1 may not be a significant predictor of y when taken in conjunction with other variables.

Using a forward selection procedure to develop multiple regression models for the world crude oil production example would result in the same outcome as that provided by stepwise regression because neither x_1 nor x_2 were removed from the model in that particular stepwise regression. The difference in the two procedures is more apparent in examples where variables selected at earlier steps in the process are removed during later steps in stepwise regression.

Backward Elimination

The **backward elimination** search procedure is *a step-by-step process that begins with the "full" model (all k predictors)*. Using the t values, a search is made to determine whether any nonsignificant independent variables are in the model. If no nonsignificant predictors are found, the backward process ends with the full model. If nonsignificant predictors are found, the predictor with the smallest absolute value of t is eliminated and a new model is developed with $k - 1$ independent variables.

This model is then examined to determine whether it contains any independent variables with nonsignificant t values. If it does, the predictor with the smallest absolute t value is eliminated from the process and a new model is developed for the next step.

This procedure of identifying the smallest nonsignificant t value and eliminating that variable continues until all variables left in the model have significant t values. Sometimes this process yields results similar to those obtained from forward selection and other times it does not. A word of caution is in order. Backward elimination always begins with all possible predictors in the model. Sometimes the sample data do not provide enough observations to

TABLE 14.11

Step 1: Backward Elimination, Full Model

Predictor	Coefficient	*t* Ratio	*p*
Energy consumption	.8357	4.64	.000
Nuclear	−.00654	−0.66	.514
Coal	.00983	1.35	.193
Dry gas	−.1432	−0.32	.753
Fuel rate	−.7341	−1.34	.196

Variable to be dropped from the model

TABLE 14.12

Step 2: Backward Elimination, Four Predictors

Predictor	Coefficient	*t* Ratio	*p*
Energy consumption	.7843	9.85	.000
Nuclear	−.004261	−0.64	.528
Coal	.010933	1.74	.096
Fuel rate	−.8253	−1.80	.086

Variable to be dropped from the model

justify the use of all possible predictors at the same time in the model. In this case, backward elimination is not a suitable option with which to build regression models.

The following steps show how the backward elimination process can be used to develop multiple regression models to predict world crude oil production using the data and five predictors displayed in Table 14.6.

STEP 1. A full model is developed with all predictors. The results are shown in Table 14.11. The R^2 for this model is 92.1%. A study of Table 14.11 reveals that the predictor "dry gas" has the smallest absolute value of a nonsignificant t ($t = -.32$, $p = .753$). In step 2, this variable will be dropped from the model.

STEP 2. A second regression model is developed with $k - 1 = 4$ predictors. Dry gas has been eliminated from consideration. The results of this multiple regression analysis are presented in Table 14.12. The computer results in Table 14.12 indicate that the variable "nuclear" has the smallest absolute value of a nonsignificant t of the variables remaining in the model ($t = -.64$, $p = .528$). In step 3, this variable will be dropped from the model.

STEP 3. A third regression model is developed with $k - 2 = 3$ predictors. Both nuclear and dry gas variables have been removed from the model. The results of this multiple regression analysis are reported in Table 14.13. The computer results in Table 14.13 indicate that the variable "coal" has the smallest absolute value of a nonsignificant t of the variables remaining in the model ($t = 1.71$, $p = .102$). In step 4, this variable will be dropped from the model.

STEP 4. A fourth regression model is developed with $k - 3 = 2$ predictors. Nuclear, dry gas, and coal variables have been removed from the model. The results of this multiple regression analysis are reported in Table 14.14. Observe that all p-values are less than $\alpha = .05$, indicating that all t values are significant, so no additional independent variables need to be removed. The backward elimination process ends with two predictors in the model. The final model obtained from this backward elimination process is the same model as that obtained by using stepwise regression.

TABLE 14.13

Step 3: Backward Elimination, Three Predictors

Predictor	Coefficient	*t* Ratio	*p*
Energy consumption	.75394	11.94	.000
Coal	.010479	1.71	.102
Fuel rate	−1.0283	−3.14	.005

Variable to be dropped from the model

TABLE 14.14

Step 4: Backward Elimination, Two Predictors

Predictor	Coefficient	*t* Ratio	*p*
Energy consumption	.77201	11.91	.000
Fuel rate	−.5173	−3.75	.001

All variables are significant at $\alpha = .05$.
No variables will be dropped from this model.
The process stops.

14.3 PROBLEMS

14.13 Use a stepwise regression procedure and the following data to develop a multiple regression model to predict y. Discuss the variables that enter at each step, commenting on their t values and on the value of R^2.

y	x_1	x_2	x_3	y	x_1	x_2	x_3
21	5	108	57	22	13	105	51
17	11	135	34	20	10	111	43
14	14	113	21	16	20	140	20
13	9	160	25	13	19	150	14
19	16	122	43	18	14	126	29
15	18	142	40	12	21	175	22
24	7	93	52	23	6	98	38
17	9	128	38	18	15	129	40

14.14 Given here are data for a dependent variable and four potential predictors. Use these data and a stepwise regression procedure to develop a multiple regression model to predict y. Examine the values of t and R^2 at each step and comment on those values. How many steps did the procedure use? Why do you think the process stopped?

y	x_1	x_2	x_3	x_4
101	2	77	1.2	42
127	4	72	1.7	26
98	9	69	2.4	47
79	5	53	2.6	65
118	3	88	2.9	37
114	1	53	2.7	28
110	3	82	2.8	29
94	2	61	2.6	22
96	8	60	2.4	48
73	6	64	2.1	42
108	2	76	1.8	34
124	5	74	2.2	11
82	6	50	1.5	61
89	9	57	1.6	53
76	1	72	2.0	72
109	3	74	2.8	36
123	2	99	2.6	17
125	6	81	2.5	48

14.15 The computer output given here is the result of a stepwise multiple regression analysis to predict a dependent variable by using six predictor variables. The number of observations was 108. Study the output and discuss the results. How many predictors ended up in the model? Which predictors, if any, did not enter the model?

STEPWISE REGRESSION OF Y ON 6 PREDICTORS, WITH N = 108

STEP	1	2	3	4
CONSTANT	8.71	6.82	6.57	5.96
X_3	-2.85	-4.92	-4.97	-5.00
T-RATIO	2.11	2.94	3.04	3.07
X_1		4.42	3.72	3.22
T-RATIO		2.64	2.20	2.05
X_2			1.91	1.78
T-RATIO			2.07	2.02
X_6				1.56
T-RATIO				1.98
S	3.81	3.51	3.43	3.36
R-SQ	29.20	49.45	54.72	59.29

14.16 Study the output given here from a stepwise multiple regression analysis to predict y from four variables. Comment on the output at each step.

```
STEPWISE REGRESSION OF Y ON 4 PREDICTORS,WITH N = 63
STEP             1        2
CONSTANT     27.88    22.30
X3            0.89
T-RATIO       2.26
X2                    12.38
T-RATIO                2.64
X4                   0.0047
T-RATIO                2.01
S            16.52     9.47
R-SQ         42.39    68.20
```

14.17 The National Underwriter Company in Cincinnati, Ohio, publishes property and casualty insurance data. Given here is a portion of the data published. These data include information from the U.S. insurance industry about (1) net income after taxes, (2) dividends to policyholders, (3) net underwriting gain/loss, and (4) premiums earned. Use the data and stepwise regression to predict premiums earned from the other three variables.

Premiums Earned	Net Income	Dividends	Underwriting Gain/Loss
30.2	1.6	.6	.1
47.2	.6	.7	−3.6
92.8	8.4	1.8	−1.5
95.4	7.6	2.0	−4.9
100.4	6.3	2.2	−8.1
104.9	6.3	2.4	−10.8
113.2	2.2	2.3	−18.2
130.3	3.0	2.4	−21.4
161.9	13.5	2.3	−12.8
182.5	14.9	2.9	−5.9
193.3	11.7	2.9	−7.6

14.18 The U.S. Energy Information Administration releases figures in their publication, *Monthly Energy Review*, about the cost of various fuels and electricity. Shown here are the figures for four different items over a 12-year period. Use the data and stepwise regression to predict the cost of residential electricity from the cost of residential natural gas, residual fuel oil, and leaded regular gasoline. Examine the data and discuss the output.

Residential Electricity (kWh)	Residential Natural Gas (1000 ft^3)	Residual Fuel Oil (gal)	Leaded Regular Gasoline (gal)
2.54	1.29	.21	.39
3.51	1.71	.31	.57
4.64	2.98	.44	.86
5.36	3.68	.61	1.19
6.20	4.29	.76	1.31
6.86	5.17	.68	1.22
7.18	6.06	.65	1.16
7.54	6.12	.69	1.13
7.79	6.12	.61	1.12
7.41	5.83	.34	.86
7.41	5.54	.42	.90
7.49	4.49	.33	.90

14.4 MULTICOLLINEARITY

One problem that can arise in multiple regression analysis is multicollinearity. **Multicollinearity** is *when two or more of the independent variables of a multiple regression model are highly correlated.* Technically, if two of the independent variables are correlated, we have collinearity; when three or more independent variables are correlated, we have multicollinearity. However, the two terms are frequently used interchangeably.

The reality of business research is that most of the time some correlation between predictors (independent variables) will be present. The problem of multicollinearity arises when the intercorrelation between predictor variables is high. This relationship causes several other problems, particularly in the interpretation of the analysis.

1. It is difficult, if not impossible, to interpret the estimates of the regression coefficients.
2. Inordinately small t values for the regression coefficients may result.
3. The standard deviations of regression coefficients are overestimated.
4. The algebraic sign of estimated regression coefficients may be the opposite of what would be expected for a particular predictor variable.

The problem of multicollinearity can arise in regression analysis in a variety of business research situations. For example, suppose a model is being developed to predict salaries in a given industry. Independent variables such as years of education, age, years in management, experience on the job, and years of tenure with the firm might be considered as predictors. It is obvious that several of these variables are correlated (virtually all of these variables have something to do with number of years, or time) and yield redundant information. Suppose a financial regression model is being developed to predict bond market rates by such independent variables as Dow Jones average, prime interest rates, GNP, producer price index, and consumer price index. Several of these predictors are likely to be intercorrelated.

In the world crude oil production example used in section 14.3, several of the independent variables are intercorrelated, leading to the potential of multicollinearity problems. Table 14.15 gives the correlations of the predictor variables for this example. Note that r values are quite high ($r > .90$) for fuel rate and nuclear (.972), fuel rate and coal (.968), and coal and nuclear (.952).

Table 14.15 shows that fuel rate and coal production are highly correlated. Using fuel rate as a single predictor of crude oil production produces the following simple regression model.

$$\hat{y} = 44.869 + .7838(\text{fuel rate})$$

Notice that the estimate of the regression coefficient, .7838, is positive, indicating that as fuel rate increases, oil production increases. Using coal as a single predictor of crude oil production yields the following simple regression model.

$$\hat{y} = 45.072 + .0157(\text{coal})$$

TABLE 14.15
Correlations Among Oil Production Predictor Variables

	Energy Consumption	Nuclear	Coal	Dry Gas	Fuel Rate
Energy consumption	1	.856	.791	.057	.791
Nuclear	.856	1	.952	−.404	.972
Coal	.791	.952	1	−.448	.968
Dry gas	.057	−.404	−.448	1	−.423
Fuel rate	.796	.972	.968	−.423	1

The multiple regression model developed using both fuel rate and coal to predict crude oil production is

$$\hat{y} = 45.806 + .0227(\text{coal}) - .3934(\text{fuel rate})$$

Observe that this regression model indicates a *negative* relationship between fuel rate and oil production (−.3934), which is in opposition to the *positive* relationship shown in the regression equation for fuel rate as a single predictor. Because of the multicollinearity between coal and fuel rate, these two independent variables interact in the regression analysis in such a way as to produce regression coefficient estimates that are difficult to interpret. Extreme caution should be exercised before interpreting these regression coefficient estimates.

The problem of multicollinearity can also affect the t values that are used to evaluate the regression coefficients. Because the problems of multicollinearity among predictors can result in an overestimation of the standard deviation of the regression coefficients, the t values tend to be underrepresentative when multicollinearity is present. In some regression models containing multicollinearity in which all t values are nonsignificant, the overall F value for the model is highly significant. In Section 14.1, an example was given of how including interaction when it is significant strengthens a regression model. The computer output for the regression models both with and without the interaction term was shown in Figure 14.5. The model without interaction produced a statistically significant F value but neither predictor variable was significant. Further investigation of this model reveals that the correlation between the two predictors, x_1 and x_2, is .945. This extremely high correlation indicates a strong collinearity between the two predictor variables.

This collinearity may explain the fact that the overall model is significant but neither predictor is significant. It also underscores one of the problems with multicollinearity: underrepresented t values. The t values test the strength of the predictor given the other variables in the model. If a predictor is highly correlated with other independent variables, it will appear not to add much to the explanation of y and produce a low t value. However, had the predictor not been in the presence of these other correlated variables, the predictor might have explained a high proportion of variation of y.

Many of the problems created by multicollinearity are interpretation problems. The business researcher should be alert to and aware of multicollinearity potential with the predictors in the model and view the model outcome in light of such potential.

The problem of multicollinearity is not a simple one to overcome. However, several methods offer an approach to the problem. One way is to examine a correlation matrix like the one in Table 14.15 to search for possible intercorrelations among potential predictor variables. If several variables are highly correlated, the researcher can select the variable that is most correlated to the dependent variable and use that variable to represent the others in the analysis. One problem with this idea is that correlations can be more complex than simple correlation among variables. In other words, simple correlation values do not always reveal multiple correlation between variables. In some instances, variables may not appear to be correlated as pairs, but one variable is a linear combination of several other variables. This situation is also an example of multicollinearity, and a cursory observation of the correlation matrix will probably not reveal the problem.

Stepwise regression is another way to prevent the problem of multicollinearity. The search process enters the variables one at a time and compares the new variable to those in solution. If a new variable is entered and the t values on old variables become nonsignificant, the old variables are dropped out of solution. In this manner, it is more difficult for the problem of multicollinearity to affect the regression analysis. Of course, because of multicollinearity, some important predictors may not enter in to the analysis.

Other techniques are available to attempt to control for the problem of multicollinearity. One is called a **variance inflation factor**, in which a regression analysis is conducted to predict an independent variable by the other independent variables. In this case, the independent variable being predicted becomes the dependent variable. As this process is done for each of the independent variables, it is possible to determine whether any of the independent variables are a function of the other independent variables, yielding

evidence of multicollinearity. By using the results from such a model, a variance inflation factor (VIF) can be computed to determine whether the standard errors of the estimates are inflated:

$$\text{VIF} = \frac{1}{1 - R_i^2}$$

where R_i^2 is the coefficient of determination for any of the models, used to predict an independent variable by the other $k - 1$ independent variables. Some researchers follow a guideline that any variance inflation factor greater than 10 or R_i^2 value more than .90 for the largest variance inflation factors indicates a severe multicollinearity problem.**

14.4 PROBLEMS

14.19 Develop a correlation matrix for the independent variables in Problem 14.13. Study the matrix and make a judgment as to whether substantial multicollinearity is present among the predictors. Why or why not?

14.20 Construct a correlation matrix for the four independent variables for Problem 14.14 and search for possible multicollinearity. What did you find and why?

14.21 In Problem 14.17, you were asked to use stepwise regression to predict premiums earned by net income, dividends, and underwriting gain or loss. Study the stepwise results, including the regression coefficients, to determine whether there may be a problem with multicollinearity. Construct a correlation matrix of the three variables to aid you in this task.

14.22 Study the three predictor variables in Problem 14.18 and attempt to determine whether substantial multicollinearity is present between the predictor variables. If there is a problem of multicollinearity, how might it affect the outcome of the multiple regression analysis?

Determining Compensation for CEOs

One statistical tool that can be used to study CEO compensation is multiple regression analysis. Regression

models can be developed using predictor variables, such as age, years of experience, worth of company, or others, for analyzing CEO compensation. Search procedures such as stepwise regression can be used to sort out the more significant predictors of CEO compensation.

The researcher prepares for the multiple regression analysis by conducting a study of CEOs and gathering data on several variables. The data presented in the Decision Dilemma could be used for such an analysis. It seems reasonable to believe that CEO compensation is related to the size and worth of a company, therefore it makes sense to attempt to develop a regression model or models to predict CEO compensation by the variables company sales, number of employees in the company, and the capital investment of a company. Qualitative or dummy variables can also be used in such an analysis. In the database given in the Decision Dilemma, one variable indicates whether a company is a manufacturing company. One way to recode this variable for regression analysis is to assign a 1 to companies that are manufacturers and a 0 to others.

A stepwise regression procedure can sort out the variables that seem to be more important predictors of CEO compensation. A stepwise regression analysis was conducted on the Decision Dilemma database using sales, number of employees, capital investment, and whether a company is in manufacturing as the four independent variables. The result of this analysis follows.

**William Mendenhall and Terry Sincich, *A Second Course in Business Statistics: Regression Analysis*. San Francisco: Dellen Publishing Company, 1989; John Neter, William Wasserman, and Michael H. Kutner, *Applied Linear Regression Models*, 2nd ed. Homewood, IL: Richard D. Irwin, 1989.

```
Stepwise Regression: Cash Compen versus Sales,
No. of Emp.,...
Alpha-to-Enter: 0.15 Alpha-to-Remove: 0.15
Response is Cash Com on 4 predictors, with N = 20
Step            1        2        3        4
Constant    243.9    232.2    223.8    223.3

No. of E   0.0696   0.1552   0.0498
T-Value     13.67     4.97     0.98
P-Value     0.000    0.000    0.343

Cap. Inv             −1.66    −2.92    −3.06
T-Value              −2.77    −3.97    −4.27
P-Value              0.013    0.001    0.001

Sales                         1.08     1.45
T-Value                       2.46     6.10
P-Value                      0.026    0.000

S            32.6     27.9     24.5     24.5
R-Sq        91.22    93.95    95.61    95.34
R-Sq(adj)   90.73    93.24    94.78    94.80
```

The stepwise regression analysis produces a single predictor model at step 1 with a high R^2 value of .9122. The number of employees variable used in a simple regression model accounts for over 91.2% of the variation of CEO compensation data. An examination of the regression coefficient of number of employees at the first step (.0696) indicates that a one-employee increase results in a predicted increase of (.0696 · $1,000) about $70 in the CEO's compensation.

At step 2, the company's capital investment enters the model. Notice that the R^2 increases only by .0273 and that the regression coefficient on capital investment is negative. This result seems counterintuitive because we would expect that the more capital investment a company has, the more the CEO should be compensated for the responsibility. A Minitab simple regression analysis using only capital investment produces the following model:

The regression equation is
CashCompen = 257 + 1.29 CapInv

Notice that the regression coefficient in this model is positive as we would suppose. Multicollinearity is likely. In fact, multicollinearity is evident among sales, number of employees, and capital investment. Each is a function or determiner of company size. Examine the following correlation coefficient:

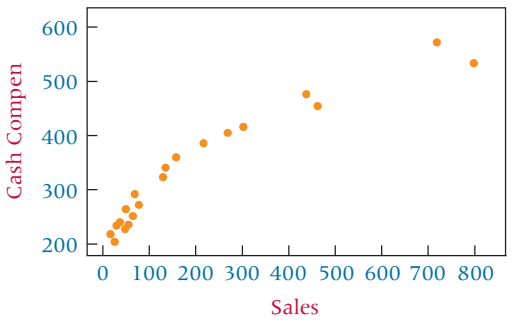
Correlations

	Sales	No. Employees
No. Employees	0.997	1
Cap. Invest	0.995	.999

Notice that these three predictors are highly interrelated. Therefore, the interpretation of the regression coefficients and the order of entry of these variables in the stepwise regression become more difficult. Nevertheless, number of employees is most highly related to CEO compensation in these data. Observe also in the stepwise regression output that number of employees actually drops out of the model at step 4. The *t* ratio for number of employees is not significant ($t = 0.98$) at step 3. However, the R^2 actually drops slightly when number of employees are removed. In searching for a model that is both parsimonious and explanatory, the researcher could do worse than to merely select the model at step 1.

Researchers might want to explore more complicated nonlinear models. Some of the independent variables might be related to CEO compensation but in some nonlinear manner.

A brief study of the predictor variables in the Decision Dilemma database reveals that as compensation increases, the values of the data in the independent variables do not increase at a linear rate. Scatter plots of sales, number of employees, and capital investment with CEO compensation confirm this suspicion. Shown here is a scatter plot of sales with cash compensation.



Observe that the graph suggests more of a logarithmic fit than a linear one. We can use recoding techniques presented in the chapter to conduct a multiple regression analysis to predict compensation using the log of each of these variables. In the analysis, the compensation figures remain the same, but each of the three quantitative independent variables are recoded by taking the log of each value and entering the resultant variable in the model. A second stepwise regression analysis is under- taken with the log variables in the mix along with the original variables. The results follow:

```
Stepwise Regression: Cash Compen versus Sales,
No. of Emp., ...
Alpha-to-Enter: 0.1 Alpha-to-Remove: 0.1
Response is Cash Com on 7 predictors, with N = 20
Step            1        2        3        4        5
Constant   −129.61   −13.23  −122.53  −147.22  −120.74

Log sale     224.3    152.2    281.4    307.8    280.8
T-Value      22.22     8.75    11.08    32.75    26.81
P-Value      0.000    0.000    0.000    0.000    0.000

NO. Emp             0.0251   0.0233   0.0903   0.0828
T-Value               4.53     6.97    13.94    15.52
P-Value              0.000    0.000    0.000    0.000

Log cap                      −106.4   −126.0   −109.8
T-Value                      −5.58   −17.87   −15.56
P-Value                      0.000    0.000    0.000

Sales                                −0.434   −0.250
T-Value                              −10.52    −4.11
P-Value                               0.000    0.001

Cap. Inv                                      −0.37
T-Value                                       −3.51
P-Value                                       0.003

S            20.7     14.3     8.59     3.07     2.32
R-Sq        96.48    98.41    99.46    99.94    99.97
R-Sq(adj)   96.29    98.22    99.36    99.92    99.95
```

Note that in this stepwise regression analysis, the variable log sales has the highest single predictability of compensation producing an R^2 of .9648, which is higher than the value at step 1 in the first stepwise regression analysis. Number of employees enters at step 2 and log of capital investment at step 3. However, such a high R^2 at step 1 leaves little room for improved predictability. Our search through the variables may well end with the decision to use the log of sales as the efficient, predictable model of compensation. The final model might be:

$$\text{CEO Compensation} = -129.61 + 224.3 \text{ Log sales}$$

Human resource managers sometimes use compensation tables to assist them in determining ranges and ballparks for salary offers. Company boards of directors can use such models as the one developed here to assist them in negotiations with possible candidates for CEO positions or to aid them in determining whether a presently employed CEO is over- or undercompensated. In addition, candidates who are searching for new CEO opportunities can use models like these to determine the potential compensation for a new position and to help them be more adequately prepared for salary negotiations should they be offered a CEO position.

Some of the variables in this study will undoubtedly produce redundant information. The use of a correlation matrix and a stepwise regression process can protect the analysis from some of the problems of multicollinearity. The use of multiple regression analysis on a large sample of CEO compensation data with many independent variables could provide some interesting and exciting results.

ETHICAL CONSIDERATIONS

Some business researchers misuse the results of search procedures by using the order in which variables come into a model (on stepwise and forward selection) to rank the variables in importance. They state that the variable entered at step 1 is the most important predictor of y, the variable entering at step 2 is second most important, and so on. In actuality, variables entering the analysis after step 1 are being analyzed by how much of the unaccounted-for variation (residual variation) they are explaining, not how much they are related to y by themselves. A variable that comes into the model at the fourth step is the variable that most greatly accounts for the variation of the y values left-over after the first three variables have explained the rest. However, the fourth variable taken by itself might explain more variation of y than the second or third variable when seen as single predictors.

Some people use the estimates of the regression coefficients to compare the worth of the predictor variables; the larger the coefficient is, the greater its worth. At least two problems plague this approach. The first is that most variables are measured in different units. Thus, regression coefficient weights are partly a function of the unit of measurement of the variable. Second, if multicollinearity is present, the interpretation of the regression coefficients is questionable. In addition, the presence of multicollinearity raises several issues about the interpretation of other regression output. Researchers who ignore this problem are at risk of presenting spurious results.

SUMMARY

Multiple regression analysis can handle nonlinear independent variables. One way to accommodate this issue is to recode the data and enter the variables into the analysis in the normal way. Other nonlinear regression models, such as exponential models, require that the entire model be transformed. Often the transformation involves the use of logarithms. In some cases, the resulting value of the regression model is in logarithmic form and the antilogarithm of the answer must be taken to determine the predicted value of y.

Indicator, or dummy, variables are qualitative variables used to represent categorical data in the multiple regression model. These variables are coded as 0, 1 and are often used to represent nominal or ordinal classification data that the researcher wants to include in the regression analysis. If a qualitative variable contains more than two categories, it generates multiple dummy variables. In general, if a qualitative variable contains c categories, $c - 1$ dummy variables should be created.

Search procedures are used to help sort through the independent variables as predictors in the examination of various possible models. Several search procedures are available, including all possible regressions, stepwise regression, forward selection, and backward elimination. The all possible regressions procedure computes every possible regression model for a set of data. The drawbacks of this procedure include the time and energy required to compute all possible regressions and the difficulty of deciding which models are most appropriate. The stepwise regression procedure involves selecting and adding one independent variable at a time to the regression process after beginning with a one-predictor model. Variables are added to the model at each step if they contain the most significant t value associated with the remaining variables. If no additional t value is statistically significant at any given step, the procedure stops. With stepwise regression, at each step the process examines the variables already in the model to determine whether their t values are still significant.

If not, they are dropped from the model, and the process searches for other independent variables with large, significant t values to replace the variable(s) dropped. The forward selection procedure is the same as stepwise regression but does not drop variables out of the model once they have been included. The backward elimination procedure begins with a "full" model, a model that contains all the independent variables. The sample size must be large enough to justify a full model, which can be a limiting factor. Backward elimination drops out the least important predictors one at a time until only significant predictors are left in the regression model. The variable with the smallest absolute t value of the statistically nonsignificant t values is the independent variable that is dropped out of the model at each step.

One of the problems in using multiple regression is multicollinearity, or correlations among the predictor variables. This problem can cause overinflated estimates of the standard deviations of regression coefficients, misinterpretation of regression coefficients, undersized t values, and misleading signs on the regression coefficients. It can be lessened by using an intercorrelation matrix of independent variables to help recognize bivariate correlation, by using stepwise regression to sort the variables one at a time, or by using statistics such as a variance inflation factor.

KEY TERMS

all possible regressions
backward elimination
dummy variable
forward selection
indicator variable
multicollinearity
quadratic model
qualitative variable
search procedures
stepwise regression
Tukey's four-quadrant approach
Tukey's ladder of transformations
variance inflation factor

FORMULAS

Variance inflation factor

$$\text{VIF} = \frac{1}{1 - R_i^2}$$

SUPPLEMENTARY PROBLEMS

CALCULATING THE STATISTICS

14.23 Given here are the data for a dependent variable, y, and independent variables. Use these data to develop a regression model to predict y. Discuss the output. Which variable is an indicator variable? Was it a significant predictor of y?

x_1	x_2	x_3	y
0	51	16.4	14
0	48	17.1	17
1	29	18.2	29
0	36	17.9	32
0	40	16.5	54
1	27	17.1	86
1	14	17.8	117
0	17	18.2	120
1	16	16.9	194
1	9	18.0	203
1	14	18.9	217
0	11	18.5	235

14.24 Use the following data and a stepwise regression analysis to predict y. In addition to the two independent variables given here, include three other predictors in your analysis: the square of each x as a predictor and an interaction predictor. Discuss the results of the process.

x_1	x_2	y	x_1	x_2	y
10	3	2002	5	12	1750
5	14	1747	6	8	1832
8	4	1980	5	18	1795
7	4	1902	7	4	1917
6	7	1842	8	5	1943
7	6	1883	6	9	1830
4	21	1697	5	12	1786
11	4	2021			

14.25 Use the x_1 values and the log of the x_1 values given here to predict the y values by using a stepwise regression procedure. Discuss the output. Were either or both of the predictors significant?

y	x_1	y	x_1
20.4	850	13.2	204
11.6	146	17.5	487
17.8	521	12.4	192
15.3	304	10.6	98
22.4	1029	19.8	703
21.9	910	17.4	394
16.4	242	19.4	647

TESTING YOUR UNDERSTANDING

14.26 The U.S. Commodities Futures Trading Commission reports on the volume of trading in the U.S. commodity futures exchanges. Shown here are the figures for grain, oilseeds, and livestock products over a period of several years. Use these data to develop a multiple regression model to predict grain futures volume of trading from oilseeds volume and livestock products volume. All figures are given in units of millions. Graph each of these predictors separately with the response variable and use Tukey's four-quadrant approach to explore possible recoding schemes for nonlinear relationships. Include any of these in the regression model. Comment on the results.

Grain	Oilseeds	Livestock
2.2	3.7	3.4
18.3	15.7	11.8
19.8	20.3	9.8
14.9	15.8	11.0
17.8	19.8	11.1
15.9	23.5	8.4
10.7	14.9	7.9
10.3	13.8	8.6
10.9	14.2	8.8
15.9	22.5	9.6
15.9	21.1	8.2

14.27 The U.S. Bureau of Mines produces data on the price of minerals. Shown here are the average prices per year for several minerals over a decade. Use these data and a stepwise regression procedure to produce a model to predict the average price of gold from the other variables. Comment on the results of the process.

Gold ($ per oz.)	Copper (cents per lb.)	Silver ($ per oz.)	Aluminum (cents per lb.)
161.1	64.2	4.4	39.8
308.0	93.3	11.1	61.0
613.0	101.3	20.6	71.6
460.0	84.2	10.5	76.0
376.0	72.8	8.0	76.0
424.0	76.5	11.4	77.8
361.0	66.8	8.1	81.0
318.0	67.0	6.1	81.0
368.0	66.1	5.5	81.0
448.0	82.5	7.0	72.3
438.0	120.5	6.5	110.1
382.6	130.9	5.5	87.8

14.28 The Shipbuilders Council of America in Washington, D.C., publishes data about private shipyards. Among the variables reported by this organization are the employment figures (per 1000), the number of naval vessels under construction, and the number of repairs or conversions done to commercial ships (in $ millions). Shown here are the data for these three variables over a seven-year period. Use the data to develop a regression model to predict private shipyard employment from number of naval vessels under construction and repairs or conversions of commercial ships. Graph each of these predictors separately with the response variable and use Tukey's four-quadrant approach to explore possible recoding schemes for nonlinear relationships. Include any of these in the regression model. Comment on the regression model and its strengths and its weaknesses.

Employment	Naval Vessels	Commercial Ship Repairs or Conversions
133.4	108	431
177.3	99	1335
143.0	105	1419
142.0	111	1631
130.3	100	852
120.6	85	847
120.4	79	806

14.29 The U.S. Bureau of Labor Statistics produces consumer price indexes for several different categories. Shown here are the percentage changes in consumer price indexes over a period of 20 years for food, shelter, apparel, and fuel oil. Also displayed are the percentage changes in consumer price indexes for all commodities. Use these data and a stepwise regression procedure to develop a model that attempts to predict all commodities by the other four variables. Construct scatter plots of each of these variables with all commodities. Examine the graphs in light of Tukey's four-quadrant approach. Develop any other appropriate predictor variables by recoding data and include them in the analysis. Comment on the result of this analysis.

All Commodities	Food	Shelter	Apparel	Fuel Oil
.9	1.0	2.0	1.6	3.7
.6	1.3	.8	.9	2.7
.9	.7	1.6	.4	2.6
.9	1.6	1.2	1.3	2.6
1.2	1.3	1.5	.9	2.1
1.1	2.2	1.9	1.1	2.4
2.6	5.0	3.0	2.5	4.4
1.9	.9	3.6	4.1	7.2
3.5	3.5	4.5	5.3	6.0
4.7	5.1	8.3	5.8	6.7
4.5	5.7	8.9	4.2	6.6
3.6	3.1	4.2	3.2	6.2
3.0	4.2	4.6	2.0	3.3
7.4	14.5	4.7	3.7	4.0
11.9	14.3	9.6	7.4	9.3
8.8	8.5	9.9	4.5	12.0
4.3	3.0	5.5	3.7	9.5
5.8	6.3	6.6	4.5	9.6
7.2	9.9	10.2	3.6	8.4
11.3	11.0	13.9	4.3	9.2

14.30 The U.S. Department of Agriculture publishes data annually on various selected farm products. Shown here are the unit production figures for three farm

products for 10 years during a 20-year period. Use these data and a stepwise regression analysis to predict corn production by the production of soybeans and wheat. Comment on the results.

Corn (million bushels)	Soybeans (million bushels)	Wheat (million bushels)
4152	1127	1352
6639	1798	2381
4175	1636	2420
7672	1861	2595
8876	2099	2424
8226	1940	2091
7131	1938	2108
4929	1549	1812
7525	1924	2037
7933	1922	2739

14.31 The American Chamber of Commerce Researchers Association compiles cost-of-living indexes for selected metropolitan areas. Shown here are cost-of-living indexes for 25 different cities on five different items for a recent year. Use the data to develop a regression model to predict the grocery cost-of-living index by the indexes of housing, utilities, transportation, and healthcare. Discuss the results, highlighting both the significant and nonsignificant predictors.

City	Grocery Items	Housing	Utilities	Transportation	Healthcare
Albany	108.3	106.8	127.4	89.1	107.5
Albuquerque	96.3	105.2	98.8	100.9	102.1
Augusta, GA	96.2	88.8	115.6	102.3	94.0
Austin	98.0	83.9	87.7	97.4	94.9
Baltimore	106.0	114.1	108.1	112.8	111.5
Buffalo	103.1	117.3	127.6	107.8	100.8
Colorado Springs	94.5	88.5	74.6	93.3	102.4
Dallas	105.4	98.9	108.9	110.0	106.8
Denver	91.5	108.3	97.2	105.9	114.3
Des Moines	94.3	95.1	111.4	105.7	96.2
El Paso	102.9	94.6	90.9	104.2	91.4
Indianapolis	96.0	99.7	92.1	102.7	97.4
Jacksonville	96.1	90.4	96.0	106.0	96.1
Kansas City	89.8	92.4	96.3	95.6	93.6
Knoxville	93.2	88.0	91.7	91.6	82.3
Los Angeles	103.3	211.3	75.6	102.1	128.5
Louisville	94.6	91.0	79.4	102.4	88.4
Memphis	99.1	86.2	91.1	101.1	85.5
Miami	100.3	123.0	125.6	104.3	137.8
Minneapolis	92.8	112.3	105.2	106.0	107.5
Mobile	99.9	81.1	104.9	102.8	92.2
Nashville	95.8	107.7	91.6	98.1	90.9
New Orleans	104.0	83.4	122.2	98.2	87.0
Oklahoma City	98.2	79.4	103.4	97.3	97.1
Phoenix	95.7	98.7	96.3	104.6	115.2

INTERPRETING THE OUTPUT

14.32 A stepwise regression procedure was used to analyze a set of 20 observations taken on four predictor variables to predict a dependent variable. The results of this procedure are given next. Discuss the results.

```
STEPWISE REGRESSION OF Y ON 4 PREDICTORS,
WITH N = 20
STEP              1        2
CONSTANT      152.2    124.5

X1            −50.6    −43.4
T-RATIO        7.42     6.13

X2                      1.36
T-RATIO                 2.13

S             15.2     13.9
R-SQ         75.39    80.59
```

14.33 Shown here are the data for y and three predictors, x_1, x_2, and x_3. A stepwise regression procedure has been done on these data; the results are also given. Comment on the outcome of the stepwise analysis in light of the data.

y	x_1	x_2	x_3
94	21	1	204
97	25	0	198
93	22	1	184
95	27	0	200
90	29	1	182
91	20	1	159
91	18	1	147
94	25	0	196
98	26	0	228
99	24	0	242
90	28	1	162
92	23	1	180
96	25	0	219

```
Step            1        2        3
Constant    74.81    82.18    87.89

X3          0.099    0.067    0.071
T-Value      6.90     3.65     5.22
P-Value     0.000    0.004    0.001

X2                  -2.26    -2.71
T-Value             -2.32    -3.71
P-Value              0.043    0.005

X1                           -0.256
T-Value                      -3.08
P-Value                      0.013

S           1.37     1.16     0.850
R-Sq       81.24    87.82    94.07
R-Sq(adj)  79.53    85.38    92.09
```

14.34 Shown below is output from two Excel regression analyses on the same problem. The first output was done on a "full" model. In the second output, the variable with the smallest absolute *t* value has been removed, and the regression has been rerun like a second step of a backward elimination process. Examine the two outputs. Explain what happened, what the results mean, and what might happen in a third step.

FULL MODEL

Regression Statistics	
Multiple R	0.567
R Square	0.321
Adjusted R Square	0.208
Standard Error	159.681
Observations	29

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	289856.08	72464.02	2.84	0.046
Residual	24	611955.23	25498.13		
Total	28	901811.31			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	336.79	124.0800	2.71	0.012
X_1	1.65	1.7800	0.93	0.363
X_2	–5.63	13.4700	–0.42	0.680
X_3	0.26	1.6800	0.16	0.878
X_4	185.50	66.2200	2.80	0.010

SECOND MODEL

Regression Statistics	
Multiple R	0.566
R Square	0.321
Adjusted R Square	0.239
Standard Error	156.534
Observations	29

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	289238.1	96412.70	3.93	0.020
Residual	25	612573.20	24502.90		
Total	28	901811.3			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	342.92	11.34	2.97	0.006
X_1	1.83	1.31	1.40	0.174
X_2	–5.75	13.18	–0.44	0.667
X_4	181.22	59.05	3.07	0.005

ANALYZING THE DATABASES

see www.wiley.com/college/black and WileyPLUS

1. Use the Manufacturing database to develop a multiple regression model to predict Cost of Materials by Number of Employees, New Capital Expenditures, Value Added by Manufacture, Value of Industry Shipments, and End-of-Year Inventories. Create indicator variables for values of industry shipments that have been coded from 1 to 4. Use a stepwise regression procedure. Does multicollinearity appear to be a problem in this analysis? Discuss the results of the analysis.

2. Construct a correlation matrix for the Hospital database variables. Are some of the variables highly correlated? Which ones and why? Perform a stepwise multiple regression analysis to predict Personnel by Control, Service, Beds, Admissions, Census, Outpatients, and Births. The variables Control and Service will need to be coded as indicator variables. Control has four subcategories, and Service has two.
3. Develop a regression model using the Financial database. Use Total Revenues, Total Assets, Return on Equity, Earnings per Share, Average Yield, and Dividends per Share to predict the average P/E ratio for a company. How strong is the model? Use stepwise regression to help sort out the variables. Several of these variables may be measuring similar things. Construct a correlation matrix to explore the possibility of multicollinearity among the predictors.
4. Using the International Stock Market database, conduct a stepwise a multiple regression procedure to predict the DJIA by the Nasdaq, the S&P 500, the Nikkei, the Hang Seng, the FTSE 100, and the IPC. Discuss the outcome of the analysis including the model, the strength of the model, and the predictors.

CASE

VIRGINIA SEMICONDUCTOR

Virginia Semiconductor is a leading manufacturer of prime silicon substrates. The company, situated in Fredericksburg, Virginia, was founded in 1978 by Dr. Thomas G. Digges and his brother, Robert. Virginia Semiconductor (VSI) was growing and prospering in the early 1980s by selling a high volume of low-profit-margin wafers in the microelectronic industry. However, in 1985, without notice, VSI lost two major customers that represented 65% of its business. Left with only 35% of its sales base, the company desperately needed customers.

Dr. Digges, CEO of VSI, decided to seek markets where his company's market share would be small but profit margin would be high because of the value of its engineering research and its expertise. This decision turned out to be a wise direction for the small, versatile company. VSI developed a silicon wafer that was two inches in diameter, 75 microns thick, and polished on both sides. Such wafers were needed by several customers but had never been produced before. The company produced a number of these wafers and sold them for more than 10 times the price of conventional wafers.

Soon the company was making wafers from 2 to 4 microns thick (extremely thin), wafers with textured surfaces for infrared applications, and wafers with micromachined holes or shapes and selling them in specialized markets. It was able to deliver these products faster than competitors were able to deliver standard wafers.

Having made inroads at replacing lost sales, Virginia Semiconductor still had to streamline operations and control inventory and expenses. No layoffs occurred, but the average work-week dropped to 32 hours and the president took an 80% pay reduction for a time. Expenses were cut as far as seemed possible. The company had virtually no long-term debt and fortunately was able to make it through this period without incurring any additional significant debt. The absence of large monthly debt payments enabled the company to respond quickly to new production needs.

Virginia Semiconductor improved production quality by cross-training employees. In addition, the company participated in the state of Virginia's economic development efforts to find markets in Europe, Japan, Korea, and Israel. Exports, which were only 1% of the company's business in 1985, grew to over 40%.

The company continues to find new customers because of product development. VSI has distributors of their products in 29 different countries. Underscoring a core value of VSI, it is stated on the company's Web page: "As always, VSI can actually make any silicon wafer to any specification and continues to supply small, complicated orders to valued customers."

Discussion

1. It is often useful to decision makers at a company to determine what factors enter into the size of a customer's purchase. Suppose decision makers at Virginia Semiconductor want to determine from past data what variables might be predictors of size of purchase and are able to gather some data on various customer companies. Assume the following data represent information gathered for 16 companies on five variables: the total amount of purchases made during a one-year period (size of purchase), the size of the purchasing company (in total sales volume), the percentage of all purchases made by the customer company that were imports, the distance of the customer company from Virginia Semiconductor, and whether the customer company had a single central purchasing agent. Use these data to generate a multiple regression model to predict size of purchase by the other variables. Summarize your findings in terms of the strength of the model, significant predictor variables, and any new variables generated by recoding.

Size of Purchase ($1,000)	Company Size ($ million sales)	Percent of Customer Imports	Distance from Virginia Semiconductor	Central Purchaser?
27.9	25.6	41	18	1
89.6	109.8	16	75	0
12.8	39.4	29	14	0
34.9	16.7	31	117	0
408.6	278.4	14	209	1
173.5	98.4	8	114	1
105.2	101.6	20	75	0
510.6	139.3	17	50	1
382.7	207.4	53	35	1
84.6	26.8	27	15	1
101.4	13.9	31	19	0
27.6	6.8	22	7	0
234.8	84.7	5	89	1
464.3	180.3	27	306	1
309.8	132.6	18	73	1
294.6	118.9	16	11	1

2. Suppose that the next set of data is Virginia Semiconductor's sales figures for the past 11 years, along with the average number of hours worked per week by a full-time employee and the number of different customers the company has for its unique wafers. How do the average workweek length and number of customers relate to total sales figures? Use scatter plots to examine possible relationships between sales and hours per week and sales and number of customers. Use Tukey's four-quadrant approach for possible ways to recode the data. Use stepwise regression analysis to explore the relationships. Let the response variable be "sales" and the predictors be "average number of hours worked per week," "number of customers," and any new variables created by recoding. Explore quadratic relationships, interaction, and other relationships that seem appropriate by using stepwise regression. Summarize your findings in terms of model strength and significant predictors.

Average Sales ($ million)	Hours Worked per Week	Number of Customers
15.6	44	54
15.7	43	52
15.4	41	55
14.3	41	55
11.8	40	39
9.7	40	28
9.6	40	37
10.2	38	58
11.3	38	67
14.3	32	186
14.8	37	226

3. As Virginia Semiconductor continues to grow and prosper, the potential for slipping back into inefficient ways is always present. Suppose that after a few years the company's sales begin to level off, but it continues hiring employees. Such figures over a 10-year period of time may look like the data given here. Graph these data, using sales as the response variable and number of employees as the predictor. Study the graph in light of Tukey's four-quadrant approach. Using the information learned, develop a regression model to predict sales by the number of employees. On the basis of what you find, what would you recommend to management about the trend if it were to continue? What do you see in these data that would concern management?

Sales ($ million)	Number of Employees
20.2	120
24.3	122
28.6	127
33.7	135
35.2	142
35.9	156
36.3	155
36.2	167
36.5	183
36.6	210

Source: Adapted from "Virginia Semiconductor: A New Beginning," *RealWorld Lessons for America's Small Businesses: Insights from the Blue Chip Enterprise Initiative 1994.* Published by *Nation's Business* magazine on behalf of Connecticut Mutual Life Insurance Company and the U.S. Chamber of Commerce in association with the Blue Chip Enterprise Initiative, 1994. Virginia Semiconductor's Web site (2009) at: http://www.virginiasemi.com.

USING THE COMPUTER

EXCEL

- Excel does not have Model Building-Search Procedure capability. However, Excel can perform multiple regression analysis. The commands are essentially the same as those for simple regression except that the x range of data may include several columns. Excel will determine the number of predictor variables from the number of columns entered in to **Input X Range**.

- Begin by selecting the **Data** tab on the Excel worksheet. From the **Analysis** panel at the right top of the **Data** tab worksheet, click on **Data Analysis**. If your Excel worksheet does not show the **Data Analysis** option, then you can load it as an add-in following directions given in Chapter 2. From the **Data Analysis** pulldown menu, select **Regression**. In the **Regression** dialog box, input the location of the *y* values in **Input Y Range**. Input the location of the *x* values in **Input X Range**. Input **Labels** and input **Confidence Level**. To pass the line through the origin, check **Constant is Zero**. To printout the raw residuals, check **Residuals**. To printout residuals converted to *z* scores, check **Standardized Residuals**. For a plot of the residuals, check **Residual Plots**. For a plot of the line through the points check **Line Fit Plots**.
- Standard output includes R, R^2, s_e, and an ANOVA table with the F test, the slope and intercept, t statistics with associated p-values, and any optionally requested output, such as graphs or residuals.

MINITAB

- Minitab does have Model Building–Search Procedure capability procedures including both forward and backward Stepwise regression, Forward Selection, and Backward Elimination.
- To begin, select **Stat** from the menu bar. Select **Regression** from the **Stat** pulldown menu. Select **Stepwise** from the **Regression** pulldown menu. Place the column name or column location of the *y* variable in **Response**. Place the column name or column location of the *x* variable(s) in **Predictors**. If you want to guarantee inclusion of particular variables in the model, place the column name or column locations of such variables in **Predictors to include in every model**. This is optional. Select **Methods** for Model Building options and selection of criterion for adding or removing a variable.
- In the **Methods** dialog box, Check **Use alpha values** to use alpha as the criterion for adding or removing a variable. Check **Use F values** to use *F* values as the criterion for adding or removing a variable. Check **Stepwise (forward and backward)** to run a standard forward or backward stepwise regression procedure. To specify particular variables to be included in the initial model, place the column name or column location of such variables in the box labeled **Predictors in initial model**. Check **Forward selection** to run a forward selection regression. Check **Backward elimination** to run a backward elimination regression. In each of these model-building procedures, you have the option of setting particular values of alpha or *F* for the entering and/or removing variables from the model. Minitab defaults to an alpha of 0.15 and an *F* of 4.