CHAPTER 13

Multiple Regression Analysis

LEARNING OBJECTIVES

This chapter presents the potential of multiple regression analysis as a tool in business decision-making and its applications, thereby enabling you to:

1. Explain how, by extending the simple regression model to a multiple regression model with two or more independent variables, it is possible to determine the multiple regression equation for any number of unknowns.
2. Examine significance tests of both the overall regression model and the regression coefficients.
3. Calculate the residual, standard error of the estimate, coefficient of multiple determination, and adjusted coefficient of multiple determination of a regression model.
4. Use a computer to find and interpret multiple regression outputs.

DECISION DILEMMA Are You Going to Hate Your New Job?

Getting a new job can be an exciting and energizing event in your life.

But what if you discover after a short time on the job that you hate your job? Is there any way to determine ahead of time whether you will love or hate your job? Sue Shellenbarger of *The Wall Street Journal* discusses some of the things to look for when interviewing for a position that may provide clues as to whether you will be happy on that job.

Among other things, work cultures vary from hip, freewheeling start-ups to old-school organizational-driven domains. Some organizations place pressure on workers to feel tense and to work long hours while others place more emphasis on creativity and the bottom line. Shellenbarger suggests that job interviewees pay close attention to how they are treated in an interview. Are they just another cog in the wheel or are they valued as an individual? Is a work-life balance apparent within the company? Ask what a typical workday is like at that firm. Inquire about the values that undergird the management by asking questions such as "What is your proudest accomplishment?" Ask about flexible schedules and how job training is managed. For example, do workers have to go to job training on their own time?

A "Work Trends" survey undertaken by the John J. Heldrich Center for Workforce Development at Rutgers University and the Center for Survey Research and Analysis at the University of Connecticut posed several questions to employees in a survey to ascertain their job satisfaction. Some of the themes included in these questions were relationship with your supervisor, overall quality of the work environment, total hours worked each week, and opportunities for advancement at the job.

iStock.com/Jacob Wackerhausen

Suppose another researcher gathered survey data from 19 employees on these questions and also asked the employees to rate their job satisfaction on a scale from 0 to 100 (with 100 being perfectly satisfied). Suppose the following data represent the results of this survey. Assume that relationship with supervisor is rated on a scale from 0 to 50 (0 represents poor relationship and 50 represents an excellent relationship), overall quality of the work environment is rated on a scale from 0 to 100 (0 represents poor work environment and 100 represents an excellent work environment), and opportunities for advancement is rated on a scale from 0 to 50 (0 represents no opportunities and 50 represents excellent opportunities).

Job Satisfaction	Relationship with Supervisor	Overall Quality of Work Environment	Total Hours Worked per Week	Opportunities for Advancement
55	27	65	50	42
20	12	13	60	28
85	40	79	45	7
65	35	53	65	48
45	29	43	40	32
70	42	62	50	41
35	22	18	75	18
60	34	75	40	32
95	50	84	45	48
65	33	68	60	11
85	40	72	55	33
10	5	10	50	21
75	37	64	45	42
80	42	82	40	46
50	31	46	60	48
90	47	95	55	30
75	36	82	70	39
45	20	42	40	22
65	32	73	55	12

Managerial, Statistical, and Analytical Questions

1. Several variables are presented that may be related to job satisfaction. Which variables are stronger predictors of job satisfaction? Might other variables not mentioned here be related to job satisfaction?
2. Is it possible to develop a mathematical model to predict job satisfaction using the data given? If so, how strong is the model? With four independent variables, will we need to develop four different simple regression models and compare their results?

Source: Adapted from Sue Shellenbarger, "How to Find Out if You're Going to Hate a New Job Before You Agree to Take It," *The Wall Street Journal*, June 13, 2002, D1 at https://www.wsj.com/articles/SB10239108069478908 40?mod=searchresults&page=3&pos=10; www.heldrich.rutgers.edu/research/topics/work-trends-surveys.

Simple regression analysis (discussed in Chapter 12) is bivariate linear regression in which one **dependent variable**, y, is predicted by one **independent variable**, x. Examples of simple regression applications include models to predict retail sales by population density, Dow Jones averages by prime interest rates, crude oil production by energy consumption, and CEO compensation by quarterly sales. However, in many cases, other independent variables, taken in conjunction with these variables, can make the regression model a better fit in predicting the dependent variable. For example, sales could be predicted by the size of store and number of competitors in addition to population density. A model to predict the Dow Jones Industrial Average of 30 companies could include, in addition to the prime interest rate, such predictors as yesterday's volume, the bond interest rate, and the producer price index. A model to predict CEO compensation could be developed by using variables such as company earnings per share, age of CEO, and size of company, in addition to quarterly sales. A model could perhaps be developed to predict the cost of outsourcing by such variables as unit price, export taxes, cost of money, damage in transit, and other factors. Each of these examples contains only one dependent variable, y, as with simple regression analysis. However, multiple independent variables, x (predictors), are involved. *Regression analysis with two or more independent variables or with at least one nonlinear predictor* is called **multiple regression** analysis.

13.1 | The Multiple Regression Model

Multiple regression analysis is similar in principle to simple regression analysis. However, it is more complex conceptually and computationally. Recall from Chapter 12 that the equation of the probabilistic simple regression model is

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

where

y = the value of the dependent variable
β_0 = the population y intercept
β_1 = the population slope
ϵ = the error of prediction

Extending this notion to multiple regression gives the general equation for the probabilistic multiple regression model.

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

where

y = the value of the dependent variable
β_0 = the regression constant
β_1 = the partial regression coefficient for independent variable 1
β_2 = the partial regression coefficient for independent variable 2
β_3 = the partial regression coefficient for independent variable 3
β_k = the partial regression coefficient for independent variable k
k = the number of independent variables

In multiple regression analysis, the dependent variable, y, is sometimes referred to as the **response variable**. The **partial regression coefficient** of an independent variable, β_i, *represents the increase that will occur in the value of y from a one-unit increase in that independent variable if all other variables are held constant.* The "full" (versus partial) regression coefficient of an independent variable is a coefficient obtained from the bivariate model (simple regression) in which the independent variable is the sole predictor of y. The partial regression coefficients occur because more than one predictor is included in a model. The partial regression coefficients are analogous to β_1, the slope of the simple regression model in Chapter 12.

In actuality, the partial regression coefficients and the regression constant of a multiple regression model are population values and are unknown. In virtually all research, these values are estimated by using sample information. Shown here is the form of the equation for estimating y with sample information.

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

where

$\hat{y}$ = the predicted value of y
b_0 = the estimate of the regression constant
b_1 = the estimate of regression coefficient 1
b_2 = the estimate of regression coefficient 2
b_3 = the estimate of regression coefficient 3
b_k = the estimate of regression coefficient k
k = the number of independent variables

Multiple Regression Model with Two Independent Variables (First-Order)

The simplest multiple regression model is one constructed with two independent variables, where the highest power of either variable is 1 (first-order regression model). This regression model is

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

The constant and coefficients are estimated from sample information, resulting in the following model.

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

Figure 13.1 is a three-dimensional graph of a series of points (x_1, x_2, y) representing values from three variables used in a multiple regression model to predict the sales price of a house by the number of square feet in the house and the age of the house. Simple regression models yield a line that is fit through data points in the xy plane. In multiple regression analysis, the resulting model produces a **response surface**. In the multiple regression model shown here with two independent first-order variables, the response surface is a **response plane**. The response plane for such a model is fit in a three-dimensional space (x_1, x_2, y).

If such a response plane is fit into the points shown in Figure 13.1, the result is the graph in **Figure 13.2**. Notice that most of the points are not on the plane. As in simple regression, an error in the fit of the model in multiple regression is usually present. The distances shown in the graph from the points to the response plane are the errors of fit, or residuals $(y - \hat{y})$. Multiple regression models with three or more independent variables involve more than three dimensions and are difficult to depict geometrically.

Observe in Figure 13.2 that the regression model attempts to fit a plane into the three-dimensional plot of points. Notice that the plane intercepts the y axis. Figure 13.2 depicts some values of y for various values of x_1 and x_2. The error of the response plane (ϵ) in predicting or determining the y values is the distance from the points to the plane.

Determining the Multiple Regression Equation

The simple regression equations for determining the sample slope and intercept given in Chapter 12 are the result of using methods of calculus to minimize the sum of squares of error for the regression model. The procedure for developing these equations involves solving two simultaneous equations with two unknowns, b_0 and b_1. Finding the sample slope and intercept from these formulas requires the values of Σx, Σy, Σxy, and Σx^2.

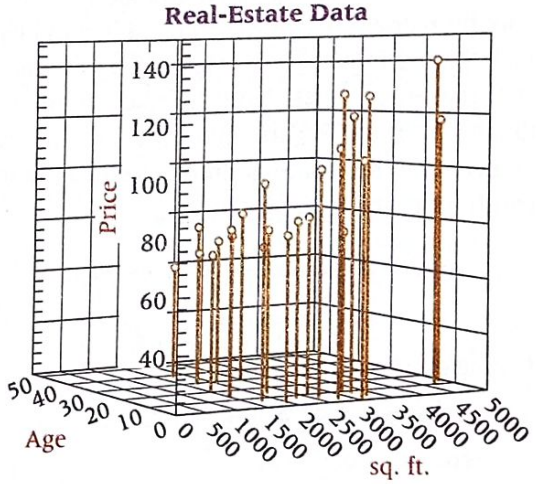


FIGURE 13.1 **Points in a Sample Space**

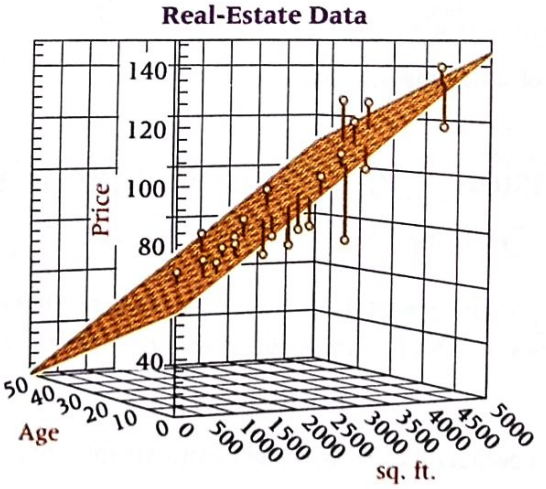


FIGURE 13.2 **Response Plane for a First-Order Two-Predictor Multiple Regression Model**

The procedure for determining formulas to solve for multiple regression coefficients is similar. The formulas are established to meet an objective of *minimizing the sum of squares of error for the model.* Hence, the regression analysis shown here is referred to as **least squares analysis**. Methods of calculus are applied, resulting in $k + 1$ equations with $k + 1$ unknowns (b_0 and k values of b_i) for multiple regression analyses with k independent variables. Thus, a regression model with six independent variables will generate seven simultaneous equations with seven unknowns (b_0, b_1, b_2, b_3, b_4, b_5, b_6).

For multiple regression models with two independent variables, the result is three simultaneous equations with three unknowns (b_0, b_1, and b_2).

$$b_0 n + b_1 \Sigma x_1 + b_2 \Sigma x_2 = \Sigma y$$

$$b_0 \Sigma x_1 + b_1 \Sigma x_1^2 + b_2 \Sigma x_1 x_2 = \Sigma x_1 y$$

$$b_0 \Sigma x_2 + b_1 \Sigma x_1 x_2 + b_2 \Sigma x_2^2 = \Sigma x_2 y$$

The process of solving these equations by hand is tedious and time-consuming. Solving for the regression coefficients and regression constant in a multiple regression model with two independent variables requires Σx_1, Σx_2, Σy, Σx_1^2, Σx_2^2, $\Sigma x_1 x_2$, $\Sigma x_1 y$, and $\Sigma x_2 y$. In actuality, virtually all business analysts use computer statistical software packages to solve for the regression coefficients, the regression constant, and other pertinent information. In this chapter, we will discuss computer output and assume little or no hand calculation. The emphasis will be on the interpretation of the computer output.

A Multiple Regression Model

A real-estate study was conducted in a small Louisiana city to determine what variables, if any, are related to the market price of a home. Several variables were explored, including the number of bedrooms, the number of bathrooms, the age of the house, the number of square feet of living space, the total number of square feet of space, and the number of garages. Suppose the business analyst wants to develop a regression model to predict the market price of a home by two variables, "total number of square feet in the house" and "the age of the house." Listed in **Table 13.1** are the data for these three variables.

A number of statistical software packages can perform multiple regression analysis, including Excel and Minitab. The output for the Minitab multiple regression analysis on the real-estate data is given in **Figure 13.3**. (An example of Excel output is shown in Demonstration Problem 13.1.)

This Minitab output for regression analysis ends with "Regression Equation." From Figure 13.3, the regression equation for the real-estate data in Table 13.1 is

$$\hat{y} = 57.4 + .0177x_1 - .666x_2$$

The regression constant, 57.4, is the y-intercept. The y-intercept is the value of $\hat{y}$ if both x_1 (number of square feet) and x_2 (age) are zero. In this example, a practical understanding of the y-intercept is meaningless. It makes little sense to say that a house containing no square feet ($x_1 = 0$) and no years of age ($x_2 = 0$) would cost \$57,400. Note in Figure 13.2 that the response plane crosses the y-axis (price) at 57.4.

The coefficient of x_1 (total number of square feet in the house) is .0177, which means that a one-unit increase in square footage would result in a predicted increase of .0177· (\$1,000) = \$17.70 in the price of the home if age were held constant. All other variables being held constant, the addition of 1 square foot of space in the house results in a predicted increase of \$17.70 in the price of the home.

The coefficient of x_2 (age) is –.666. The negative sign on the coefficient denotes an inverse relationship between the age of a house and the price of the house: the older the house, the lower the price. In this case, if the total number of square feet in the house is kept constant, a one-unit increase in the age of the house (1 year) will result in –.666· (\$1,000) = –\$666, a predicted \$666 drop in the price.

TABLE 13.1 Real-Estate Data

Market Price ($1,000s) y	Total Number of Square Feet x_1	Age of House (years) x_2
63.0	1605	35
65.1	2489	45
69.9	1553	20
76.8	2404	32
73.9	1884	25
77.9	1558	14
74.9	1748	8
78.0	3105	10
79.0	1682	28
83.4	2470	30
79.5	1820	2
83.9	2143	6
79.7	2121	14
84.5	2485	9
96.0	2300	19
109.5	2714	4
102.5	2463	5
121.0	3076	7
104.9	3048	3
128.0	3267	6
129.0	3069	10
117.9	4765	11
140.0	4540	8

Regression Analysis: Price Versus Square Feet, Age

Analysis of Variance

Source	df	Adj SS	Adj MS	F-Value	P-Value
Regression	2	8190	4094.9	28.63	0.000
Square Feet	1	4538	4538.5	31.73	0.000
Age	1	1222	1221.9	8.54	0.008
Error	20	2861	143.1		
Total	22	11051			

Model Summary

S	R-sq	R-sq(adj)
11.9604	74.11%	71.52%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	57.4	10.0	5.73	0.000
Square Feet	0.0177	0.0032	5.63	0.000
Age	-0.666	0.228	-2.92	0.008

Regression Equation

Price = 57.4 + 0.0177 Square Feet - 0.666 Age

FIGURE 13.3 Minitab Output of Regression for the Real-Estate Example

In examining the regression coefficients, it is important to remember that the independent variables are often measured in different units. It is usually not wise to compare the regression coefficients of predictors in a multiple regression model and decide that the variable with the largest regression coefficient is the best predictor. In this example, the two variables are in different units, square feet and years. Just because x_2 has the larger coefficient (.666) does not necessarily make x_2 the strongest predictor of y.

This regression model can be used to predict the price of a house in this small Louisiana city. If the house has 2500 square feet total and is 12 years old, $x_1 = 2500$ and $x_2 = 12$. Substituting these values into the regression model yields

$$\hat{y} = 57.4 + .0177x_1 - .666x_2$$
$$= 57.4 + .0177(2500) - .666(12) = 93.658$$

The predicted price of the house is $93,658. Figure 13.2 is a graph of these data with the response plane and the residual distances.

DEMONSTRATION PROBLEM 13.1

Since 1980, the prime interest rate in the United States has varied from less than 5% to over 15%. What factor in the U.S. economy seems to be related to the prime interest rate? Two possible predictors of the prime interest rate are the annual unemployment rate and the savings rate in the United States. Shown below are data for the annual prime interest rate over a 15-year period in the United States along with the annual unemployment rate and the annual average personal saving rate (as a percentage of disposable personal income). Use these data to develop a multiple regression model to predict the annual prime interest rate by the unemployment rate and the average personal saving. Determine the predicted prime interest rate if the unemployment rate is 6.5 and the average personal saving is 5.0.

Year	Prime Interest Rate	Unemployment Rate	Personal Saving Rate
1	8.33	7.0	8.2
2	9.32	5.5	7.3
3	10.01	5.6	7.0
4	6.25	7.5	7.7
5	7.15	6.1	4.8
6	8.27	5.4	4.0
7	8.35	4.5	4.3
8	9.23	4.0	2.3
9	4.67	5.8	2.4
10	4.34	5.5	2.1
11	7.96	4.6	0.7
12	5.09	5.8	1.8
13	3.25	9.6	5.8
14	3.25	8.1	7.6
15	3.25	6.2	4.8

Solution The following output shows the results of analyzing the data by using the regression portion of Excel.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.820
R Square	0.672
Adjusted R Square	0.617
Standard Error	1.496
Observations	15

ANOVA

	df	SS	MS	F	Significance F
Regression	2	54.9835	27.4917	12.29	0.0012
Residual	12	26.8537	2.2378		
Total	14	81.8372			

	Coefficients	Standard Error	*t* Stat	*P*-value
Intercept	13.5786	1.728	7.86	0.0000
Unemployment Rates	−1.6622	0.337	−4.93	0.0003
Personal Savings	0.6586	0.199	3.31	0.0062

The regression equation is

$$\hat{y} = 13.5786 - 1.6622x_1 + 0.6586x_2$$

where

$\hat{y}$ = prime interest rate

x_1 = unemployment rate

x_2 = personal saving

The model indicates that for every one-unit (1%) increase in the unemployment rate, the predicted prime interest rate decreases by 1.6622%, if personal saving is held constant. The model also indicates that for every one-unit (1%) increase in personal saving, the predicted prime interest rate increases by 0.6586%, if unemployment is held constant.

If the unemployment rate is 6.5 and the personal saving rate is 5.0, the predicted prime interest rate is 6.07%.

$$\hat{y} = 13.5786 - 1.6622(6.5) + 0.6586(5.0) = 6.07$$

13.1 Problems

13.1. Use a computer to develop the equation of the regression model for the following data. Comment on the regression coefficients. Determine the predicted value of y for $x_1 = 230$ and $x_2 = 9$.

y	x_1	x_2
22	204	5
28	311	11
41	219	6
38	232	10
62	179	11
57	218	14
48	245	7
32	180	13
46	197	10
27	165	7

13.2. Use a computer to develop the equation of the regression model for the following data. Comment on the regression coefficients. Determine the predicted value of y for $x_1 = 41$, $x_2 = 33$, and $x_3 = 15$.

y	x_1	x_2	x_3
139	28	12	9
119	51	31	12
112	64	48	29
123	27	33	13
126	37	26	16
110	42	51	25
119	48	39	18
132	40	20	15
144	24	10	11
118	26	37	20
133	35	18	14
142	39	9	12

13.3. Using the following data, determine the equation of the regression model. How many independent variables are there? Comment on the meaning of these regression coefficients.

Predictor	Coefficient
Constant	121.62
x_1	−.174
x_2	6.02
x_3	.00026
x_4	.0041

13.4. Use the following data to determine the equation of the multiple regression model. Comment on the regression coefficients.

Predictor	Coefficient
Constant	32,563.2
x_1	.06785
x_2	311.26
x_3	−.0853

13.5. Is there a particular product that is an indicator of per capita personal consumption for countries around the world? Shown here are data on per capita personal consumption, paper consumption, fish consumption, and gasoline consumption for 11 countries. Use the data to develop a multiple regression model to predict per capita personal consumption by paper consumption, fish consumption, and gasoline consumption. Discuss the meaning of the partial regression weights.

Country	Per Capita Personal Consumption ($ U.S.)	Paper Consumption (kg per person)	Fish Consumption (lbs per person)	Gasoline Consumption (liters per person)
Bangladesh	836	1	23	2
Greece	3,145	85	53	394
Italy	21,785	204	48	368
Japan	37,931	250	141	447
Kenya	276	4	12	16
Norway	1,913	156	113	477
Philippines	2,195	19	65	43
Portugal	3,154	116	133	257
United Kingdom	19,539	207	44	460
United States	109,521	308	47	1,624
Venezuela	622	27	40	528

13.6. Jensen, Solberg, and Zorn investigated the relationship of insider ownership, debt, and dividend policies in companies. One of their findings was that firms with high insider ownership choose lower levels of both debt and dividends. Shown here is a sample of data of these three variables for 11 different industries. Use the data to develop the equation of the regression model to predict insider ownership by debt ratio and dividend payout. Comment on the regression coefficients.

Industry	Insider Ownership	Debt Ratio	Dividend Payout
Mining	8.2	14.2	10.4
Food and beverage	18.4	20.8	14.3
Furniture	11.8	18.6	12.1
Publishing	28.0	18.5	11.8
Petroleum refining	7.4	28.2	10.6
Glass and cement	15.4	24.7	12.6
Motor vehicle	15.7	15.6	12.6
Department store	18.4	21.7	7.2
Restaurant	13.4	23.0	11.3
Amusement	18.1	46.7	4.1
Hospital	10.0	35.8	9.0

13.2 | Significance Tests of the Regression Model and Its Coefficients

Multiple regression models can be developed to fit almost any data set if the level of measurement is adequate and enough data points are available. Once a model has been constructed, it is important to test the model to determine whether it fits the data well and whether the assumptions underlying regression analysis are met. Assessing the adequacy of the regression

model can be done in several ways, including testing the overall significance of the model, studying the significance tests of the regression coefficients, computing the residuals, examining the standard error of the estimate, and observing the coefficient of determination. In this section, we examine significance tests of the regression model and of its coefficients.

Testing the Overall Model

With simple regression, a t test of the slope of the regression line is used to determine whether the population slope of the regression line is different from zero—that is, whether the independent variable contributes significantly in linearly predicting the dependent variable.

The hypotheses for this test, presented in Chapter 12 are

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

For multiple regression, an analogous test makes use of the F statistic. The overall significance of the multiple regression model is tested with the following hypotheses.

$$H_0: \beta_1 = \beta_2 = \beta_3 = \ldots = \beta_k = 0$$
H_a: At least one of the regression coefficients is $\neq 0$.

If we fail to reject the null hypothesis, we are stating that the regression model has no significant predictability for the dependent variable. A rejection of the null hypothesis indicates that at least one of the independent variables is adding significant predictability for y.

This F test of overall significance is often given as a part of the standard multiple regression output from statistical software packages. The output appears as an analysis of variance (ANOVA) table. Shown here is the ANOVA table for the real-estate example taken from the Minitab output in Figure 13.3.

Analysis of Variance

Source	df	SS	MS	F	p
Regression	2	8189.7	4094.9	28.63	0.000
Residual Error	20	2861.0	143.1		
Total	22	11050.7			

The F value is 28.63; because $p = .000$, the F value is significant at $\alpha = .001$. The null hypothesis is rejected, and there is at least one significant predictor of house price in this analysis.

The F value is calculated by the following equation.

$$F = \frac{MS_{reg}}{MS_{err}} = \frac{SS_{reg} / df_{reg}}{SS_{err} / df_{err}} = \frac{SSR/k}{SSE/(n-k-1)}$$

where

MS = mean square
SS = sum of squares
df = degrees of freedom
k = number of independent variables
n = number of observations

Note that in the ANOVA table for the real-estate example, $df_{reg} = 2$. The degrees of freedom formula for regression is the number of regression coefficients plus the regression constant minus 1. The net result is the number of regression coefficients, which equals the number of independent variables, k. The real-estate example uses two independent variables, so $k = 2$. Degrees of freedom error in multiple regression equals the total number of observations minus the number of regression coefficients minus the regression constant, or $n - k - 1$. For the real-estate example, $n = 23$; thus, $df_{err} = 23 - 2 - 1 = 20$.

As shown in Chapter 11, MS = SS/df. The F ratio is formed by dividing MS_{reg} by MS_{err}. In using the F distribution table to determine a critical value against which to test the observed F value, the degrees of freedom numerator is df_{reg} and the degrees of freedom denominator

is df_{err}. The table F value is obtained in the usual manner, as presented in Chapter 11. With $\alpha = .01$ for the real-estate example, the table value is

$$F_{.01,2,20} = 5.85$$

Comparing the observed F of 28.63 to this table value shows that the decision is to reject the null hypothesis. This same conclusion was reached using the p-value method from the computer output.

If a regression model has only one linear independent variable, it is a simple regression model. In that case, the F test for the overall model is the same as the t test for significance of the population slope. The F value displayed in the regression ANOVA table is related to the t test for the slope in the simple regression case as follows.

$$F = t^2$$

In simple regression, the F value and the t value give redundant information about the overall test of the model.

Most analysts who use multiple regression analysis will observe the value of F and its p-value rather early in the process. If F is not significant, then no population regression coefficient is significantly different from zero, and the regression model has no predictability for the dependent variable.

Significance Tests of the Regression Coefficients

In multiple regression, individual significance tests can be computed for each regression coefficient using a t test. Each of these t tests is analogous to the t test for the slope used in Chapter 12 for simple regression analysis. The hypotheses for testing the regression coefficient of each independent variable take the following form.

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

$$H_0: \beta_2 = 0$$
$$H_a: \beta_2 \neq 0$$

$$\vdots$$

$$H_0: \beta_k = 0$$
$$H_a: \beta_k \neq 0$$

Most multiple regression computer packages yield observed t values to test the individual regression coefficients as standard output. Shown here are the t values and their associated probabilities for the real-estate example as displayed with the multiple regression output in Figure 13.3.

Variable	T	P
Square feet	5.63	.000
Age	−2.92	.008

At $\alpha = .05$, the null hypothesis is rejected for both variables because the probabilities (p) associated with their t values are less than .05. If the t ratios for any predictor variables are not significant (fail to reject the null hypothesis), the business analyst might decide to drop that variable(s) from the analysis as a nonsignificant predictor(s). Other factors can enter into this decision. In Chapter 14, we will explore techniques for model building in which some variable sorting is required.

The degrees of freedom for each of these individual tests of regression coefficients are $n - k - 1$. In this particular example because there are $k = 2$ predictor variables, the degrees of freedom are $23 - 2 - 1 = 20$. With $\alpha = .05$ and a two-tailed test, the critical table t value is

$$t_{.025,20} = \pm 2.086$$

Notice from the t ratios shown here that if this critical table t value had been used as the hypothesis test criterion instead of the p-value method, the results would have been the same. Testing the regression coefficients not only gives the analyst some insight into the fit of the regression model, but it also helps in the evaluation of how worthwhile individual independent variables are in predicting y.

13.2 Problems

13.7. Examine the Minitab output shown here for a multiple regression analysis. How many predictors were there in this model? Comment on the overall significance of the regression model. Discuss the t ratios of the variables and their significance.

The regression equation is

$$Y = 4.096 - 5.111X_1 + 2.662X_2 + 1.557X_3 + 1.141X_4 + 1.650X_5 - 1.248X_6 + 0.436X_7 + 0.962X_8 + 1.289X_9$$

Predictor	Coef	Stdev	T	p
Constant	4.096	1.2884	3.24	.006
X_1	-5.111	1.8700	2.73	.011
X_2	2.662	2.0796	1.28	.212
X_3	1.557	1.2811	1.22	.235
X_4	1.141	1.4712	0.78	.445
X_5	1.650	1.4994	1.10	.281
X_6	-1.248	1.2735	0.98	.336
X_7	0.436	0.3617	1.21	.239
X_8	0.962	1.1896	0.81	.426
X_9	1.289	1.9182	0.67	.508

$S = 3.503$ $R\text{-sq} = 40.8\%$ $R\text{-sq(adj.)} = 20.3\%$

Analysis of Variance

Source	df	SS	MS	F	p
Regression	9	219.746	24.416	1.99	.0825
Error	26	319.004	12.269		
Total	35	538.750			

13.8. Displayed here is the Minitab output for a multiple regression analysis. Study the ANOVA table and the t ratios and use these to discuss the strengths of the regression model and the predictors. Does this model appear to fit the data well? From the information here, what recommendations would you make about the predictor variables in the model?

The regression equation is

$$Y = 34.7 + 0.0763X_1 + 0.00026X_2 - 1.12X_3$$

Predictor	Coef	Stdev	T	p
Constant	34.672	5.256	6.60	.000
X_1	0.07629	0.02234	3.41	.005
X_2	0.000259	0.001031	0.25	.805
X_3	-1.1212	0.9955	-1.13	.230

$S = 9.722$ $R\text{-sq} = 51.5\%$ $R\text{-sq(adj)} = 40.4\%$

Analysis of Variance

Source	df	SS	MS	F	p
Regression	3	1306.99	435.66	4.61	.021
Error	13	1228.78	94.52		
Total	16	2535.77			

13.9. Using the data in Problem 13.5, develop a multiple regression model to predict per capita personal consumption by the consumption of paper, fish, and gasoline. Discuss the output and pay particular attention to the F test and the t tests.

13.10. Using the data from Problem 13.6, develop a multiple regression model to predict insider ownership from debt ratio and dividend payout. Comment on the strength of the model and the predictors by examining the ANOVA table and the t tests.

13.11. Develop a multiple regression model to predict y from x_1, x_2, and x_3 using the following data. Discuss the values of F and t.

y	x_1	x_2	x_3
5.3	44	11	401
3.6	24	40	219
5.1	46	13	394
4.9	38	18	362
7.0	61	3	453
6.4	58	5	468
5.2	47	14	386
4.6	36	24	357
2.9	19	52	206
4.0	31	29	301
3.8	24	37	243
3.8	27	36	228
4.8	36	21	342
5.4	50	11	421
5.8	55	9	445

13.12. Use the following data to develop a regression model to predict y from x_1 and x_2. Comment on the output. Develop a regression model to predict y from x_1 only. Compare the results of this model with those of the model using both predictors. What might you conclude by examining the output from both regression models?

y	x_1	x_2
31	11.7	142
46	13.6	134
48	14.8	151
52	12.2	160
60	12.6	151
71	9.5	155
77	9.9	136
84	8.5	127
85	9.8	138
89	7.9	143
104	7.1	149
115	6.6	131
117	6.8	129
122	6.4	137
127	5.4	143

13.13. Study the following Excel multiple regression output. How many predictors are in this model? How many observations? What is the equation of the regression line? Discuss the strength of the model in terms of F. Which predictors, if any, are significant? Why or why not? Comment on the overall effectiveness of the model.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.842
R Square	0.710
Adjusted R Square	0.630
Standard Error	109.430
Observations	15

ANOVA

	df	SS	MS	F	Significance F
Regression	3	321946.82	107315.6	8.96	0.0027
Residual	11	131723.20	11974.8		
Total	14	453670.00			

	Coefficients	Standard Error	t Stat	P-value
Intercept	657.053	167.46	3.92	0.0024
x Variable 1	5.7103	1.792	3.19	0.0087
x Variable 2	−0.4169	0.322	−1.29	0.2222
x Variable 3	−3.4715	1.443	−2.41	0.0349

13.3 | Residuals, Standard Error of the Estimate, and R^2

Three more statistical tools for examining the strength of a regression model are the residuals, the standard error of the estimate, and the coefficient of multiple determination.

Residuals

The **residual**, or error, of the regression model is *the difference between the y value and the predicted value, $\hat{y}$.*

$$\text{Residual} = y - \hat{y}$$

The residuals for a multiple regression model are solved for in the same manner as they are with simple regression. First, a predicted value, $\hat{y}$, is determined by entering the value for each independent variable for a given set of observations into the multiple regression equation and solving for $\hat{y}$. Next, the value of $y - \hat{y}$ is computed for each set of observations. Shown here are the calculations for the residuals of the first set of observations from Table 13.1. The predicted value of y for $x_1 = 1605$ and $x_2 = 35$ is

$$\hat{y} = 57.4 + .0177(1605) - .666(35) = 62.499$$

Actual value of $y = 63.0$

Residual $= y - \hat{y} = 63.0 - 62.499 = 0.501$

All residuals for the real-estate data and the regression model displayed in Table 13.1 and Figure 13.3 are displayed in **Table 13.2.**

An examination of the residuals in Table 13.2 can reveal some information about the fit of the real-estate regression model. The business analyst can observe the residuals and decide whether the errors are small enough to support the accuracy of the model. The house price figures are in units of $1,000. Two of the 23 residuals are more than 20.00, or more than $20,000 off in their prediction. On the other hand, two residuals are less than 1, or $1,000 off in their prediction.

Residuals are also helpful in locating outliers. **Outliers** are *data points that are apart, or far, from the mainstream of the other data.* They are sometimes data points that were mistakenly recorded or measured. Because every data point influences the regression model, outliers can exert an overly important influence on the model based on their distance from other points. In examining the residuals in Table 13.2 for outliers, the eighth residual listed is −27.699. This error indicates that the regression model was not nearly as successful in predicting house price

TABLE 13.2 Residuals for the Real-Estate Regression Model

y	$\hat{y}$	$y - \hat{y}$
63.0	62.499	.501
65.1	71.485	−6.385
69.9	71.568	−1.668
76.8	78.639	−1.839
73.9	74.097	−.197
77.9	75.653	2.247
74.9	83.012	−8.112
78.0	105.699	−27.699
79.0	68.523	10.477
83.4	81.139	2.261
79.5	88.282	−8.782
83.9	91.335	−7.435
79.7	85.618	−5.918
84.5	95.391	−10.891
96.0	85.456	10.544
109.5	102.774	6.726
102.5	97.665	4.835
121.0	107.183	13.817
104.9	109.352	−4.452
128.0	111.230	16.770
129.0	105.061	23.939
117.9	134.415	−16.515
140.0	132.430	7.570

on this particular house as it was with others (an error of more than $27,000). For whatever reason, this data point stands somewhat apart from other data points and may be considered an outlier.

Residuals are also useful in testing the assumptions underlying regression analysis. **Figure 13.4** displays Minitab diagnostic techniques for the real-estate example. In the top right is a graph of the residuals. Notice that residual variance seems to increase in the right half of the plot, indicating potential heteroscedasticity. As discussed in Chapter 12, one of the assumptions underlying regression analysis is that the error terms have homoscedasticity or homogeneous variance. That assumption might be violated in this example. The normal plot of residuals is nearly a straight line, indicating that the assumption of normally distributed error terms probably has not been violated.

SSE and Standard Error of the Estimate

One of the properties of a regression model is that the residuals sum to zero. As pointed out in Chapter 12, this property precludes the possibility of computing an "average" residual as a single measure of error. In an effort to compute a single statistic that can represent the error in a regression analysis, the zero-sum property can be overcome by *squaring the residuals and then summing the squares.* Such an operation produces the **sum of squares of error (SSE).**

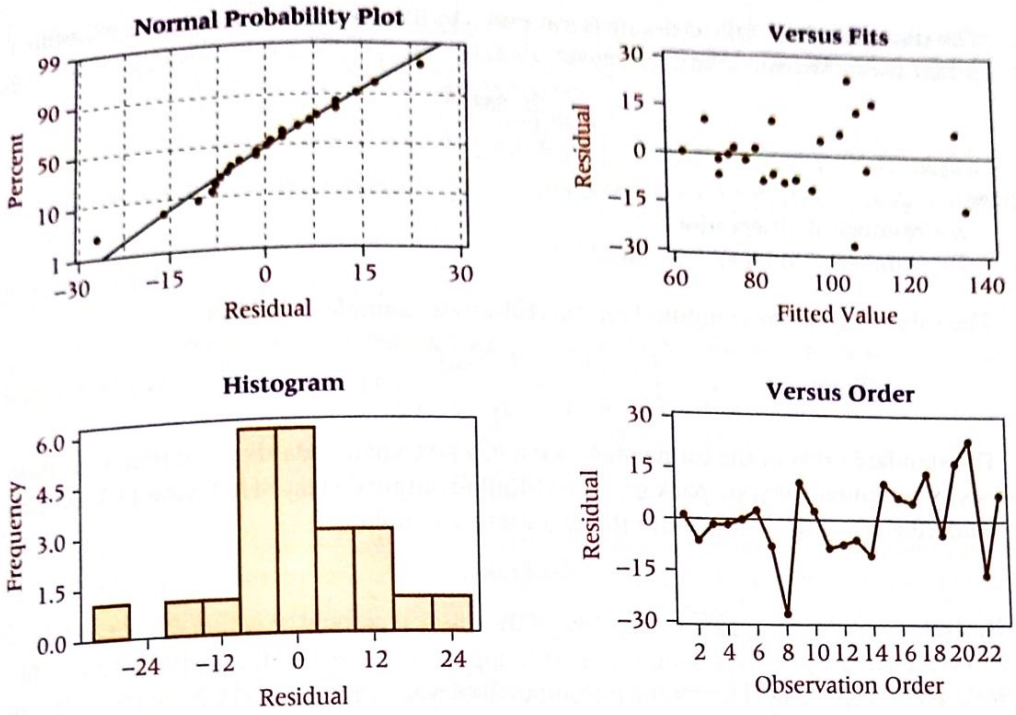


FIGURE 13.4 **Minitab Residual Diagnosis for the Real-Estate Example**

The formula for computing the sum of squares of error (SSE) for multiple regression is the same as it is for simple regression.

$$\text{SSE} = \Sigma(y - \hat{y})^2$$

For the real-estate example, SSE can be computed by squaring and summing the residuals displayed in Table 13.2.

$$\begin{aligned}\text{SSE} = [&(.501)^2 (-6.385)^2 + (-1.668)^2 + (-1.839)^2 \\ &+ (-.197)^2 + (2.247)^2 + (-8.112)^2 + (-27.699)^2 \\ &+ (10.477)^2 + (2.261)^2 + (-8.782)^2 + (-7.435)^2 \\ &+ (-5.918)^2 + (-10.891)^2 + (10.544)^2 + (6.726)^2 \\ &+ (4.835)^2 + (13.817)^2 + (-4.452)^2 + (16.770)^2 \\ &+ (23.939)^2 + (-16.515)^2 + (7.570)^2] \\ = 2861.0\end{aligned}$$

SSE can also be obtained directly from the multiple regression computer output by selecting the value of SS (sum of squares) listed beside error. Shown here is the ANOVA portion of the output displayed in Figure 13.3, which is the result of a multiple regression analysis model developed to predict house prices. Note that the SS for error shown in the ANOVA table equals the value of $\Sigma(y - \hat{y})^2$ just computed (2861.0).

SSE

Analysis of Variance

Source	df	SS	MS	F	P
Regression	2	8189.7	4094.9	28.63	.000
Error	20	2861.0	143.1		
Total	22	11050.7			

SSE has limited usage as a measure of error. However, it is a tool used to solve for other, more useful measures. One of those is the **standard error of the estimate, s_e,** which is essentially *the standard deviation of residuals (error) for the regression model.* As explained in Chapter 12, an assumption underlying regression analysis is that the error terms are approximately normally distributed with a mean of zero. With this information and by the empirical rule, approximately 68% of the residuals should be within $\pm 1s_e$ and 95% should be within $\pm 2s_e$. This property makes the standard error of the estimate a useful tool in estimating how accurately a regression model is fitting the data.

The standard error of the estimate is computed by dividing SSE by the degrees of freedom of error for the model and taking the square root.

$$s_e = \sqrt{\frac{\text{SSE}}{n-k-1}}$$

where

n = number of observations
k = number of independent variables

The value of s_e can be computed for the real-estate example as follows.

$$s_e = \sqrt{\frac{\text{SSE}}{n-k-1}} = \sqrt{\frac{2861}{23-2-1}} = 11.96$$

The standard error of the estimate, s_e, is usually given as standard output from regression analysis by computer software packages. The Minitab output displayed in Figure 13.3 contains the standard error of the estimate for the real-estate example.

$$S = 11.96$$

By the empirical rule, approximately 68% of the residuals should be within $\pm 1s_e = \pm 1(11.96) = \pm 11.96$. Because house prices are in units of $1,000, approximately 68% of the predictions are within ±11.96($1,000), or ±$11,960. Examining the output displayed in Table 13.2, 18/23, or about 78%, of the residuals are within this span. According to the empirical rule, approximately 95% of the residuals should be within $\pm 2s_e$, or ±2(11.96) = ±23.92. Further examination of the residual values in Table 13.2 shows that 21 of 23, or 91%, fall within this range. The business analyst can study the standard error of the estimate and these empirical rule–related ranges and decide whether the error of the regression model is sufficiently small to justify further use of the model.

Coefficient of Multiple Determination (R^2)

The **coefficient of multiple determination (R^2)** is analogous to the coefficient of determination (r^2) discussed in Chapter 12. R^2 represents *the proportion of variation of the dependent variable, y, accounted for by the independent variables in the regression model.* As with r^2, the range of possible values for R^2 is from 0 to 1. An R^2 of 0 indicates no relationship between the predictor variables in the model and y. An R^2 of 1 indicates that 100% of the variability of y has been accounted for by the predictors. Of course, it is desirable for R^2 to be high, indicating the strong predictability of a regression model. The coefficient of multiple determination can be calculated by the following formula.

$$R^2 = \frac{\text{SSR}}{\text{SS}_{yy}} = 1 - \frac{\text{SSE}}{\text{SS}_{yy}}$$

R^2 can be calculated in the real-estate example by using the sum of squares regression (SSR), the sum of squares error (SSE), and sum of squares total (SS_{yy}) from the ANOVA portion of Figure 13.3.

Analysis of Variance

Source	df	SS	MS	F	P
Regression	2	8189.7	4094.9	28.63	.000
Error	20	2861.0	143.1		
Total	22	11050.7			

SSR → 8189.7; SSE → 2861.0; SS_{yy} → 11050.7

$$R^2 = \frac{\text{SSR}}{\text{SS}_{yy}} = \frac{8189.7}{11050.7} = .741$$

or

$$R^2 = 1 - \frac{\text{SSE}}{\text{SS}_{yy}} = 1 - \frac{2861.0}{11050.7} = .741$$

In addition, virtually all statistical software packages print out R^2 as standard output with multiple regression analysis. A reexamination of Figure 13.3 reveals that R^2 is given as

$$R\text{-sq} = 74.1\%$$

This result indicates that a relatively high proportion of the variation of the dependent variable, house price, is accounted for by the independent variables in this regression model.

Adjusted R^2

As additional independent variables are added to a regression model, the value of R^2 cannot decrease, and in most cases it will increase. In the formulas for determining R^2,

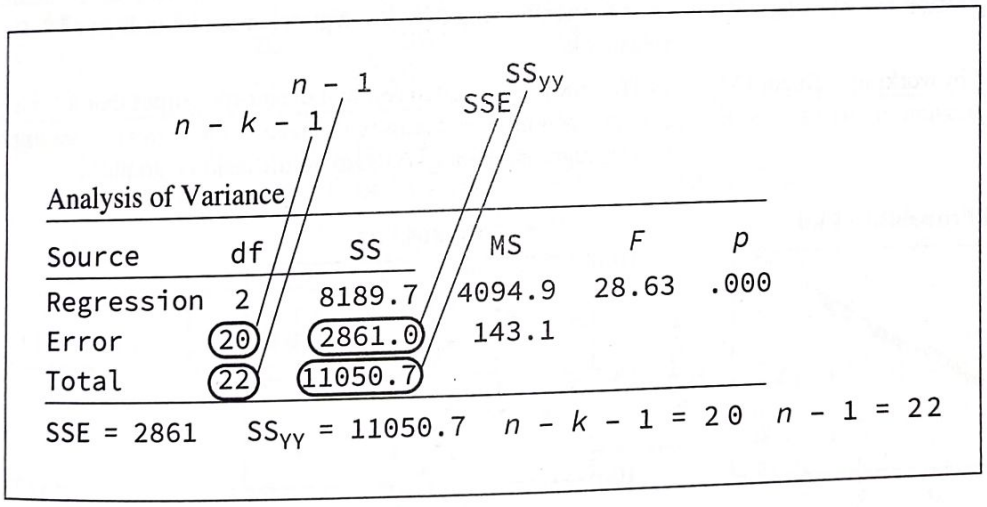
$$R^2 = \frac{\text{SSR}}{\text{SS}_{yy}} = 1 - \frac{\text{SSE}}{\text{SS}_{yy}}$$

the value of SS_{yy} for a given set of observations will remain the same as independent variables are added to the regression analysis because SS_{yy} is the sum of squares for the dependent variable. Because additional independent variables are likely to increase SSR at least by some amount, the value of R^2 will probably increase for any additional independent variables.

However, sometimes additional independent variables add no *significant* information to the regression model, yet R^2 increases. R^2 therefore may yield an inflated figure. Statisticians have developed an **adjusted R^2** *to take into consideration both the additional information each new independent variable brings to the regression model and the changed degrees of freedom of regression.* Many standard statistical computer packages now compute and report adjusted R^2 as part of the output. The formula for computing adjusted R^2 is

$$\text{Adjusted } R^2 = 1 - \frac{\text{SSE}/(n-k-1)}{\text{SS}_{yy}/(n-1)}$$

The value of adjusted R^2 for the real estate example can be solved by using information from the ANOVA portion of the computer output in Figure 13.3.



$$\text{Adj. } R^2 = 1 - \left[\frac{2861/20}{11050.7/22}\right] = 1 - .285 = .715$$

The standard Minitab regression output in Figure 13.3 contains the value of the adjusted R^2 already computed. For the real estate example, this value is shown as

$$R\text{-sq(adj.)} = 71.5\%$$

A comparison of R^2 (.741) with the adjusted R^2 (.715) for this example shows that the adjusted R^2 reduces the overall proportion of variation of the dependent variable accounted for by the independent variables by a factor of .026, or 2.6%. The gap between the R^2 and adjusted R^2 tends to increase as nonsignificant independent variables are added to the regression model. As n increases, the difference between R^2 and adjusted R^2 becomes less.

Thinking Critically About Statistics in Business Today

Assessing Property Values Using Multiple Regression

According to county assessor sources, a Colorado state statute requires that all county assessors in the state value residential real property solely by a market approach. Furthermore, in the statute, it is stated that such a market approach will be based on a representative body of sales sufficient to set a pattern. No specifics on market analysis methods are given in the statutes, but there are several commonly used methods, including multiple regression. In the multiple regression approach, an attempt is made to develop a model to predict recent sales (dependent variable) by such property characteristics (independent variables) as location, lot size, building square feet, construction quality, property type (single family, condominium, townhouse), garage size, basement type, and other features. One county website states that "regression does not require strict similarity between property sales because it estimates the value contribution (coefficient) for each attribute. . . ."

In producing a sound multiple regression model to predict property values, several models are developed, refined, and compared using the typical indicators of a good fit, such as R^2, standard error of the estimate, F test for the overall model, and p-values associated with the t tests of significance for predictors. The final multiple regression model is then used in the estimation of property values by the appraiser, who enters into the regression model (equation) the specific measure of each independent (predictor) variable for a given property, resulting in a predicted appraised property value for tax purposes. The models are updated for currentness and based on data that are never more than two years old.

Things to Ponder

1. How does the multiple regression method improve the validity of property value assessments over typical standard methods? Do you think it is more fair and equitable? If so, why?
2. Can you think of other similar possible applications of multiple regression in business?

13.3 Problems

13.14. Study the Minitab output shown in Problem 13.7. Comment on the overall strength of the regression model in light of S, R^2, and adjusted R^2.

13.15. Study the Minitab output shown in Problem 13.8. Comment on the overall strength of the regression model in light of S, R^2, and adjusted R^2.

13.16. Using the regression output obtained by working Problem 13.5, comment on the overall strength of the regression model using S, R^2, and adjusted R^2.

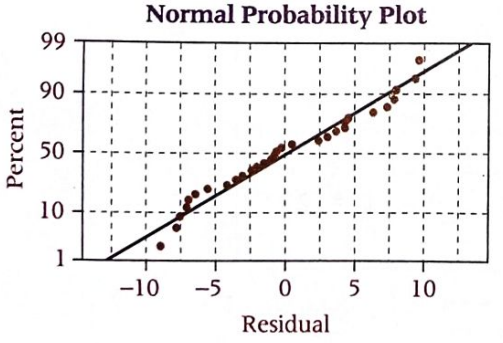
13.17. Using the regression output obtained by working Problem 13.6, comment on the overall strength of the regression model using S, R^2, and adjusted R^2.

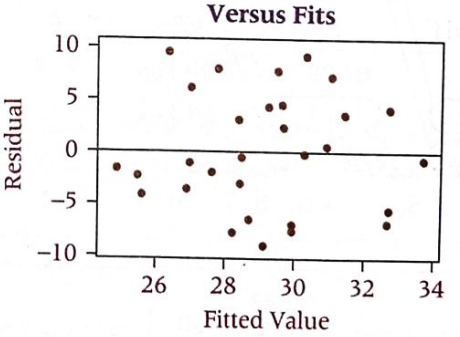
13.18. Using the regression output obtained by working Problem 13.11, comment on the overall strength of the regression model using S, R^2, and adjusted R^2.

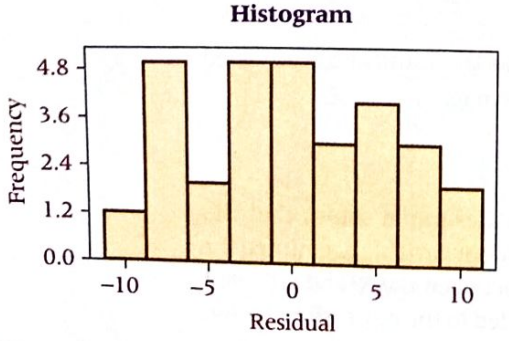
13.19. Using the regression output obtained by working Problem 13.12, comment on the overall strength of the regression model using S, R^2, and adjusted R^2.

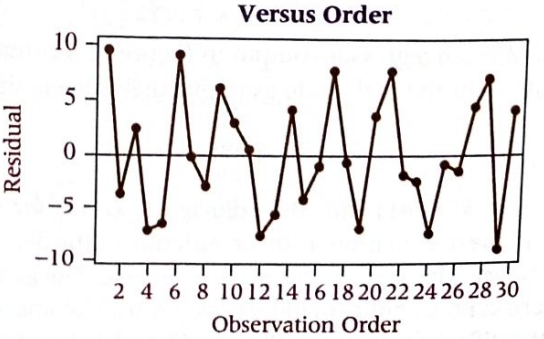
13.20. Study the Excel output shown in Problem 13.13. Comment on the overall strength of the regression model in light of S, R^2, and adjusted R^2.

13.21. Study the Minitab residual diagnostic output that follows. Discuss any potential problems with meeting the regression assumptions for this regression analysis based on the residual graphics.









13.4 | Interpreting Multiple Regression Computer Output

A Reexamination of the Multiple Regression Output

Figure 13.5 shows again the Minitab multiple regression output for the real estate example. Many of the concepts discussed thus far in the chapter are highlighted. Note the following items:

1. The equation of the regression model
2. The ANOVA table with the F value for the overall test of the model
3. The t ratios, which test the significance of the regression coefficients
4. The value of SSE
5. The value of s_e
6. The value of R^2
7. The value of adjusted R^2

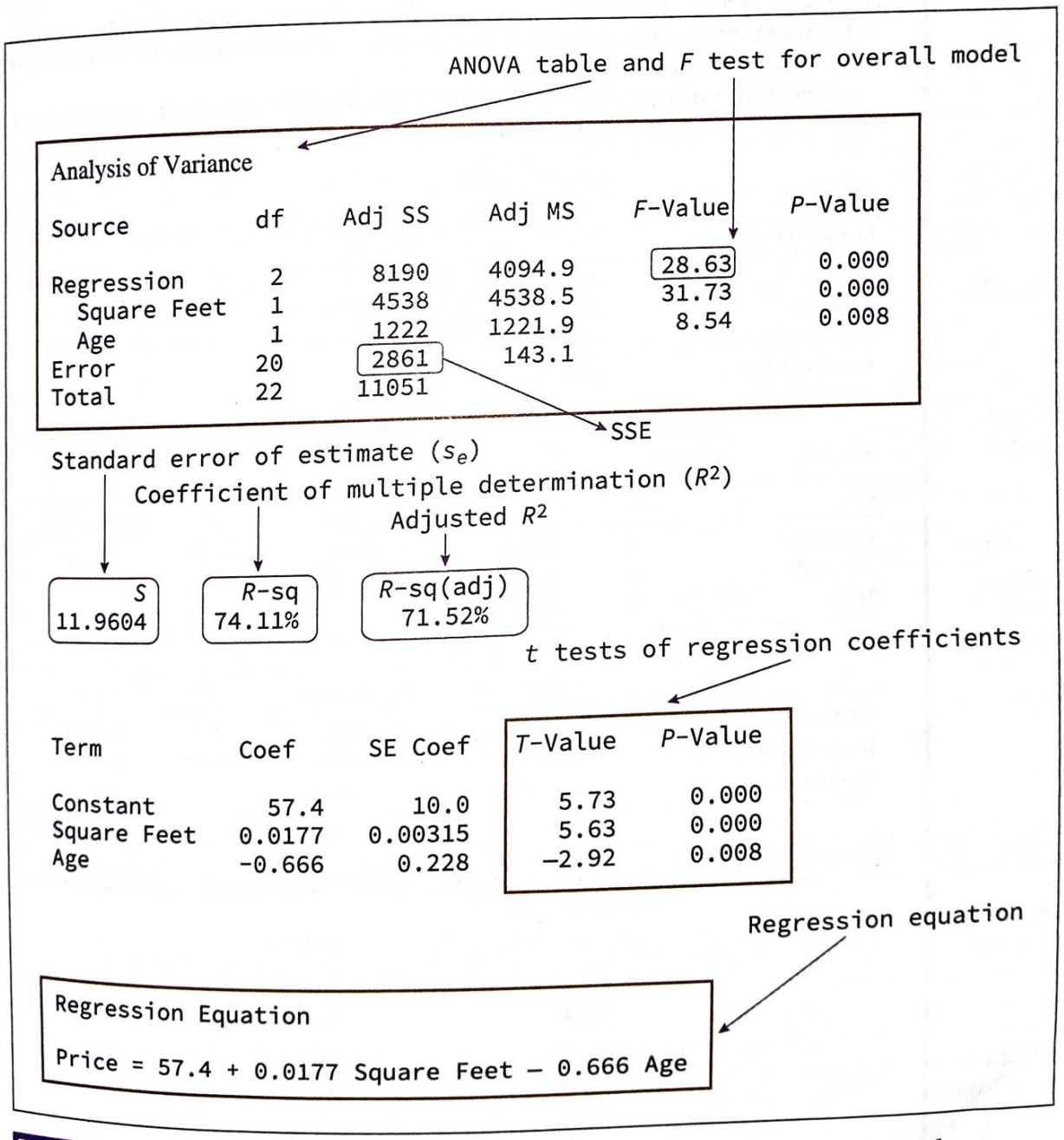


FIGURE 13.5 **Annotated Version of the Minitab Output of Regression for the Real-Estate Example**

DEMONSTRATION PROBLEM 13.2

Discuss the Excel multiple regression output for Demonstration Problem 13.1. Comment on the F test for the overall significance of the model, the t tests of the regression coefficients, and the values of s_e, R^2, and adjusted R^2.

Solution This multiple regression analysis was done to predict the prime interest rate using the predictors of unemployment and personal saving. The equation of the regression model was presented in the solution of Demonstration Problem 13.1. Shown here is the complete multiple regression output from the Excel analysis of the data.

The value of F for this problem is 12.29, with a p-value of .0012, which is significant at $\alpha = .01$. On the basis of this information, the null hypothesis is rejected for the overall test of significance. At least one of the predictor variables is **statistically significant**, and there is significant predictability of the prime interest rate by this model.

An examination of the t ratios reveals that unemployment rate is a significant predictor at $\alpha = .001$ ($t = -4.93$ with a p-value of .0003) and that personal savings is a significant predictor at $\alpha = .01$ ($t = 3.31$ with a p-value of .0062). The positive signs on the regression coefficient and the t value for personal savings indicate that as personal savings increase, the prime interest rate tends to get higher. On the other hand, the negative signs on the regression coefficient and the t value for unemployment rates indicate that as the unemployment rate increases, the prime interest rate tends to decrease.

The standard error of the estimate is $s_e = 1.496$, indicating that approximately 68% of the residuals are within ±1.496. An examination of the Excel-produced residuals shows that actually 11 out of 15, or 73.3%, fall in this interval. Approximately 95% of the residuals should be within ±2(1.496) = ±2.992, and an examination of the Excel-produced residuals shows that 14 out of 15, or 93.3%, of the residuals are within this interval.

R^2 for this regression analysis is .672, or 67.2%. The adjusted R^2 is .617, indicating that there is some inflation in the R^2 value. Overall, there is modest predictability in this model.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.820
R Square	0.672
Adjusted R Square	0.617
Standard Error	1.496
Observations	15

ANOVA

	df	SS	MS	F	Significance F
Regression	2	54.9835	27.4917	12.29	0.0012
Residual	12	26.8537	2.2378		
Total	14	81.8372			

	Coefficients	Standard Error	*t* Stat	*P*-value
Intercept	13.5786	1.728	7.86	0.0000
Unemployment Rates	−1.6622	0.337	−4.93	0.0003
Personal Savings	0.6586	0.199	3.31	0.0062

RESIDUAL OUTPUT

Observation	Predicted Prime Interest Rate	Residuals
1	7.3441	0.9859
2	9.2446	0.0754
3	8.8808	1.1292
4	6.1837	0.0663
5	6.6008	0.5492

6	7.2374	1.0326
7	8.9309	−0.5809
8	8.4448	0.7852
9	5.5187	−0.8487
10	5.8198	−1.4798
11	6.3937	1.5663
12	5.1236	−0.0336
13	1.4418	1.8082
14	5.1206	−1.8706
15	6.4346	−3.1846

13.4 Problems

13.22. Study the Minitab regression output that follows. How many predictors are there? What is the equation of the regression model? Using the key statistics discussed in this chapter, discuss the strength of the model and the predictors.

Regression Analysis: Y versus X_1, X_2, X_3, X_4

Analysis of Variance

Source	df	Adj SS	Adj MS	F-Value	P-Value
Regression	4	18088.5	4522.13	55.52	0.000
Error	55	4479.7	81.45		
Total	59	22568.2			

S	R-sq	R-sq(adj)
9.025	80.2%	78.7%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	-55.93	24.22	-2.31	0.025
X_1	0.01049	0.02100	0.50	0.619
X_2	-0.10720	0.03503	-3.06	0.003
X_3	0.57922	0.07633	7.59	0.000
X_4	-0.8695	0.1498	-5.81	0.000

Regression Equation

$$Y = -55.9 + 0.0105\,X_1 - 0.107\,X_2 + 0579\,X_3 - 0.870\,X_4$$

13.23. Study the Excel regression output that follows. How many predictors are there? What is the equation of the regression model? Using the key statistics discussed in this chapter, discuss the strength of the model and its predictors.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.814
R Square	0.663
Adjusted R Square	0.636
Standard Error	51.761
Observations	28

ANOVA

	df	SS	MS	F	Significance F
Regression	2	131567.02	65783.51	24.55	0.0000013
Residual	25	66979.65	2679.19		
Total	27	198546.68			

	Coefficients	Standard Error	*t* Stat	*P*-value
Intercept	203.3937	67.518	3.01	0.0059
X_1	1.1151	0.528	2.11	0.0448
X_2	−2.2115	0.567	−3.90	0.0006

DECISION DILEMMA SOLVED Are You Going to Hate Your New Job?

In the Decision Dilemma, several variables are considered in attempting to determine whether a person will like his or her new job. Four predictor (independent) variables are given with the data set: relationship with supervisor, overall quality of work environment, total hours worked per week, and opportunities for advancement. Other possible variables might include openness of work culture, amount of pressure, how the interviewee is treated during the interview, availability of flexible scheduling, size of office, amount of time allotted for lunch, availability of management, interesting work, and many others.

Using the data that are given, a multiple regression model can be developed to predict job satisfaction from the four independent variables. Such an analysis allows the business analyst to study the entire data set in one model rather than constructing four different simple regression models, one for each independent variable. In the multiple regression model, job satisfaction is the dependent variable. There are 19 observations. The Excel regression output for this problem follows.

The test for overall significance of the model produced an F of 87.79 with a p-value of .000000001 (significant at $\alpha = .00000001$). The R^2 of .962 and adjusted R^2 of .951 indicate very strong predictability in the model. The standard error of the estimate, 5.141, can be viewed in light of the job satisfaction values that ranged from 10 to 95 and the residuals, which are not shown here. Fourteen of the 19 residuals (73.7%) are within the standard error of the estimate. Examining the t statistics and their associated p-values reveals that two independent variables, relationship with supervisor ($t = 5.33$, p-value = .0001) and overall quality of work environment ($t = 2.73$, p-value = .0162) are significant predictors at $\alpha = .05$. Judging by their large p-values, it appears that total hours worked per week and opportunities for advancement are not good predictors of job satisfaction.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.981
R Square	0.962
Adjusted R Square	0.951
Standard Error	5.141
Observations	19

ANOVA

	df	SS	MS	F	Significance F
Regression	4	9282.569	2320.642	87.79	0.000000001
Residual	14	370.062	26.433		
Total	18	9652.632			

	Coefficients	Standard Error	t Stat	P-value
Intercept	−1.469	8.116	−0.18	0.8590
Relationship with Supervisor	1.391	0.261	5.33	0.0001
Overall Quality of Work Environment	0.317	0.116	2.73	0.0162
Total Hours Worked per Week	0.043	0.121	0.36	0.7263
Opportunities for Advancement	−0.094	0.102	−0.92	0.3711

Ethical Considerations

Multiple regression analysis can be used either intentionally or unintentionally in questionable or unethical ways. When degrees of freedom are small, an inflated value of R^2 can be obtained, leading to overenthusiastic expectations about the predictability of a regression model. To prevent this type of reliance, a business analyst should take into account the nature of the data, the variables, and the value of the adjusted R^2.

Another misleading aspect of multiple regression can be the tendency of analysts to assume cause-and-effect relationships between the dependent variable and predictors. Just because independent variables produce a significant R^2 does not necessarily mean those variables are causing the deviation of the y values. Indeed, some other force not in the model may be driving both the independent variables and the dependent variable over the range of values being studied.

Some people use the estimates of the regression coefficients to compare the worth of the predictor variables; the larger the coefficient, the greater is its worth. At least two problems can be found in 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 (discussed in Chapter 14) 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. Analysts who ignore this problem are at risk of presenting spurious results.

Another danger in using regression analysis is in the extrapolation of the model to values beyond the range of values used to derive the model. A regression model that fits data within a given range does not necessarily fit data outside that range. One of the uses of regression analysis is in the area of forecasting. Users need to be aware that what has occurred in the past is not guaranteed to continue to occur in the future. Unscrupulous and sometimes even well-intentioned business decision makers can use regression models to project conclusions about the future that have little or no basis. The receiver of such messages should be cautioned that regression models may lack validity outside the range of values in which the models were developed.

Summary

Multiple regression analysis is a statistical tool in which a mathematical model is developed in an attempt to predict a dependent variable by two or more independent variables or in which at least one predictor is nonlinear. Because doing multiple regression analysis by hand is extremely tedious and time-consuming, it is almost always done on a computer.

The standard output from a multiple regression analysis is similar to that of simple regression analysis. A regression equation is produced with a constant that is analogous to the y-intercept in simple regression and with estimates of the regression coefficients that are analogous to the estimate of the slope in simple regression. An F test for the overall model is computed to determine whether at least one of the regression coefficients is significantly different from zero. This F value is usually displayed in an ANOVA table, which is part of the regression output. The ANOVA table also contains the sum of squares of error and sum of squares of regression, which are used to compute other statistics in the model.

Most multiple regression computer output contains t values, which are used to determine the significance of the regression coefficients. Using these t values, business analysts can make decisions about including or excluding variables from the model.

Residuals, standard error of the estimate, and R^2 are also standard computer regression output with multiple regression. The coefficient of determination for simple regression models is denoted r^2, whereas for multiple regression it is R^2. The interpretation of residuals, standard error of the estimate, and R^2 in multiple regression is similar to that in simple regression. Because R^2 can be inflated with nonsignificant variables in the mix, an adjusted R^2 is often computed. Unlike R^2, adjusted R^2 takes into account the degrees of freedom and the number of observations.

Key Terms

adjusted R^2
coefficient of multiple determination (R^2)
dependent variable
independent variable
least squares analysis
multiple regression
outliers
partial regression coefficient
residual
response plane
response surface
response variable
standard error of the estimate (s_e)
sum of squares of error (SSE)

Formulas

The F value

$$F = \frac{MS_{reg}}{MS_{err}} = \frac{SS_{reg}/df_{reg}}{SS_{err}/df_{err}} = \frac{SSR/k}{SSE/(n-k-1)}$$

Sum of squares of error

$$SSE = \Sigma(y - \hat{y})^2$$

Standard error of the estimate

$$s_e = \sqrt{\frac{SSE}{n-k-1}}$$

Coefficient of multiple determination

$$R^2 = \frac{SSR}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}}$$

Adjusted R^2

$$\text{Adjusted } R^2 = 1 - \frac{SSE/(n-k-1)}{SS_{yy}/(n-1)}$$

Supplementary Problems

Calculating the Statistics

13.24. Use the following data to develop a multiple regression model to predict y from x_1 and x_2. Discuss the output, including comments about the overall strength of the model, the significance of the regression coefficients, and other indicators of model fit.

y	x_1	x_2
198	29	1.64
214	71	2.81
211	54	2.22
219	73	2.70
184	67	1.57
167	32	1.63
201	47	1.99
204	43	2.14
190	60	2.04
222	32	2.93
197	34	2.15

13.25. 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.

y	x_1	x_2	x_3
14	51	16.4	56
17	48	17.1	64
29	29	18.2	53
32	36	17.9	41
54	40	16.5	60
86	27	17.1	55
117	14	17.8	71
120	17	18.2	48
194	16	16.9	60
203	9	18.0	77
217	14	18.9	90
235	11	18.5	67

Testing Your Understanding

13.26. 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 multiple regression 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

13.27. 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. Comment on the regression model and its strengths and its weaknesses.

	Commercial Ship	
Employment	Naval Vessels	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

13.28. 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 multiple regression to develop a model that attempts to predict all commodities by the other four variables. 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

13.29. The U.S. Department of Agriculture publishes data annually on various selected farm products. Shown here are the unit production figures (in millions of bushels) for three farm products for 10 years during a 20-year period. Use these data and multiple regression analysis to predict corn production by the production of soybeans and wheat. Comment on the results.

Corn	Soybeans	Wheat
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

13.30. 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

13.31. Shown here are the data for y and three predictors, x_1, x_2, and x_3. A multiple regression analysis has been done on these data; the Minitab results are given. Comment on the outcome of the 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

Regression Analysis: Y versus X_1, X_2, X_3

Analysis of Variance

Source	df	Adj SS	Adj MS	F-Value	P-Value
Regression	3	103.185	34.3950	47.57	0.000
X_1	1	6.857	6.8570	9.48	0.013
X_2	1	9.975	9.9752	13.80	0.005
X_3	1	19.685	19.6849	27.23	0.001
Error	9	6.507	0.7230		
Total	12	109.692			

S	R-sq	R-sq(adj)	R-sq(pred)
0.850311	94.07%	92.09%	88.90%

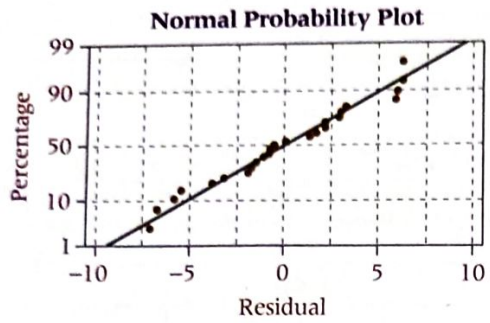
Coefficients

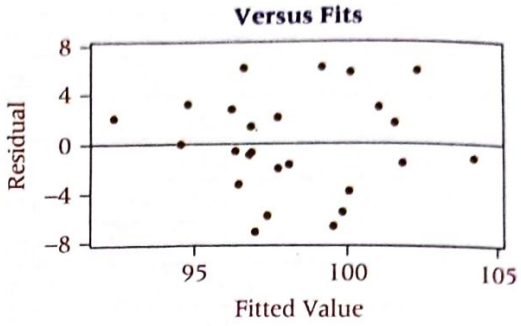
Term	Coef	SE Coef	T-Value	P-Value
Constant	87.89	3.45	25.51	0.000
X_1	-0.2561	0.0832	-3.08	0.013
X_2	-2.714	0.731	-3.71	0.005
X_3	0.0706	0.0135	5.22	0.001

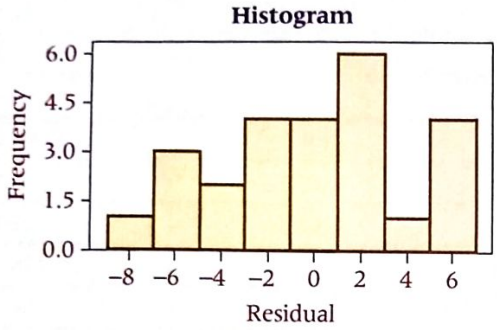
Regression Equation

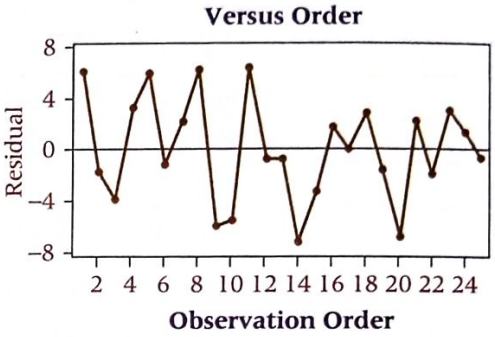
$$Y = 87.89 - 0.2561\ X_1 - 2.714\ X_2 + 0.0706\ X_3$$

13.32. Minitab residual diagnostic output from the multiple regression analysis for the data given in Problem 13.30 follows. Discuss any potential problems with meeting the regression assumptions for this regression analysis based on the residual graphics.









Exploring the Databases with Business Analytics

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, and End-of-Year Inventories. Discuss the results of the analysis.

2. 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? Which variables seem to be the best predictors?

3. Using the International Stock Market database, develop a multiple regression model to predict the Nikkei by the DJIA, the Nasdaq, the S&P 500, the Hang Seng, the FTSE 100, and the IPC. Discuss the outcome, including the model, the strength of the model, and the strength of the predictors.

4. Develop a multiple regression model to predict Annual Food Spending by Annual Household Income and Non-Mortgage Household Debt using the Consumer Food database. How strong is the model? Which of the two predictors seems to be stronger? Why?

Chapter Case

Starbucks Introduces Debit Card

Starbucks is a resounding restaurant success story. Beginning with its first coffee house in 1971, Starbucks has grown to more than 28,218 stores. Opening up its first international outlet in the mid-1990s, Starbucks now operates in more than 67 countries outside of North America. Besides selling beverages, pastries, confections, and coffee-related accessories and equipment at its retail outlets, Starbucks also purchases and roasts high-quality coffee beans in several locations. The company's objective is to become the most recognized and respected brand in the world. Starbucks maintains a strong environmental orientation and is committed to taking a leadership position environmentally. In addition, the company has won awards for corporate social responsibility through its community-building programs, its strong commitment to its origins (coffee producers, family, community), and the Starbucks Foundation, which is dedicated to creating hope, discovery, and opportunity in the communities where Starbucks resides.

In November 2001, Starbucks launched its prepaid (debit) Starbucks Card. The card, which holds between $5 and $500, can be used at virtually any Starbucks location. The card was so popular when it was first released that many stores ran out. By mid-2002, Starbucks had activated more than 5 million of these cards. The Starbucks Card has surpassed the $25 billion mark for total activations and reloads since its introduction. As customers "reload" the cards, it appears they are placing more money on them than the initial value of the card.

Starbucks has gone on to promote their Starbucks Card as a flexible marketing tool that can be used by individuals as a gift of thanks and appreciation for friendship or service and can be used by companies to reward loyal customers and as an incentive to employees.

Discussion

1. Starbucks enjoyed considerable success with its debit cards, which they sell for $5 to $500. Suppose Starbucks management wants to study the reasons why some people purchase debit cards with higher prepaid amounts than do other people. Suppose a study of 25 randomly selected prepaid card purchasers is taken. Respondents are asked the amount of the prepaid card, the customer's age, the number of days per month the customer makes a purchase at Starbucks, the number of cups of coffee the customer drinks per day, and the customer's income. The data follow. Using these data, develop a multiple regression model to study how well the amount of the prepaid card can be predicted by the other variables and which variables seem to be more promising in doing the prediction. What sales implications might be evident from this analysis?

Amount of Prepaid Card ($)	Age	Days per Month at Starbucks	Cups of Coffee per Day	Income ($1,000s)
5	25	4	1	20
25	30	12	5	35
10	27	10	4	30
5	42	8	5	30
15	29	11	8	25
50	25	12	5	60
10	50	8	3	30
15	45	6	5	35
5	32	16	7	25
5	23	10	1	20
20	40	18	5	40
35	35	12	3	40
40	28	10	3	50
15	33	12	2	30
200	40	15	5	80
15	37	3	1	30
40	51	10	8	35
5	20	8	4	25
30	26	15	5	35
100	38	19	10	45
30	27	12	3	35
25	29	14	6	35
25	34	10	4	45
50	30	6	3	55
15	22	8	5	30

2. Suppose marketing wants to be able to describe frequent visitors to a Starbucks store. Using the same data set already provided, develop a multiple regression model to predict Days per month at Starbucks by Age, Income, and Number of cups of coffee per day. How strong is the model? Which particular independent variables seem to have more promise in predicting how many days per month a customer visits Starbucks? What marketing implications might be evident from this analysis?

3. Over the past decade or so, Starbucks has grown quite rapidly. As they add stores and increase the number of drinks, their sales revenues increase. In reflecting about this growth, think about some other variables that might be related to the increase in Starbucks sales revenues. Some data for the past seven years on the number of Starbucks stores (worldwide), approximate sales revenue (in $ millions), number of different drinks sold, and average weekly earnings of U.S. production workers are given here. Most figures are approximate. Develop a multiple regression model to predict sales revenue by number of drinks sold, number of stores, and average weekly earnings. How strong is the model? What are the key predictors, if any? How might this analysis help Starbucks management in attempting to determine what drives sales revenues?

Sales Year	Revenue ($ millions)	Number of Stores	Number of Drinks	Average Weekly Earnings ($)
1	400	676	15	386
2	700	1015	15	394
3	1000	1412	18	407
4	1350	1886	22	425
5	1650	2135	27	442
6	2200	3300	27	457
7	2600	4709	30	474

Source: Adapted from Shirley Leung, "Starbucks May Indeed Be a Robust Staple," *The Wall Street Journal*, July 26, 2002, B4; James Peters, "Starbucks' Growth Still Hot; Gift Card Jolts Chain's Sales," *Nation's Restaurant News*, February 11, 2002, 1–2. https://www.starbucks.com/card, Starbucks' website (2019) at https://www.starbucks.com/about-us/company-information

Big Data Case

Further exploring the American Hospital Association (AHA) database of over 2000 hospitals, perform the following analyses:

1. Develop a multiple regression model using all-hospital data to predict the number of outpatient visits by number of beds, number of admissions, and census. Discuss the strength of the model including R^2 and overall F. Are any of the predictor variables significant? If so, why, both statistically and logically?

2. Using the Small Hospitals Only AHA sub-database, conduct a multiple regression analysis in an attempt to predict number of personnel by births, admissions, and outpatient visits. How strong is the model? Explain. Which, if any, of the predictor variables are significant? How does R^2 compare to adjusted R^2 and what does it mean in this model?