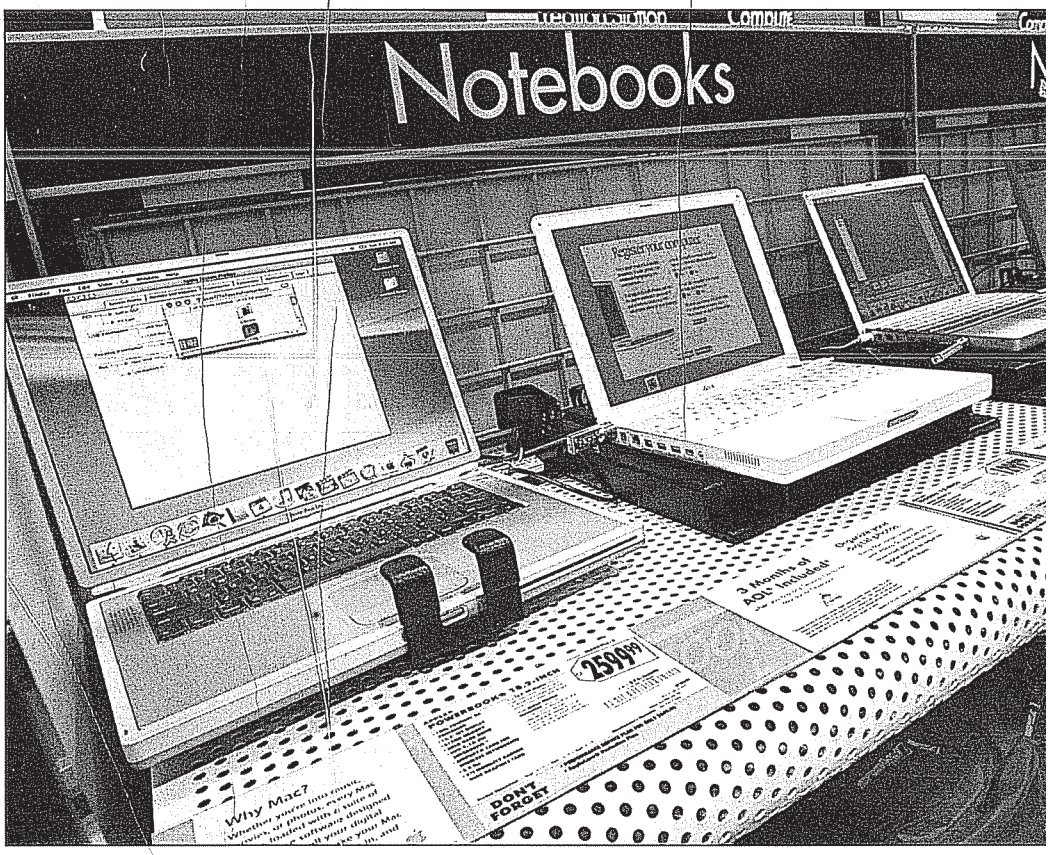
13

Linear Regression and Correlation

GOALS

When you have completed this chapter, you will be able to:

1. **Understand and interpret the terms *dependent* and *independent variable.***
2. **Calculate and interpret the *coefficient of correlation*, *the coefficient of determination*, and the *standard error of estimate.***
3. **Conduct a test of hypothesis to determine whether the coefficient of correlation in the population is zero.**
4. **Calculate the *least squares regression line.***
5. **Construct and interpret *confidence* and *prediction intervals* for the dependent variable.**
6. **Set up and interpret an ANOVA table.**



Use the data given in Exercise 55 showing the retail price for 12 randomly selected laptop computers with their corresponding processor speeds to develop a linear equation that can be used to describe how the price depends on the processor speed. (See Goal 4 and Exercise 55.)

Introduction

Chapters 2 through 4 dealt with *descriptive statistics.* We organized raw data into a frequency distribution, and computed several measures of location and measures of dispersion to describe the major characteristics of the data. Chapter 5 started the study of *statistical inference.* The main emphasis was on inferring something about a population parameter, such as the population mean, on the basis of a sample. We tested for the reasonableness of a population mean or a population proportion, the difference between two population means, or whether several population means were equal. All of these tests involved just *one* interval- or ratio-level variable, such as the weight of a plastic soft drink bottle, the income of bank presidents, or the number of patients admitted to a particular hospital.

We shift the emphasis in this chapter to the study of two variables. Recall in Chapter 4 we introduced the idea of showing the relationship between two variables with a scatter diagram. We plotted the price of vehicles sold at Whitner Autoplex on the vertical axis and the age of the buyer on the horizontal axis. See the statistical software output on page 108. In that case we observed that as the age of the buyer increased, the amount spent for the vehicle also increased. In this chapter we carry this idea further. That is, we develop numerical measures to express the relationship between two variables. Is the relationship strong or weak, is it direct or inverse? In addition we develop an equation to express the relationship between variables. This will allow us to estimate one variable on the basis of another. Here are some examples.

- Is there a relationship between the amount Healthtex spends per month on advertising and the sales in the month?
- Can we base an estimate of the cost to heat a home in January on the number of square feet in the home?
- Is there a relationship between the miles per gallon achieved by large pickup trucks and the size of the engine?
- Is there a relationship between the number of hours that students studied for an exam and the score earned?

Note in each of these cases there are two variables observed for each sampled observation. For the last example, we find, for each student selected for the sample, the hours studied and the score earned.

We begin this chapter by examining the meaning and purpose of **correlation analysis.** We continue our study by developing a mathematical equation that will allow us to estimate the value of one variable based on the value of another. This is called **regression analysis.** We will (1) determine the equation of the line that best fits the data, (2) use the equation to estimate the value of one variable based on another, (3) measure the error in our estimate, and (4) establish confidence and prediction intervals for our estimate.

What Is Correlation Analysis?

Correlation analysis is the study of the relationship between variables. To explain, suppose the sales manager of Copier Sales of America, which has a large sales force throughout the United States and Canada, wants to determine whether there is a

Statistics in Action

The space shuttle Challenger exploded on January 28, 1986. An investigation of the cause examined four contractors: Rockwell International for the shuttle and engines, Lockheed for ground support, Martin Marietta for the external fuel tanks, and Morton Thiokol for the solid fuel booster rockets. After several months, the investigation blamed the explosion on defective O-rings produced by Morton Thiokol. A study of the contractor's stock prices showed an interesting happenstance. On the day of the crash, Morton Thiokol stock was down 11.86% and the stock of the other three lost only 2 to 3%. Can we conclude that financial markets predicted the outcome of the investigation?

relationship between the number of sales calls made in a month and the number of copiers sold that month. The manager selects a random sample of 10 representatives and determines the number of sales calls each representative made last month and the number of copiers sold. The sample information is shown in Table 13–1.

TABLE 13–1 Sales Calls and Copiers Sold for 10 Salespeople

Sales Representative	Number of Sales Calls	Number of Copiers Sold
Tom Keller	20	30
Jeff Hall	40	60
Brian Virost	20	40
Greg Fish	30	60
Susan Welch	10	30
Carlos Ramirez	10	40
Rich Niles	20	40
Mike Kiel	20	50
Mark Reynolds	20	30
Soni Jones	30	70

By reviewing the data we observe that there does seem to be some relationship between the number of sales calls and the number of units sold. That is, the salespeople who made the most sales calls sold the most units. However, the relationship is not "perfect" or exact. For example, Soni Jones made fewer sales calls than Jeff Hall, but she sold more units.

Instead of talking in generalities as we did in Chapter 4 and have so far in this chapter, we will develop some statistical measures to portray more precisely the relationship between the two variables, sales calls and copiers sold. This group of statistical techniques is called **correlation analysis.**

CORRELATION ANALYSIS A group of techniques to measure the association between two variables.

The basic idea of correlation analysis is to report the association between two variables. The usual first step is to plot the data in a **scatter diagram** as we described in Chapter 4. An example will show how a scatter diagram is used.

EXAMPLE

Copier Sales of America sells copiers to businesses of all sizes throughout the United States and Canada. Ms. Marcy Bancer was recently promoted to the position of national sales manager. At the upcoming sales meeting, the sales representatives from all over the country will be in attendance. She would like to impress upon them the importance of making that extra sales call each day. She decides to gather some information on the relationship between the number of sales calls and the number of copiers sold. She selected a random sample of 10 sales representatives and determined the number of sales calls they made last month and the number of copiers they sold. The sample information is reported in Table 13–1. What observations can you make about the relationship between the number of sales calls and the number of copiers sold? Develop a scatter diagram to display the information.

SOLUTION

Based on the information in Table 13–1, Ms. Bancer suspects there is a relationship between the number of sales calls made in a month and the number of copiers sold. Soni Jones sold the most copiers last month, and she was one of three representatives making 30 or more sales calls. On the other hand, Susan Welch and Carlos

Ramirez made only 10 sales calls last month. Ms. Welch had the lowest number of copiers sold among the sampled representatives.

The implication is that the number of copiers sold is related to the number of sales calls made. As the number of sales calls increases, it appears the number of copiers sold also increases. We refer to number of sales calls as the **independent variable** and number of copiers sold as the **dependent variable.**

DEPENDENT VARIABLE The variable that is being predicted or estimated.

INDEPENDENT VARIABLE A variable that provides the basis for estimation. It is the predictor variable.

It is common practice to scale the dependent variable (copiers sold) on the vertical or *Y*-axis and the independent variable (number of sales calls) on the horizontal or *X*-axis. To develop the scatter diagram of the Copier Sales of America sales information, we begin with the first sales representative, Tom Keller. Tom made 20 sales calls last month and sold 30 copiers, so $X = 20$ and $Y = 30$. To plot this point, move along the horizontal axis to $X = 20$, then go vertically to $Y = 30$ and place a dot at the intersection. This process is continued until all the paired data are plotted, as shown in Chart 13–1.

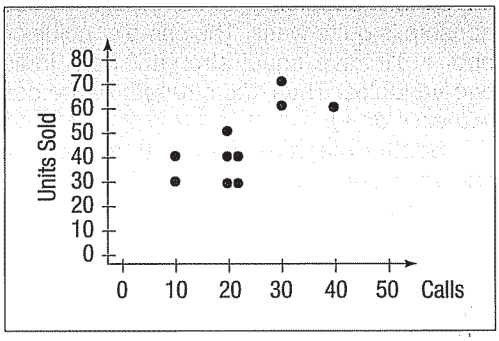


CHART 13–1 Scatter Diagram Showing Sales Calls and Copiers Sold

The scatter diagram shows graphically that the sales representatives who make more calls tend to sell more copiers. It is reasonable for Ms. Bancer, the national sales manager at Copier Sales of America, to tell her salespeople that the more sales calls they make the more copiers they can expect to sell. Note that while there appears to be a positive relationship between the two variables, all the points do not fall on a line. In the following section you will measure the strength and direction of this relationship between two variables by determining the coefficient of correlation.

The Coefficient of Correlation

Interval- or ratio-level data are required

Originated by Karl Pearson about 1900, the **coefficient of correlation** describes the strength of the relationship between two sets of interval-scaled or ratio-scaled variables. Designated *r,* it is often referred to as *Pearson's r* and as the *Pearson product-moment correlation coefficient.* It can assume any value from −1.00 to +1.00 inclusive. A correlation coefficient of −1.00 or +1.00 indicates *perfect correlation.* For example, a correlation coefficient for the preceding example computed to be +1.00

Characteristics of *r*

would indicate that the number of sales calls and the number of copiers sold are perfectly related in a positive linear sense. A computed value of −1.00 reveals that sales calls and the number of copiers sold are perfectly related in an inverse linear sense. How the scatter diagram would appear if the relationship between the two sets of data were linear and perfect is shown in Chart 13–2.

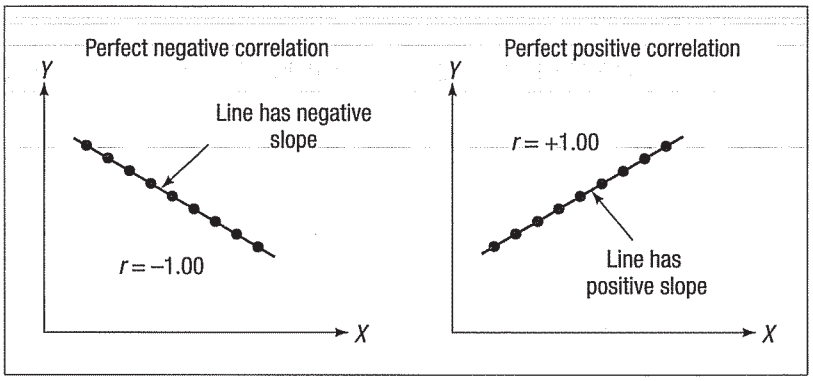


CHART 13–2 Scatter Diagrams Showing Perfect Negative Correlation and Perfect Positive Correlation

If there is absolutely no relationship between the two sets of variables, Pearson's *r* is zero. A coefficient of correlation *r* close to 0 (say, .08) shows that the linear relationship is quite weak. The same conclusion is drawn if $r = -.08$. Coefficients of −.91 and +.91 have equal strength; both indicate very strong correlation between the two variables. Thus, *the strength of the correlation does not depend on the direction (either − or +).*

Scatter diagrams for $r = 0$, a weak *r* (say, −.23), and a strong *r* (say, +.87) are shown in Chart 13–3. Note that if the correlation is weak, there is considerable scatter about a line drawn through the center of the data. For the scatter diagram representing a strong relationship, there is very little scatter about the line. This indicates, in the example shown on the chart, that hours studied is a good predictor of exam score.

Examples of degrees of correlation

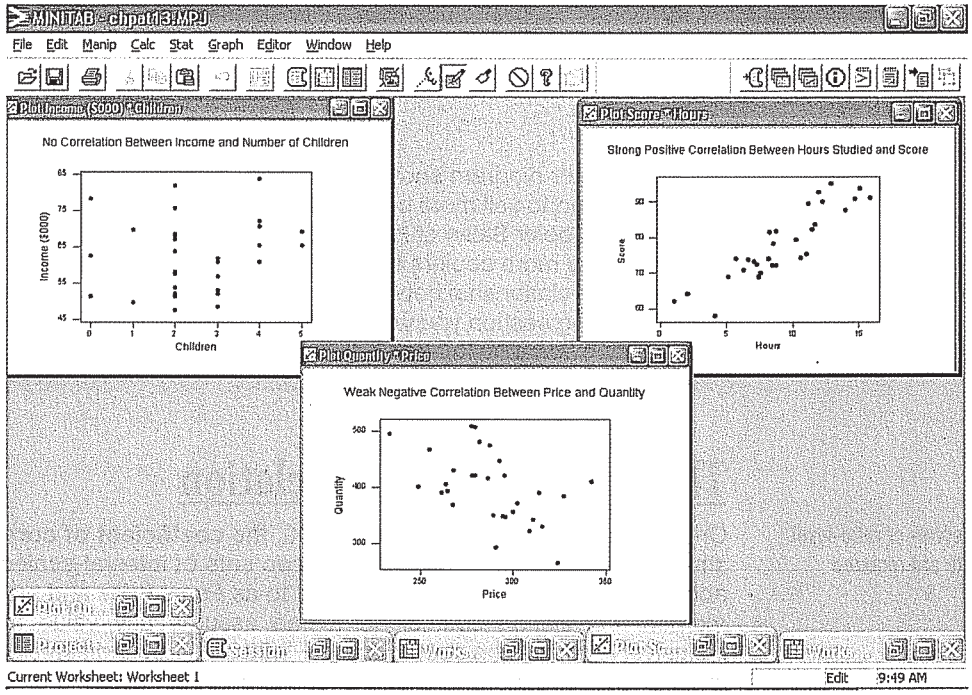
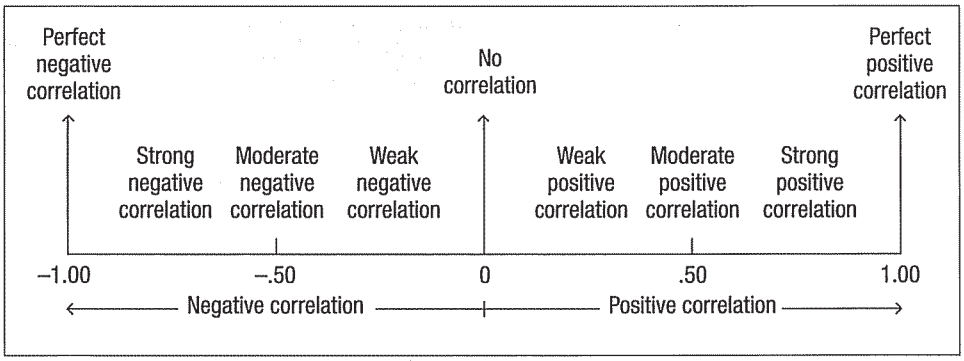


CHART 13–3 Scatter Diagrams Depicting Zero, Weak, and Strong Correlation

The following drawing summarizes the strength and direction of the coefficient of correlation.



COEFFICIENT OF CORRELATION A measure of the strength of the linear relationship between two variables.

How is the value of the coefficient of correlation determined? We will use the Copier Sales of America data, which are reported in Table 13–2, as an example. We begin with a scatter diagram, similar to Chart 13–2. Draw a vertical line through the data values at the mean of the X-values and a horizontal line at the mean of the Y-values. In Chart 13–4 we've added a vertical line at 22.0 calls ($\bar{X} = \Sigma X/n = 220/10 = 22$) and a horizontal line at 45.0 copiers ($\bar{Y} = \Sigma Y/n = 450/10 = 45.0$). These lines pass through the "center" of the data and divide the scatter diagram into four quadrants. Think of moving the origin from (0, 0) to (22, 45).

TABLE 13–2 Sales Calls and Copiers Sold for 10 Salespeople

Sales Representative	Sales Calls (X)	Copiers Sold (Y)
Tom Keller	20	30
Jeff Hall	40	60
Brian Virost	20	40
Greg Fish	30	60
Susan Welch	10	30
Carlos Ramirez	10	40
Rich Niles	20	40
Mike Kiel	20	50
Mark Reynolds	20	30
Soni Jones	30	70
Total	220	450

Two variables are positively related, when the number of copiers sold is above the mean and the number of sales calls is also above the mean. These points appear in the upper-right quadrant of Chart 13–4 on the next page. Similarly, when the number of copiers sold is less than the mean, so is the number of sales calls. These points fall in the lower-left quadrant of Chart 13–4. For example, the last person on the list in Table 13–2, Soni Jones, made 30 sales calls and sold 70 copiers. These values are above their respective means, so this point is located in the upper-right quadrant. She made 8 ($X - \bar{X} = 30 - 22$) more sales calls than the mean and sold 25 ($Y - \bar{Y} =$

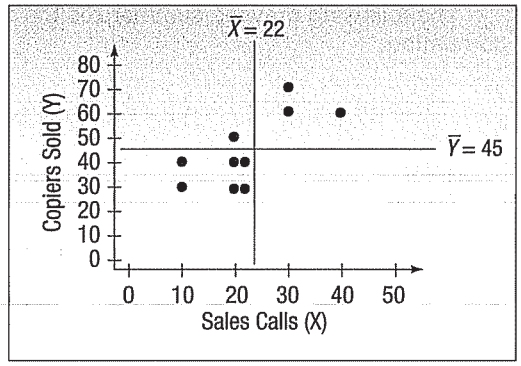


CHART 13–4 Computation of the Coefficient of Correlation

70 − 45) more copiers than the mean. Tom Keller, the first name on the list in Table 13–2, made 20 sales calls and sold 30 copiers. Both of these values are less than their respective mean; hence this point is in the lower-left quadrant. Tom made 2 less sales calls and sold 15 less copiers than the respective means. The deviations from the mean number of sales calls and for the mean number of copiers sold are summarized in Table 13–3 for the 10 sales representatives. The sum of the products of the deviations from the respective means is 900. That is, the term $\Sigma(X - \bar{X})(Y - \bar{Y}) = 900$.

TABLE 13–3 Deviations from the Mean and Their Products

Sales Representative	Calls Y	Sales X	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})(Y - \bar{Y})$
Tom Keller	20	30	−2	−15	30
Jeff Hall	40	60	18	15	270
Brian Virost	20	40	−2	−5	10
Greg Fish	30	60	8	15	120
Susan Welch	10	30	−12	−15	180
Carlos Ramirez	10	40	−12	−5	60
Rich Niles	20	40	−2	−5	10
Mike Kiel	20	50	−2	5	−10
Mark Reynolds	20	30	−2	−15	30
Soni Jones	30	70	8	25	200
					900

In both the upper-right and the lower-left quadrants, the product of $(X - \bar{X})(Y - \bar{Y})$ is positive because both of the factors have the same sign. In our example this happens for all sales representatives except Mike Kiel. We can therefore expect the coefficient of correlation to have a positive value.

If the two variables are inversely related, one variable will be above the mean and the other below the mean. Most of the points in this case occur in the upper-left and lower-right quadrants. Now $(X - \bar{X})$ and $(Y - \bar{Y})$ will have opposite signs, so their product is negative. The resulting correlation coefficient is negative.

What happens if there is no linear relationship between the two variables? The points in the scatter diagram will appear in all four quadrants. The negative products of $(X - \bar{X})(Y - \bar{Y})$ offset the positive products, so the sum is near zero. This leads to a correlation coefficient near zero.

Pearson also wanted the correlation coefficient to be unaffected by the units of the two variables. For example, if we had used hundreds of copiers sold instead of the number sold, the coefficient of correlation would be the same. The coefficient of correlation is independent of the scale used if we divide the term $\Sigma(X - \bar{X})(Y - \bar{Y})$ by the

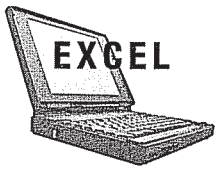
sample standard deviations. It is also made independent of the sample size and bounded by the values +1.00 and −1.00 if we divide by $(n - 1)$.

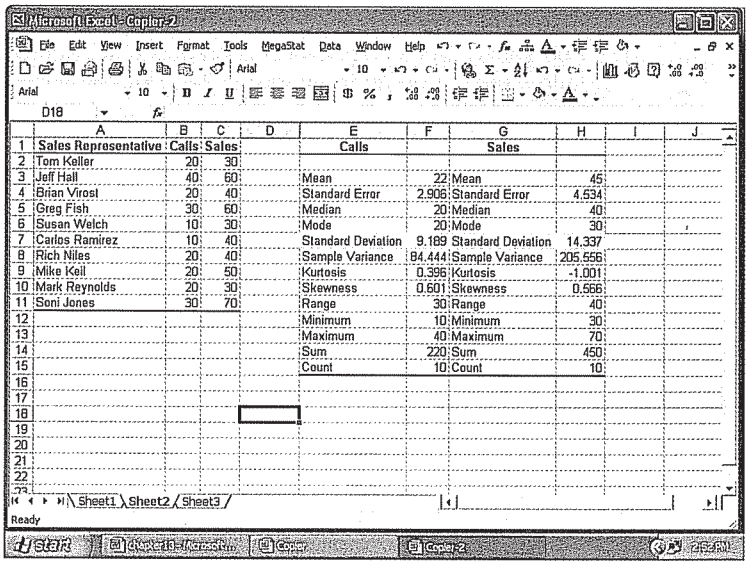
This reasoning leads to the following formula:

CORRELATION COEFFICIENT

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{(n - 1)\, s_x\, s_y} \qquad [13\text{–}1]$$

To compute the coefficient of correlation, we use the standard deviations of the sample of 10 sales calls and 10 copiers sold. We could use formula (3–11) to calculate the sample standard deviations or we could use a software package. For the specific Excel and MINITAB commands see the Software Command section at the end of Chapter 3. The following is the Excel output. The standard deviation of the number of sales calls is 9.189 and of the number of copiers sold 14.337.





Microsoft Excel - Copier-2

File Edit View Insert Format Tools MegaStat Data Window Help

D18

	A	B	C	D	E	F	G	H	I	J
1	Sales Representative	Calls	Sales		Calls		Sales			
2	Tom Keller	20	30							
3	Jeff Hall	40	60		Mean	22	Mean	45		
4	Brian Virosl	20	40		Standard Error	2.906	Standard Error	4.534		
5	Greg Fish	30	60		Median	20	Median	40		
6	Susan Welch	10	30		Mode	20	Mode	30		
7	Carlos Ramirez	10	40		Standard Deviation	9.189	Standard Deviation	14.337		
8	Rich Niles	20	40		Sample Variance	84.444	Sample Variance	205.556		
9	Mike Keil	20	50		Kurtosis	0.396	Kurtosis	-1.001		
10	Mark Reynolds	20	30		Skewness	0.601	Skewness	0.566		
11	Soni Jones	30	70		Range	30	Range	40		
12					Minimum	10	Minimum	30		
13					Maximum	40	Maximum	70		
14					Sum	220	Sum	450		
15					Count	10	Count	10		
16										
17										
18										
19										
20										
21										
22										

Sheet1 Sheet2 Sheet3

Ready

We now insert these values into formula (13–1) to determine the coefficient of correlation:

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{(n - 1)\, s_x\, s_y} = \frac{900}{(10 - 1)(9.189)(14.337)} = 0.759$$

How do we interpret a correlation of 0.759? First, it is positive, so we see there is a direct relationship between the number of sales calls and the number of copiers sold. This confirms our reasoning based on the scatter diagram, Chart 13–4. The value of 0.759 is fairly close to 1.00, so we conclude that the association is strong. To put it another way, an increase in calls will likely lead to more sales.

The Coefficient of Determination

In the previous example regarding the relationship between the number of sales calls and the units sold, the coefficient of correlation, 0.759, was interpreted as being "strong." Terms such as *weak, moderate,* and *strong,* however, do not have precise meaning. A measure that has a more easily interpreted meaning is the **coefficient of determination.** It is computed by squaring the coefficient of correlation. In the example, the coefficient of determination, r^2, is 0.576, found by $(0.759)^2$. This is a proportion or a percent; we can say that 57.6 percent of the variation in the number of copiers sold is explained, or accounted for, by the variation in the number of sales calls.

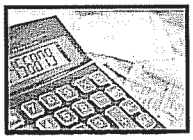
COEFFICIENT OF DETERMINATION The proportion of the total variation in the dependent variable Y that is explained, or accounted for, by the variation in the independent variable X.

Further discussion of the coefficient of determination is found later in the chapter.

Correlation and Cause

If there is a strong relationship (say, $r = .91$) between two variables, we are tempted to assume that an increase or decrease in one variable *causes* a change in the other variable. For example, it can be shown that the consumption of Georgia peanuts and the consumption of aspirin have a strong correlation. However, this does not indicate that an increase in the consumption of peanuts *caused* the consumption of aspirin to increase. Likewise, the incomes of professors and the number of inmates in mental institutions have increased proportionately. Further, as the population of donkeys has decreased, there has been an increase in the number of doctoral degrees granted. Relationships such as these are called **spurious correlations.** What we can conclude when we find two variables with a strong correlation is that there is a relationship or association between the two variables, not that a change in one causes a change in the other.

Self-Review 13–1



Haverty's Furniture is a family business that has been selling to retail customers in the Chicago area for many years. They advertise extensively on radio, TV, and the Internet emphasizing their low prices and easy credit terms. The owner would like to review the relationship between sales and the amount spent on advertising. Below is information on sales and advertising expense for the last four months.

Month	Advertising Expense ($ million)	Sales Revenue ($ million)
July	2	7
August	1	3
September	3	8
October	4	10

(a) The owner wants to forecast sales on the basis of advertising expense. Which variable is the dependent variable? Which variable is the independent variable?
(b) Draw a scatter diagram.
(c) Determine the coefficient of correlation.
(d) Interpret the strength of the correlation coefficient.
(e) Determine the coefficient of determination. Interpret.

Exercises

1. The following sample observations were randomly selected.

X:	4	5	3	6	10
Y:	4	6	5	7	7

Determine the coefficient of correlation and the coefficient of determination. Interpret.

2. The following sample observations were randomly selected.

X:	5	3	6	3	4	4	6	8
Y:	13	15	7	12	13	11	9	5

Determine the coefficient of correlation and the coefficient of determination. Interpret the association between X and Y.

3. Bi-lo Appliance Stores has outlets in several large metropolitan areas in New England. The general sales manager plans to air a commercial for a digital camera on selected local TV stations prior to a sale starting on Saturday and ending Sunday. She plans to get the information for Saturday–Sunday digital camera sales at the various outlets and pair them with the number of times the advertisement was shown on the local TV stations. The purpose is to find whether there is any relationship between the number of times the advertisement was aired and digital camera sales. The pairings are:

Location of TV Station	Number of Airings	Saturday–Sunday Sales ($ thousands)
Providence	4	15
Springfield	2	8
New Haven	5	21
Boston	6	24
Hartford	3	17

   a. What is the dependent variable?
   b. Draw a scatter diagram.
   c. Determine the coefficient of correlation.
   d. Determine the coefficient of determination.
   e. Interpret these statistics.

4. The production department of NDB Electronics wants to explore the relationship between the number of employees who assemble a subassembly and the number produced. As an experiment, two employees were assigned to assemble the subassemblies. They produced 15 during a one-hour period. Then four employees assembled them. They produced 25 during a one-hour period. The complete set of paired observations follows.

Number of Assemblers	One-Hour Production (units)
2	15
4	25
1	10
5	40
3	30

   The dependent variable is production; that is, it is assumed that the level of production depends upon the number of employees.
   a. Draw a scatter diagram.
   b. Based on the scatter diagram, does there appear to be any relationship between the number of assemblers and production? Explain.
   c. Compute the coefficient of correlation.
   d. Evaluate the strength of the relationship by computing the coefficient of determination.

5. The city council of Pine Bluffs is considering increasing the number of police in an effort to reduce crime. Before making a final decision, the council asks the Chief of Police to survey other cities of similar size to determine the relationship between the number of police and the number of crimes reported. The Chief gathered the following sample information.

City	Police	Number of Crimes	City	Police	Number of Crimes
Oxford	15	17	Holgate	17	7
Starksville	17	13	Carey	12	21
Danville	25	5	Whistler	11	19
Athens	27	7	Woodville	22	6

   a. If we want to estimate crimes on the basis of the number of police, which variable is the dependent variable and which is the independent variable?
   b. Draw a scatter diagram.

c. Determine the coefficient of correlation.
d. Determine the coefficient of determination.
e. Interpret these statistics. Does it surprise you that the relationship is inverse?

6. The owner of Maumee Ford-Mercury wants to study the relationship between the age of a car and its selling price. Listed below is a random sample of 12 used cars sold at the dealership during the last year.

Car	Age (years)	Selling Price ($000)	Car	Age (years)	Selling Price ($000)
1	9	8.1	7	8	7.6
2	7	6.0	8	11	8.0
3	11	3.6	9	10	8.0
4	12	4.0	10	12	6.0
5	8	5.0	11	6	8.6
6	7	10.0	12	6	8.0

a. If we want to estimate selling price on the basis of the age of the car, which variable is the dependent variable and which is the independent variable?
b. Draw a scatter diagram.
c. Determine the coefficient of correlation.
d. Determine the coefficient of determination.
e. Interpret these statistics. Does it surprise you that the relationship is inverse?

Testing the Significance of the Correlation Coefficient

Recall the sales manager of Copier Sales of America found the correlation between the number of sales calls and the number of copiers sold was 0.759. This indicated a strong association between the two variables. However, only 10 salespeople were sampled. Could it be that the correlation in the population is actually 0? This would mean the correlation of 0.759 was due to chance. The population in this example is all the salespeople employed by the firm.

Could the correlation in the population be zero?

Resolving this dilemma requires a test to answer the obvious question: Could there be zero correlation in the population from which the sample was selected? To put it another way, did the computed r come from a population of paired observations with zero correlation? To continue our convention of allowing Greek letters to represent a population parameter, we will let ρ represent the correlation in the population. It is pronounced "rho."

We will continue with the illustration involving sales calls and copiers sold. We employ the same hypothesis testing steps described in Chapter 10. The null hypothesis and the alternate hypothesis are:

H_0: $\rho = 0$ (The correlation in the population is zero.)
H_1: $\rho \neq 0$ (The correlation in the population is different from zero.)

From the way H_1 is stated, we know that the test is two-tailed.

The formula for t is:

t TEST FOR THE COEFFICIENT OF CORRELATION

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad \text{with } n-2 \text{ degrees of freedom} \qquad [13\text{–}2]$$

Using the .05 level of significance, the decision rule states that if the computed t falls in the area between plus 2.306 and minus 2.306, the null hypothesis is not rejected. To

locate the critical value of 2.306, refer to Appendix F for $df = n - 2 = 10 - 2 = 8$. See Chart 13–5.

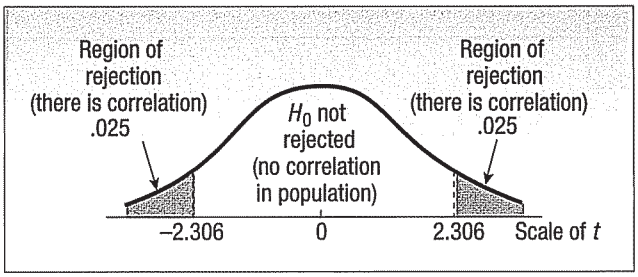


CHART 13–5 Decision Rule for Test of Hypothesis at .05 Significance Level and 8 *df*

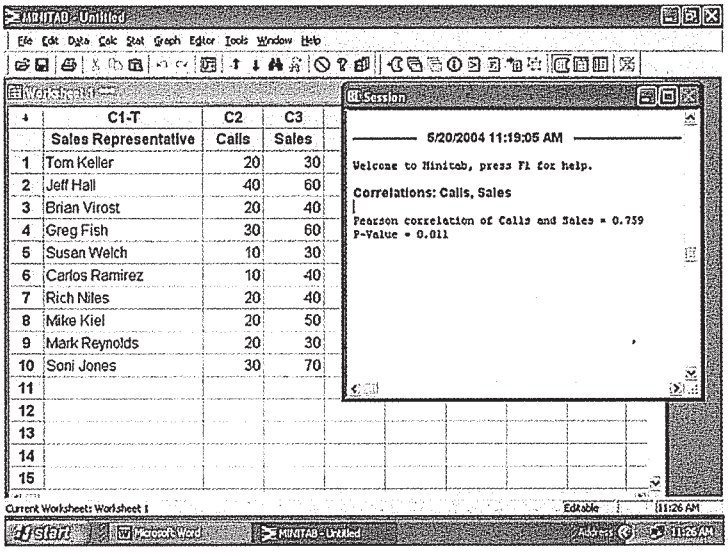
Applying formula (13–2) to the example regarding the number of sales calls and units sold:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{.759\sqrt{10-2}}{\sqrt{1-.759^2}} = 3.297$$

The computed t is in the rejection region. Thus, H_0 is rejected at the .05 significance level. This means the correlation in the population is not zero. From a practical standpoint, it indicates to the sales manager that there is correlation with respect to the number of sales calls made and the number of copiers sold in the population of salespeople.

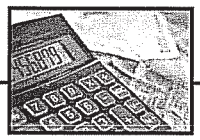
We can also interpret the test of hypothesis in terms of *p*-values. A *p*-value is the likelihood of finding a value of the test statistic more extreme than the one computed, when H_0 is true. To determine the *p*-value, go to the *t* distribution in Appendix F and find the row for 8 degrees of freedom. The value of the test statistic is 3.297, so in the row for 8 degrees of freedom and a two-tailed test, find the value closest to 3.297. For a two-tailed test at the .02 significance level, the critical value is 2.896, and the critical value at the .01 significance level is 3.355. Because 3.297 is between 2.896 and 3.355 we conclude that the *p*-value is between .01 and .02.

Both MINITAB and Excel will report the correlation between two variables. In addition to the correlation, MINITAB reports the *p*-value for the test of hypothesis that the correlation in the population between the two variables is 0. The MINITAB output showing the results is below. They are the same as those calculated earlier.



	C1-T	C2	C3
	Sales Representative	Calls	Sales
1	Tom Keller	20	30
2	Jeff Hall	40	60
3	Brian Virost	20	40
4	Greg Fish	30	60
5	Susan Welch	10	30
6	Carlos Ramirez	10	40
7	Rich Niles	20	40
8	Mike Kiel	20	50
9	Mark Reynolds	20	30
10	Soni Jones	30	70
11			
12			
13			
14			
15			

Self-Review 13–2



A sample of 25 mayoral campaigns in cities with populations larger than 50,000 showed that the correlation between the percent of the vote received and the amount spent on the campaign by the candidate was .43. At the .05 significance level, is there a positive association between the variables?

Exercises

7. The following hypotheses are given.

H_0: $\rho \leq 0$
H_1: $\rho > 0$

A random sample of 12 paired observations indicated a correlation of .32. Can we conclude that the correlation in the population is greater than zero? Use the .05 significance level.

8. The following hypotheses are given.

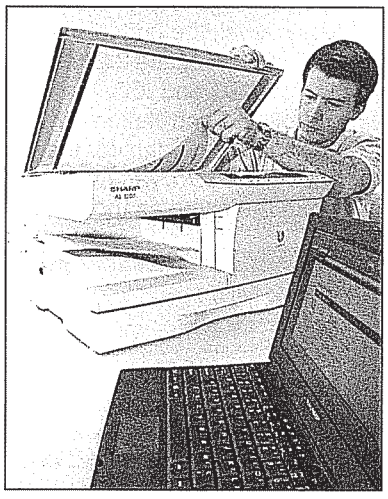
H_0: $\rho \geq 0$
H_1: $\rho < 0$

A random sample of 15 paired observations has a correlation of $-.46$. Can we conclude that the correlation in the population is less than zero? Use the .05 significance level.

9. The Pennsylvania Refining Company is studying the relationship between the pump price of gasoline and the number of gallons sold. For a sample of 20 stations last Tuesday, the correlation was .78. At the .01 significance level, is the correlation in the population greater than zero?

10. A study of 20 worldwide financial institutions showed the correlation between their assets and pretax profit to be .86. At the .05 significance level, can we conclude that there is positive correlation in the population?

Regression Analysis



In the previous section we developed measures to express the strength and the direction of the relationship between two variables. In this section we wish to develop an equation to express the *linear* (straight line) relationship between two variables. In addition we want to be able to estimate the value of the dependent variable Y based on a selected value of the independent variable X. The technique used to develop the equation and provide the estimates is called **regression analysis.**

In Table 13–1 we reported the number of sales calls and the number of units sold for a sample of 10 sales representatives employed by Copier Sales of America. Chart 13–1 portrayed this information in a scatter diagram. Now we want to develop a linear equation that expresses the relationship between the number of sales calls and the number of units sold. The equation for the line used to estimate Y on the basis of X is referred to as the **regression equation.**

REGRESSION EQUATION An equation that expresses the linear relationship between two variables.

Least Squares Principle

The scatter diagram in Chart 13–1 is reproduced in Chart 13–6, with a line drawn with a ruler through the dots to illustrate that a straight line would probably fit the data.

However, the line drawn using a straight edge has one disadvantage: Its position is based in part on the judgment of the person drawing the line. The hand-drawn lines in Chart 13–7 represent the judgments of four people. All the lines except line *A* seem to be reasonable. However, each would result in a different estimate of units sold for a particular number of sales calls.

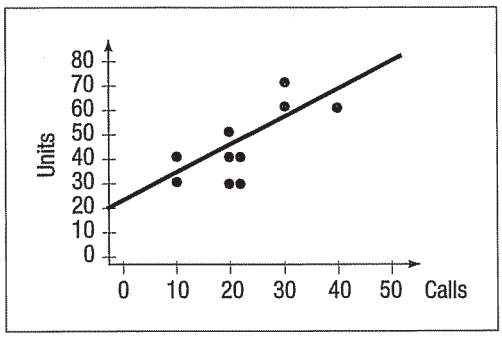


CHART 13–6 Sales Calls and Copiers Sold for 10 Sales Representatives

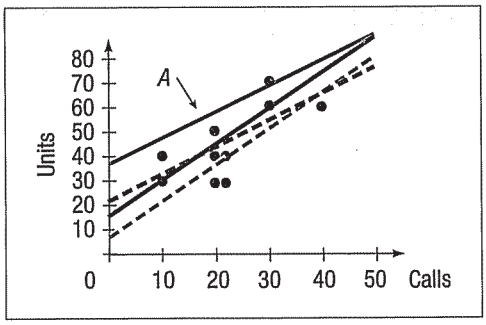


CHART 13–7 Four Lines Superimposed on the Scatter Diagram

Least squares line gives "best" fit; subjective method is unreliable.

Judgment is eliminated by determining the regression line using a mathematical method called the **least squares principle.** This method gives what is commonly referred to as the "best-fitting" line.

LEAST SQUARES PRINCIPLE Determining a regression equation by minimizing the sum of the squares of the vertical distances between the actual *Y* values and the predicted values of *Y*.

To illustrate this concept, the same data are plotted in the three charts that follow. The regression line in Chart 13–8 was determined using the least squares method. It is the best-fitting line because the sum of the squares of the vertical deviations about it is at a minimum. The first plot ($X = 3$, $Y = 8$) deviates by 2 from the line, found by $10 - 8$. The deviation squared is 4. The squared deviation for the plot $X = 4$, $Y = 18$ is 16. The squared deviation for the plot $X = 5$, $Y = 16$ is 4. The sum of the squared deviations is 24, found by $4 + 16 + 4$.

Assume that the lines in Charts 13–9 and 13–10 were drawn with a straight edge. The sum of the squared vertical deviations in Chart 13–9 is 44. For Chart 13–10 it is 132. Both sums are greater than the sum for the line in Chart 13–8, found by using the least squares method.

The equation of a straight line has the form

GENERAL FORM OF LINEAR REGRESSION EQUATION $$Y' = a + bX$$ **[13–3]**

where:

Y' read *Y* prime, is the predicted value of the *Y* variable for a selected *X* value.
a is the *Y*-intercept. It is the estimated value of *Y* when $X = 0$. Another way to put it is: *a* is the estimated value of *Y* where the regression line crosses the *Y*-axis when *X* is zero.
b is the slope of the line, or the average change in Y' for each change of one unit (either increase or decrease) in the independent variable *X*.
X is any value of the independent variable that is selected.

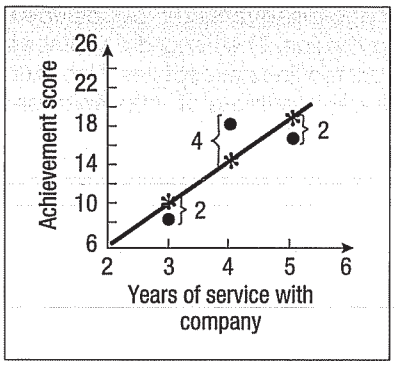


CHART 13–8 The Least Squares Line

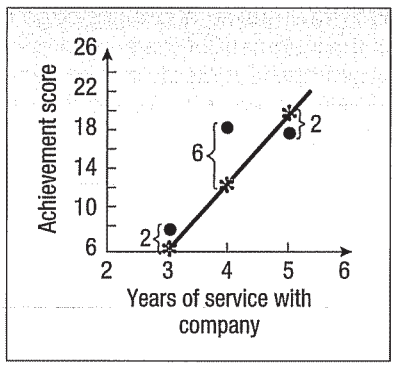


CHART 13–9 Line Drawn with a Straight Edge

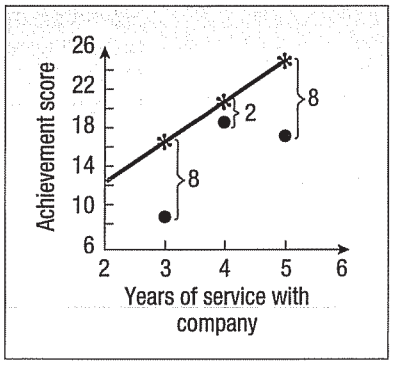


CHART 13–10 Line Drawn with a Straight Edge

The formulas for a and b are:

SLOPE OF THE REGRESSION LINE $$b = r\frac{s_y}{s_x}$$ **[13-4]**

where:

r is the correlation coefficient.
s_y is the standard deviation of Y (the dependent variable).
s_x is the standard deviation of X (the independent variable).

Y-INTERCEPT $$a = \bar{Y} - b\bar{X}$$ **[13-5]**

where:

$\bar{Y}$ is the mean of Y (the dependent variable).
$\bar{X}$ is the mean of X (the independent variable).

The following example shows the details of determining the slope and intercept values.

EXAMPLE

Recall the example involving Copier Sales of America. The sales manager gathered information on the number of sales calls made and the number of copiers sold for a random sample of 10 sales representatives. As a part of her presentation at the upcoming sales meeting, Ms. Bancer, the sales manager, would like to offer specific information about the relationship between the number of sales calls and the number of copiers sold. Use the least squares method to determine a linear equation to express the relationship between the two variables. What is the expected number of copiers sold by a representative who made 20 calls?

SOLUTION

The calculations necessary to determine the regression equation are:

$$b = r\left(\frac{s_y}{s_x}\right) = .759\left(\frac{14.337}{9.189}\right) = 1.1842$$

$$a = \bar{Y} - b\bar{X} = 45 - (1.1842)22 = 18.9476$$

The standard deviation for the sales calls (X) and the units sold (Y) as well as their respective means can be found in the Excel spreadsheet on page 381. The value of r is calculated just below the spreadsheet.

Thus, the regression equation is $Y' = 18.9476 + 1.1842X$. So if a salesperson makes 20 calls, he or she can expect to sell 42.6316 copiers, found by $Y' = 18.9476 + 1.1842X = 18.9476 + 1.1842(20)$. The b value of 1.1842 means that for each additional sales call made the sales representative can expect to increase the number of copiers sold by about 1.2. To put it another way, five additional sales calls in a month will result in about six more copiers being sold, found by $1.1842(5) = 5.921$.

The a value of 18.9476 is the point where the equation crosses the Y-axis. A literal translation is that if no sales calls are made, that is, $X = 0$, 18.9476 copiers will be sold. Note that $X = 0$ is outside the range of values included in the sample and, therefore, should not be used to estimate the number of copiers sold. The sales calls ranged from 10 to 40, so estimates should be made within that range.

Statistics in Action

In finance, investors are interested in the tradeoff between returns and risk. One technique to quantify risk is a regression analysis of a company's stock price (dependent variable) and an average measure of the stock market (independent variable). Often the Standard and Poor's (S&P) 500 Index is used to estimate the market. The regression coefficient, called Beta in finance, shows the change in a company's stock price for a one-unit change in the S&P Index. For example, if a stock has a beta of 1.5, then when the S&P index increases by 1%, the stock price will increase by 1.5%. The opposite is also true. If the S&P decreases by 1%, the stock price will decrease by 1.5%. If the beta is 1.0, then a 1% change in the index should show a 1% change in a stock price. If the beta is less than 1.0, then a 1% change in the index shows less than a 1% change in the stock price.

Drawing the Line of Regression

The least squares equation, $Y' = 18.9476 + 1.1842X$, can be drawn on the scatter diagram. The first sales representative in the sample is Tom Keller. He made 20 calls. His estimated number of copiers sold is $Y' = 18.9476 + 1.1842(20) = 42.6316$. The plot $X = 20$ and $Y = 42.6316$ is located by moving to 20 on the X-axis and then going vertically to 42.6316. The other points on the regression equation can be determined by substituting the particular value of X into the regression equation.

Sales Representative	Sales Calls (X)	Estimated Sales (Y')	Sales Representative	Sales Calls (X)	Estimated Sales (Y')
Tom Keller	20	42.6316	Carlos Ramirez	10	30.7896
Jeff Hall	40	66.3156	Rich Niles	20	42.6316
Brian Virost	20	42.6316	Mike Kiel	20	42.6316
Greg Fish	30	54.4736	Mark Reynolds	20	42.6316
Susan Welch	10	30.7896	Soni Jones	30	54.4736

All the other points are connected to give the line. See Chart 13–11.

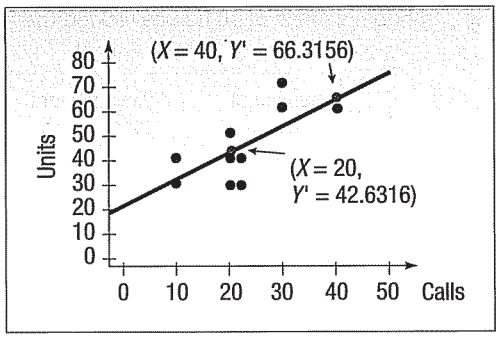
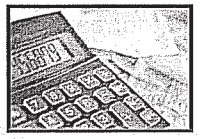


CHART 13–11 The Line of Regression Drawn on the Scatter Diagram

This line has some interesting features. As we have discussed, there is no other line through the data for which the sum of the squared deviations is smaller. In addition, this line will pass through the points represented by the mean of the X values and the mean of the Y values, that is, $\bar{X}$ and $\bar{Y}$. In this example $\bar{X} = 22.0$ and $\bar{Y} = 45.0$.

Self-Review 13–3



Refer to Self-Review 13–1, where the owner of Haverty's Furniture Company was studying the relationship between sales and the amount spent on advertising. The sales information for the last four months is repeated below.

Month	Advertising Expense ($ million)	Sales Revenue ($ million)
July	2	7
August	1	3
September	3	8
October	4	10

(a) Determine the regression equation.
(b) Interpret the values of *a* and *b*.
(c) Estimate sales when $3 million is spent on advertising.

Exercises

11. The following sample observations were randomly selected.

X:	4	5	3	6	10
Y:	4	6	5	7	7

a. Determine the regression equation.
b. Determine the value of Y' when X is 7.

12. The following sample observations were randomly selected.

X:	5	3	6	3	4	4	6	8
Y:	13	15	7	12	13	11	9	5

a. Determine the regression equation.
b. Determine the value of Y' when X is 7.

13. The Bradford Electric Illuminating Company is studying the relationship between kilowatt-hours (thousands) used and the number of rooms in a private single-family residence. A random sample of 10 homes yielded the following.

Number of Rooms	Kilowatt-Hours (thousands)	Number of Rooms	Kilowatt-Hours (thousands)
12	9	8	6
9	7	10	8
14	10	10	10
6	5	5	4
10	8	7	7

a. Determine the regression equation.
b. Determine the number of kilowatt-hours, in thousands, for a six-room house.

14. Mr. James McWhinney, president of Daniel-James Financial Services, believes there is a relationship between the number of client contacts and the dollar amount of sales. To document this assertion, Mr. McWhinney gathered the following sample information. The X column indicates the number of client contacts last month, and the Y column shows the value of sales ($ thousands) last month for each salesperson sampled.

Salesperson	Contacts (X)	Sales (Y)
Robert Armstrong	14	24
Jack Bender	12	14
Dorothy Brumley	20	28
Carmen Carella	16	30
Annette Perrault	46	80
Mary Jane Duryee	23	30
David Gwyer	48	90
Harvey Lazik	50	85
Ray Osbeck	55	120
Al Montanaro	50	110

a. Determine the regression equation.
b. Determine the estimated sales if 40 contacts are made.

15. A recent article in *Business Week* listed the "Best Small Companies." We are interested in the current results of the companies' sales and earnings. A random sample of 12 companies was selected and the sales and earnings, in millions of dollars, are reported below.

Company	Sales ($ millions)	Earnings ($ millions)	Company	Sales ($ millions)	Earnings ($ millions)
Papa John's International	$89.2	$4.9	Checkmate Electronics	$17.5	$2.6
Applied Innovation	18.6	4.4	Royal Grip	11.9	1.7
Integracare	18.2	1.3	M-Wave	19.6	3.5
Wall Data	71.7	8.0	Serving-N-Slide	51.2	8.2
Davidson Associates	58.6	6.6	Daig	28.6	6.0
Chico's Fas	46.8	4.1	Cobra Golf	69.2	12.8

Let sales be the independent variable and earnings be the dependent variable.
a. Draw a scatter diagram.
b. Compute the coefficient of correlation.
c. Compute the coefficient of determination.
d. Interpret your findings in parts b and c.
e. Determine the regression equation.
f. For a small company with $50.0 million in sales, estimate the earnings.

16. We are studying mutual bond funds for the purpose of investing in several funds. For this particular study, we want to focus on the assets of a fund and its five-year performance. The question is: Can the five-year rate of return be estimated based on the assets of the fund? Nine mutual funds were selected at random, and their assets and rates of return are shown below.

Fund	Assets ($ millions)	Return (%)	Fund	Assets ($ millions)	Return (%)
AARP High Quality Bond	$622.2	10.8	MFS Bond A	$494.5	11.6
Babson Bond L	160.4	11.3	Nichols Income	158.3	9.5
Compass Capital Fixed Income	275.7	11.4	T. Rowe Price Short-term	681.0	8.2
Galaxy Bond Retail	433.2	9.1	Thompson Income B	241.3	6.8
Keystone Custodian B-1	437.9	9.2			

a. Draw a scatter diagram.
b. Compute the coefficient of correlation.
c. Compute the coefficient of determination.
d. Write a brief report of your findings for parts b and c.
e. Determine the regression equation. Use assets as the independent variable.
f. For a fund with $400.0 million in sales, determine the five-year rate of return (in percent).

17. Refer to Exercise 5.
    a. Determine the regression equation.
    b. Estimate the number of crimes for a city with 20 police.
    c. Interpret the regression equation.
18. Refer to Exercise 6.
    a. Determine the regression equation.
    b. Estimate the selling price of a 10-year-old car.
    c. Interpret the regression equation.

The Standard Error of Estimate

Note in the preceding scatter diagram (Chart 13–11) that all of the points do not lie exactly on the regression line. If they all were on the line, there would be no error in estimating the number of units sold. To put it another way, if all the points were on the regression line, units sold could be predicted with 100 percent accuracy. Thus, there would be no error in predicting the Y variable based on an X variable. This is true in the following hypothetical case (see Chart 13–12). Theoretically, if $X = 4$, then an exact Y of 100 could be predicted with 100 percent confidence. Or if $X = 12$, then $Y = 300$. Because there is no difference between the observed values and the predicted values, there is no error in this estimate.

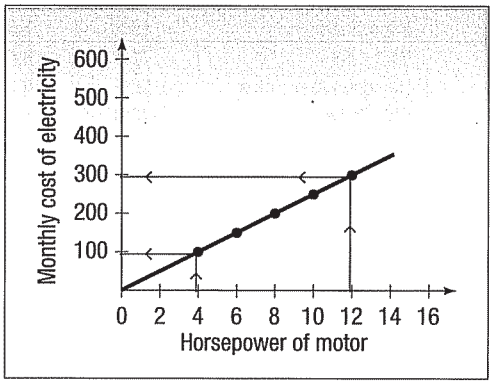


CHART 13–12 Example of Perfect Prediction: Horsepower and Cost of Electricity

Perfect prediction unrealistic in business

Perfect prediction in economics and business is practically impossible. For example, the revenue for the year from gasoline sales (Y) based on the number of automobile registrations (X) as of a certain date could no doubt be closely approximated, but the prediction would not be exact to the nearest dollar, or probably even to the nearest thousand dollars. Even predictions of tensile strength of steel wires based on the outside diameters of the wires are not always exact due to slight differences in the composition of the steel.

What is needed, then, is a measure that describes how precise the prediction of Y is based on X or, conversely, how inaccurate the estimate might be. This measure is called the **standard error of estimate.** The standard error of estimate, symbolized by $s_{y \cdot x}$, is the same concept as the standard deviation discussed in Chapter 3. The standard deviation measures the dispersion around the mean. The standard error of estimate measures the dispersion about the regression line.

STANDARD ERROR OF ESTIMATE A measure of the scatter, or dispersion, of the observed values around the line of regression.

The standard error of estimate is found by the following equation. Note that the equation is quite similar to the one for the standard deviation of a sample.

STANDARD ERROR OF ESTIMATE

$$s_{y\cdot x} = \sqrt{\frac{\Sigma(Y - Y')^2}{n - 2}} \qquad [13\text{–}6]$$

The standard deviation is based on the squared deviations from the mean, whereas the standard error of estimate is based on squared deviations between each Y and its predicted value, Y'. Remember that the regression line represents all the values of Y'. If $s_{y\cdot x}$ is small, this means that the data are relatively close to the regression line and the regression equation can be used to predict Y with little error. If $s_{y\cdot x}$ is large, this means that the data are widely scattered around the regression line and the regression equation will not provide a precise estimate Y.

EXAMPLE

Recall the example involving Copier Sales of America. The sales manager determined the least squares regression equation to be $Y' = 18.9476 + 1.1842X$, where Y' refers to the predicted number of copiers sold and X the number of sales calls made. Determine the standard error of estimate as a measure of how well the values fit the regression line.

SOLUTION

To find the standard error, we begin by finding the difference between the value, Y, and the value estimated from the regression equation, Y'. Next we square this difference, that is, $(Y - Y')^2$. We do this for each of the n observations and sum the results. That is, we compute $\Sigma(Y - Y')^2$, which is the numerator of formula (13–6). Finally, we divide by the number of observations minus 2. Why minus 2? We lose a degree of freedom each for estimating the intercept value, a, and the slope value, b. The details of the calculations are summarized in Table 13–4.

TABLE 13–4 Computations Needed for the Standard Error of Estimate

Sales Representative	Actual Sales (Y)	Estimated Sales (Y')	Deviation ($Y - Y'$)	Deviation Squared $(Y - Y')^2$
Tom Keller	30	42.6316	−12.6316	159.557
Jeff Hall	60	66.3156	−6.3156	39.887
Brian Virost	40	42.6316	−2.6316	6.925
Greg Fish	60	54.4736	5.5264	30.541
Susan Welch	30	30.7896	−0.7896	0.623
Carlos Ramirez	40	30.7896	9.2104	84.831
Rich Niles	40	42.6316	−2.6316	6.925
Mike Kiel	50	42.6316	7.3684	54.293
Mark Reynolds	30	42.6316	−12.6316	159.557
Soni Jones	70	54.4736	15.5264	241.069
			0.0000	784.211

The standard error of estimate is 9.901, found by using formula (13–6).

$$s_{y\cdot x} = \sqrt{\frac{\Sigma(Y - Y')^2}{n - 2}} = \sqrt{\frac{784.211}{10 - 2}} = 9.901$$

The deviations $(Y - Y')$ are the vertical deviations from the regression line. To illustrate, the 10 deviations from Table 13–4 are shown in Chart 13–13. Note in Table 13–4 that the sum of the signed deviations is zero. This indicates that the positive deviations (above the regression line) are offset by the negative deviations (below the regression line).

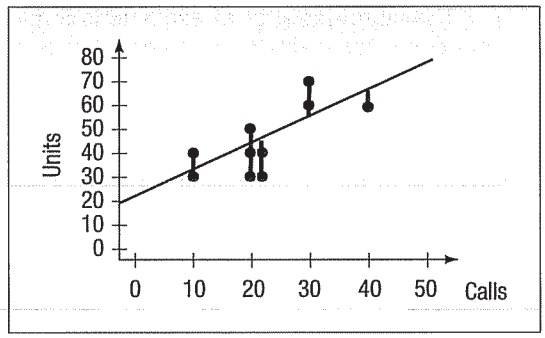
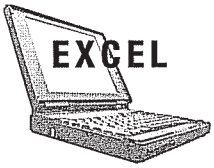
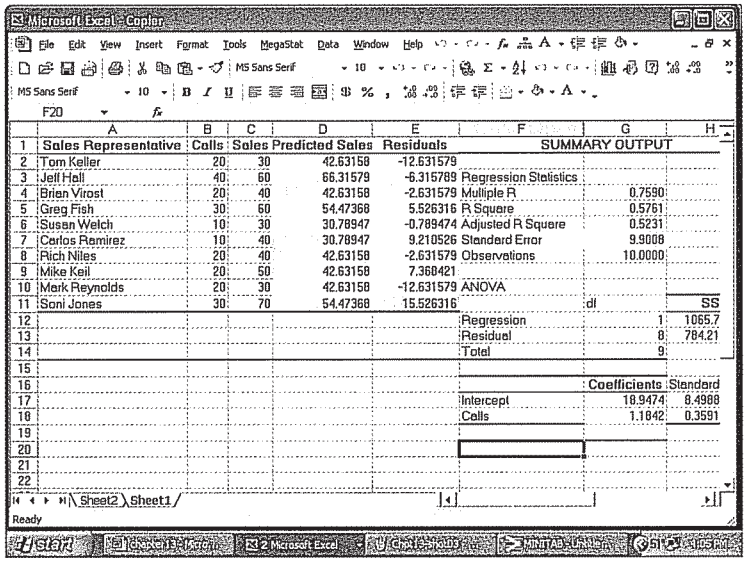


CHART 13–13 Sales Calls and Copiers Sold for 10 Salespeople

Software eases computation when you are finding the least squares regression line, calculating fitted values, or finding the standard error. The Excel output from the Copier Sales of America example is included below. The slope and intercept are in the column "Coefficients" (cells G17 and G18). The fitted values for each sales representative are the column "Predicted Sales" (cells D2:D11). The "Residuals" or differences between the actual and the estimated values are in the next column (cells E2:E11). The standard error of estimate is in cell G7. All of these values are highlighted below.





Microsoft Excel - Copier

File Edit View Insert Format Tools MegaStat Data Window Help

MS Sans Serif 10

F20

	A	B	C	D	E	F	G	H
1	Sales Representative	Calls	Sales	Predicted Sales	Residuals		SUMMARY OUTPUT	
2	Tom Keller	20	30	42.63158	-12.631579			
3	Jeff Hall	40	60	66.31579	-6.315789	Regression Statistics		
4	Brian Virost	20	40	42.63158	-2.631579	Multiple R	0.7590	
5	Greg Fish	30	60	54.47368	5.526316	R Square	0.5761	
6	Susan Welch	10	30	30.78947	-0.789474	Adjusted R Square	0.5231	
7	Carlos Ramirez	10	40	30.78947	9.210526	Standard Error	9.9008	
8	Rich Niles	20	40	42.63158	-2.631579	Observations	10.0000	
9	Mike Keil	20	50	42.63158	7.368421			
10	Mark Reynolds	20	30	42.63158	-12.631579	ANOVA		
11	Soni Jones	30	70	54.47368	15.526316		df	SS
12						Regression	1	1065.7
13						Residual	8	784.21
14						Total	9	
15								
16							Coefficients	Standard
17						Intercept	18.9474	8.4988
18						Calls	1.1842	0.3591
19								
20								
21								
22								

Sheet2 Sheet1

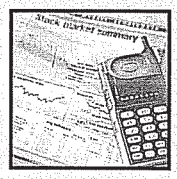
Ready

Thus far we have presented linear regression only as a descriptive tool. In other words it is a simple summary ($Y' = a + bX$) of the relationship between the dependent Y variable and the independent X variable. When our data is a sample taken from a population, we are doing inferential statistics. Then we need to recall the distinction between population parameters and sample statistics. In this case, we "model" the linear relationship in the population by the equation:

$$Y = \alpha + \beta X$$

where:

Y is any value of the dependent variable.
α is the Y-intercept (the value of Y when $X = 0$) in the population.



Statistics in Action

Studies indicate that for both men and women, those who are considered good looking earn higher wages than those who are not. In addition, for men there is a correlation between height and salary. For each additional inch of height, a man can expect to earn an additional $250 per year. So a man 6′6″ tall receives a $3,000 "stature" bonus over his 5′6″ counterpart. Being overweight or underweight is also related to earnings, particularly among women. A study of young women showed the heaviest 10 percent earned about 6 percent less than their lighter counterparts.

β is the slope (the amount by which Y changes when X increases by one unit) of the population line.
X is any value of the independent variable.

Now α and β are population parameters and a and b, respectively, are estimates of those parameters. They are computed from a particular sample taken from the population. Fortunately, the formulas given earlier in the chapter for a and b do not change when we move from using regression as a descriptive tool to regression in statistical inference.

It should be noted that the linear regression equation for the sample of salespeople is only an estimate of the relationship between the two variables for the population. Thus, the values of a and b in the regression equation are usually referred to as the **estimated regression coefficients,** or simply the **regression coefficients.**

Assumptions Underlying Linear Regression

To properly apply linear regression, several assumptions are necessary. Chart 13–14 illustrates these assumptions.

1. For each value of X, there is a group of Y values. These Y values follow the normal distribution.
2. The means of these normal distributions lie on the regression line.
3. The standard deviations of these normal distributions are all the same. The best estimate we have of this common standard deviation is the standard error of estimate ($s_{y \cdot x}$).
4. The Y values are statistically independent. This means that in selecting a sample a particular X does not depend on any other value of X. This assumption is particularly important when data are collected over a period of time. In such situations, the errors for a particular time period are often correlated with those of other time periods.

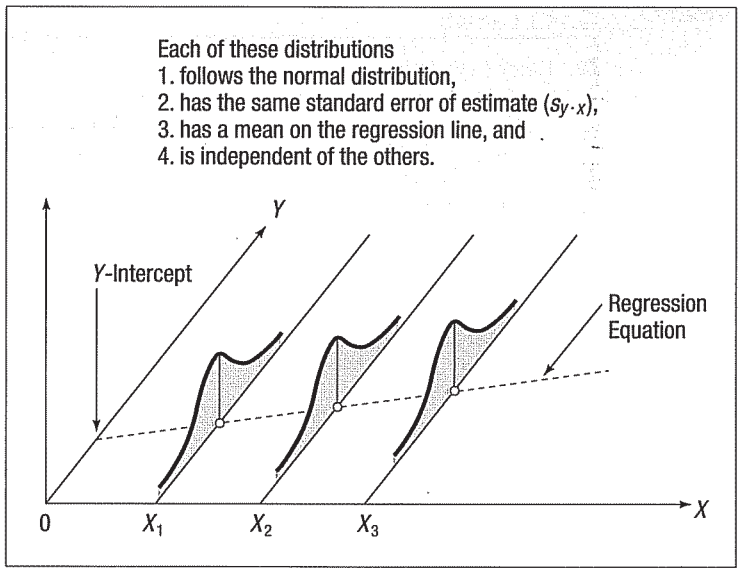


CHART 13–14 Regression Assumptions Shown Graphically

Recall from Chapter 7 that if the values follow a normal distribution, then the mean plus or minus one standard deviation will encompass 68 percent of the observations,

the mean plus or minus two standard deviations will encompass 95 percent of the observations, and the mean plus or minus three standard deviations will encompass virtually all of the observations. The same relationship exists between the predicted values Y' and the standard error of estimate ($s_{y \cdot x}$).

1. $Y' \pm s_{y \cdot x}$ will include the middle 68 percent of the observations.
2. $Y' \pm 2s_{y \cdot x}$ will include the middle 95 percent of the observations.
3. $Y' \pm 3s_{y \cdot x}$ will include virtually all the observations.

We can now relate these assumptions to Copier Sales of America, where we studied the relationship between the number of sales calls and the number of copiers sold. Assume that we took a much larger sample than $n = 10$, but that the standard error of estimate was still 9.901. If we drew a parallel line 9.901 units above the regression line and another 9.901 units below the regression line, about 68 percent of the points would fall between the two lines. Similarly, a line 19.802 [$2s_{y \cdot x} = 2(9.901)$] units above the regression line and another 19.802 units below the regression line should include about 95 percent of the data values.

As a rough check, refer to the second column from the right in Table 13–4 on page 393, i.e., the column headed "Deviation." Three of the 10 deviations exceed one standard error of estimate. That is, the deviation of −12.6316 for Tom Keller, −12.6316 for Mark Reynolds, and +15.5264 for Soni Jones all exceed the value of 9.901, which is one standard error from the regression line. All of the values are within 19.802 units of the regression line. To put it another way, 7 of the 10 deviations in the sample are within one standard error of the regression line and all are within two—a good result for a relatively small sample.

Self-Review 13–4 Refer to Self-Reviews 13–1 and 13–3, where the owner of Haverty's Furniture was studying the relationship between sales and the amount spent on advertising. Determine the standard error of estimate.

Exercises

19. Refer to Exercise 11.
- **a.** Determine the standard error of estimate.
- **b.** Suppose a large sample is selected (instead of just five). About 68 percent of the predictions would be between what two values?

20. Refer to Exercise 12.
- **a.** Determine the standard error of estimate.
- **b.** Suppose a large sample is selected (instead of just eight). About 95 percent of the predictions would be between what two values?

21. Refer to Exercise 13.
- **a.** Determine the standard error of estimate.
- **b.** Suppose a large sample is selected (instead of just 10). About 95 percent of the predictions regarding kilowatt-hours would occur between what two values?

22. Refer to Exercise 14.
- **a.** Determine the standard error of estimate.
- **b.** Suppose a large sample is selected (instead of just 10). About 95 percent of the predictions regarding sales would occur between what two values?

23. Refer to Exercise 5. Determine the standard error of estimate.

24. Refer to Exercise 6. Determine the standard error of estimate.

Confidence and Prediction Intervals

The standard error of estimate is also used to establish confidence intervals when the sample size is large and the scatter around the regression line approximates the

normal distribution. In our example involving the number of sales calls and the number of copiers sold, the sample size is small; hence, we need a correction factor to account for the size of the sample. In addition, when we move away from the mean of the independent variable, our estimates are subject to more variation, and we also need to adjust for this.

We are interested in providing interval estimates of two types. The first, which is called a **confidence interval,** reports the *mean* value of Y for a given X. The second type of estimate is called a **prediction interval,** and it reports the *range of values* of Y for a *particular* value of X. To explain further, suppose we estimate the salary of executives in the retail industry based on their years of experience. If we want an interval estimate of the mean salary of *all* retail executives with 20 years of experience, we calculate a confidence interval. If we want an estimate of the salary of Curtis Bender, a particular retail executive with 20 years of experience, we calculate a prediction interval.

To determine the confidence interval for the mean value of Y for a given X, the formula is:

CONFIDENCE INTERVAL FOR THE MEAN OF *Y*, GIVEN *X*.

$$Y' \pm t s_{y \cdot x} \sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma (X - \bar{X})^2}} \qquad [13\text{–}7]$$

where:

Y' is the predicted value for any selected X value.
X is any selected value of X.
$\bar{X}$ is the mean of the Xs, found by $\Sigma X/n$.
n is the number of observations.
$s_{y \cdot x}$ is the standard error of estimate.
t is the value of t from Appendix F with $n - 2$ degrees of freedom.

We first described the t distribution in Chapter 9. In review the concept of t was developed by William Gossett in the early 1900s. He noticed that $\bar{X} \pm z(s)$ was not precisely correct for small samples. He observed, for example, for degrees of freedom of 120, that 95 percent of the items fell within $\bar{X} \pm 1.98s$ instead of $\bar{X} \pm 1.96s$. This difference is not too critical, but note what happens as the sample size becomes smaller:

df	*t*
120	1.980
60	2.000
21	2.080
10	2.228
3	3.182

This is logical. The smaller the sample size, the larger the possible error. The increase in the t value compensates for this possibility.

EXAMPLE

We return to the Copier Sales of America illustration. Determine a 95 percent confidence interval for all sales representatives who make 25 calls and for Sheila Baker, a West Coast sales representative who made 25 calls.

SOLUTION

We use formula (13–7) to determine a confidence interval. Table 13–5 includes the necessary totals and a repeat of the information of Table 13–2 on page 379.

TABLE 13–5 Calculations Needed for Determining the Confidence Interval and Prediction Interval

Sales Representative	Sales Calls (X)	Copier Sales (Y)	$(X - \bar{X})$	$(X - \bar{X})^2$
Tom Keller	20	30	−2	4
Jeff Hall	40	60	18	324
Brian Virost	20	40	−2	4
Greg Fish	30	60	8	64
Susan Welch	10	30	−12	144
Carlos Ramirez	10	40	−12	144
Rich Niles	20	40	−2	4
Mike Kiel	20	50	−2	4
Mark Reynolds	20	30	−2	4
Soni Jones	30	70	8	64
			0	760

The first step is to determine the number of copiers we expect a sales representative to sell if he or she makes 25 calls. It is 48.5526, found by $Y' = 18.9476 + 1.1842X = 18.9476 + 1.1842(25)$.

To find the t value, we need to first know the number of degrees of freedom. In this case the degrees of freedom is $n - 2 = 10 - 2 = 8$. We set the confidence level at 95 percent. To find the value of t, move down the left-hand column to 8 degrees of freedom, then move across to the column with the 95 percent level of confidence. The value of t is 2.306.

In the previous section we calculated the standard error of estimate to be 9.901. We let $X = 25$, $\bar{X} = \Sigma X/n = 220/10 = 22$, and from Table 13–5 $\Sigma(X - \bar{X})^2 = 760$. Inserting these values in formula (13–7), we can determine the confidence interval.

$$\text{Confidence Interval} = Y' \pm ts_{y \cdot x}\sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma(X - \bar{X})^2}}$$

$$= 48.5526 \pm 2.306(9.901)\sqrt{\frac{1}{10} + \frac{(25 - 22)^2}{760}}$$

$$= 48.5526 \pm 7.6356$$

Thus, the 95 percent confidence interval for all sales representatives who make 25 calls is from 40.9170 up to 56.1882. To interpret, let's round the values. If a sales representative makes 25 calls, he or she can expect to sell 48.6 copiers. It is likely those sales will range from 40.9 to 56.2 copiers.

To determine the prediction interval for a particular value of Y for a given X, formula (13–7) is modified slightly: A 1 is added under the radical. The formula becomes:

PREDICTION INTERVAL FOR Y, GIVEN X

$$Y' \pm ts_{y \cdot x}\sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma(X - \bar{X})^2}} \qquad [13\text{–}8]$$

Suppose we want to estimate the number of copiers sold by Sheila Baker, who made 25 sales calls. The 95 percent prediction interval is determined as follows:

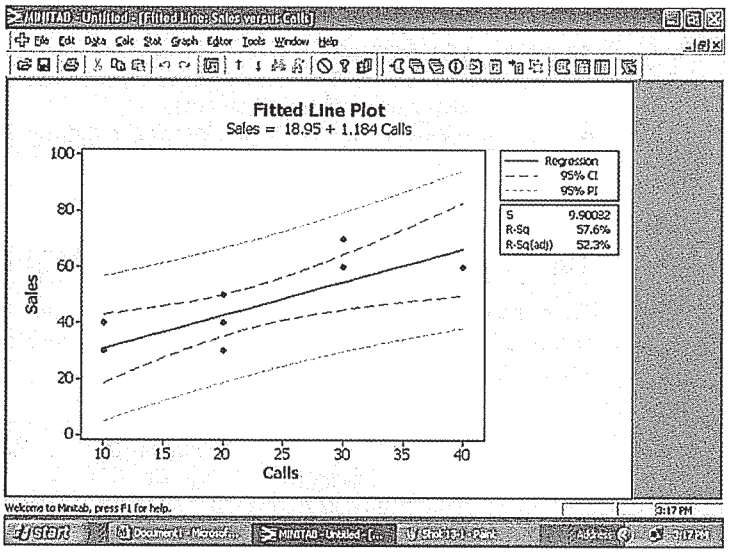
$$\text{Prediction Interval} = Y' \pm ts_{y\cdot x}\sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma(X - \bar{X})^2}}$$

$$= 48.5526 \pm 2.306(9.901)\sqrt{1 + \frac{1}{10} + \frac{(25 - 22)^2}{760}}$$

$$= 48.5526 \pm 24.0746$$

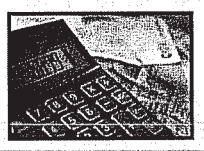
Thus, the interval is from 24.478 up to 72.627 copiers. We conclude that the number of copiers sold will be between about 24 and 73 for a particular sales representative. This interval is quite large. It is much larger than the confidence interval for all sales representatives who made 25 calls. It is logical, however, that there should be more variation in the sales estimate for an individual than for a group.

The following MINITAB graph shows the relationship between the regression line (in the center), the confidence interval (dashed lines), and the prediction interval (dotted lines). The bands for the prediction interval are always further from the regression line than for the confidence interval. Also, as the values of X move away from the mean number of calls (22) in either the positive or the negative direction the confidence interval and prediction interval bands widen. This is caused by the numerator of the right-hand term under the radical in formulas (13–7) and (13–8). That is, as the term $(X - \bar{X})^2$ increases, the widths of the confidence interval and the prediction interval also increase. To put it another way, there is less precision in our estimates as we move away, in either direction, from the mean of the independent variable.



We wish to emphasize again the distinction between a confidence interval and a prediction interval. A confidence interval refers to all cases with a given value of X and is computed by formula (13–7). A prediction interval refers to a particular case for a given value of X and is computed using formula (13–8). The prediction interval will always be wider because of the extra 1 under the radical in the second equation.

Self-Review 13–5



Refer to the sample data in Self-Reviews 13–1, 13–3, and 13–4, where the owner of Haverty's Furniture was studying the relationship between sales and the amount spent on advertising. The sales information for the last four months is repeated below.

Month	Advertising Expense ($ million)	Sales Revenue ($ million)
July	2	7
August	1	3
September	3	8
October	4	10

The regression equation was computed to be $Y' = 1.5 + 2.2X$, and the standard error is 0.9487. Both variables are reported in millions of dollars. Determine the 90 percent confidence interval for the typical sales revenue for a month in which $3 million was spent on advertising.

Exercises

25. Refer to Exercise 11.
- **a.** Determine the .95 confidence interval for the mean predicted when $X = 7$.
- **b.** Determine the .95 prediction interval for an individual predicted when $X = 7$.

26. Refer to Exercise 12.
- **a.** Determine the .95 confidence interval for the mean predicted when $X = 7$.
- **b.** Determine the .95 prediction interval for an individual predicted when $X = 7$.

27. Refer to Exercise 13.
- **a.** Determine the .95 confidence interval, in thousands of kilowatt-hours, for the mean of all six-room homes.
- **b.** Determine the .95 prediction interval, in thousands of kilowatt-hours, for a particular six-room home.

28. Refer to Exercise 14.
- **a.** Determine the .95 confidence interval, in thousands of dollars, for the mean of all sales personnel who make 40 contacts.
- **b.** Determine the .95 prediction interval, in thousands of dollars, for a particular salesperson who makes 40 contacts.

More on the Coefficient of Determination

To further examine the basic concept of the coefficient of determination, suppose there is interest in the relationship between years on the job, X, and weekly production, Y. Sample data revealed:

Employee	Years on Job, X	Weekly Production, Y
Gordon	14	6
James	7	5
Ford	3	3
Salter	15	9
Artes	11	7

The sample data were plotted in a scatter diagram. Since the relationship between X and Y appears to be linear, a line was drawn through the plots (see Chart 13–15). The equation is $Y' = 2 + 0.4X$.

Note in Chart 13–15 that if we were to use that line to predict weekly production for an employee, in no case would our prediction be exact. That is, there would be some error in each of our predictions. As an example, for Gordon, who has been with the company 14 years, we would predict weekly production to be 7.6 units; however, he produces only 6 units.

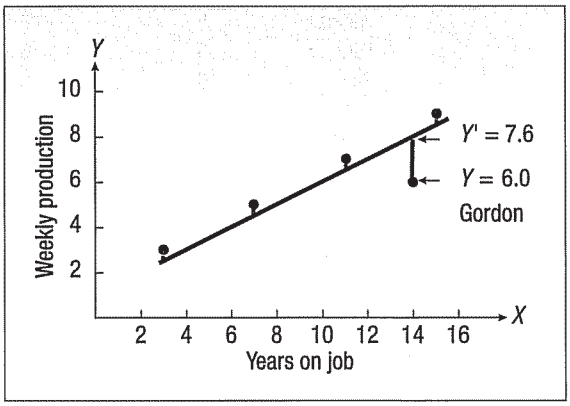


CHART 13–15 Observed Data and the Least Squares Line

Unexplained variation

To measure the overall error in our prediction, every deviation from the line is squared and the squares summed. The predicted point on the line is designated Y', read Y prime, and the observed point is designated Y. For Gordon, $(Y - Y')^2 = (6 - 7.6)^2 = (-1.6)^2 = 2.56$. Logically, this variation cannot be explained by the independent variable, so it is referred to as the *unexplained variation.* Specifically, we cannot explain why Gordon's production of 6 units is 1.6 units below his predicted production of 7.6 units, based on the number of years he has been on the job.

The sum of the squared deviations, $\Sigma(Y - Y')^2$, is 4.00. (See Table 13–6.) The term $\Sigma(Y - Y')^2 = 4.00$ is the variation in Y (production) that cannot be predicted from X. It is the "unexplained" variation in Y.

TABLE 13–6 Computations Needed for the Unexplained Variation

	X	Y	Y'	$Y - Y'$	$(Y - Y')^2$
Gordon	14	6	7.6	−1.6	2.56
James	7	5	4.8	0.2	0.04
Ford	3	3	3.2	−0.2	0.04
Salter	15	9	8.0	1.0	1.00
Artes	11	7	6.4	0.6	0.36
Total	50	30		0.0*	4.00

*Must be 0.

Total variation in Y

Now suppose *only* the Y values (weekly production, in this problem) are known and we want to predict production for every employee. The actual production figures for the employees are 6, 5, 3, 9, and 7 (from Table 13–6). To make these predictions, we could assign the mean weekly production (6 units, found by $\Sigma Y/n = 30/5 = 6$) to each employee. This would keep the sum of the squared prediction errors at a minimum. (Recall from Chapter 3 that the sum of the squared deviations from the arithmetic mean for a set of numbers is smaller than the sum of the squared deviations from any other value, such as the median.) Table 13–7 shows the necessary calculations. The sum of the squared deviations is 20, as shown in Table 13–7. The value 20 is referred to as the *total variation in Y.*

TABLE 13–7 Calculations Needed for the Total Variation in Y

Name	Weekly Production, Y	Mean Weekly Production, $\bar{Y}$	$Y - \bar{Y}$	$(Y - \bar{Y})^2$
Gordon	6	6	0	0
James	5	6	−1	1
Ford	3	6	−3	9
Salter	9	6	3	9
Artes	7	6	1	1
Total			0*	20

*Must be 0.

What we did to arrive at the total variation in Y is shown diagrammatically in Chart 13–16.

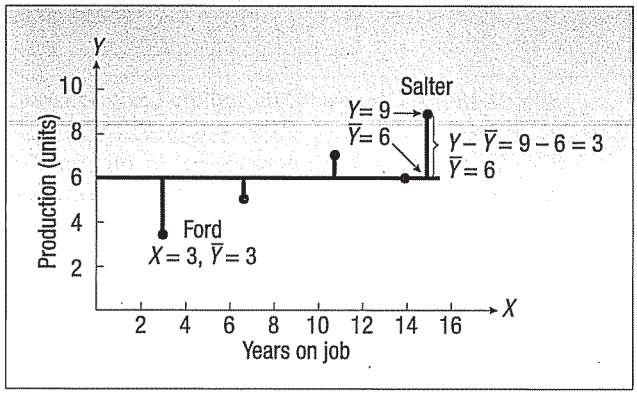


CHART 13–16 Plots Showing Deviations from the Mean of Y

Logically, the total variation in Y can be subdivided into unexplained variation and explained variation. To arrive at the explained variation, since we know the total variation and unexplained variation, we simply subtract: Explained variation = Total variation − Unexplained variation. Dividing the explained variation by the total variation gives the coefficient of determination, r^2, which is a proportion. In terms of a formula:

COEFFICIENT OF DETERMINATION

$$r^2 = \frac{\text{Total variation} - \text{Unexplained variation}}{\text{Total variation}} = \frac{\Sigma(Y - \bar{Y})^2 - \Sigma(Y - Y')^2}{\Sigma(Y - \bar{Y})^2} \qquad [13\text{–}9]$$

In this problem:

$$r^2 = \frac{20 - 4}{20} = \frac{16}{20} = .80$$

Table 13–7 → 20 (Total variation)
Table 13–6 → 4 (Unexplained variation)
16 = Explained variation

As mentioned, .80 is a proportion. It is not a probability. We say that 80 percent of the variation in weekly production, Y, is determined, or accounted for, by its linear relationship with X (years on the job).

As a check, formula (13–1) for the coefficient of correlation could be used. Squaring r gives the coefficient of determination. Exercise 29 offers a check on the preceding problem.

Exercises

29. Using the preceding problem, involving years on the job and weekly production, verify that the coefficient of determination is in fact .80.

30. The number of shares of Icom, Inc., turned over during a month, and the price at the end of the month, are listed in the following table. Also given are the Y' values.

Turnover (thousands of shares), X	Actual Price, Y	Estimated Price, Y'
4	$2	$2.7
1	1	0.6
5	4	3.4
3	2	2.0
2	1	1.3

a. Draw a scatter diagram. Plot a line through the dots.
b. Compute the coefficient of determination using formula (13–10).
c. Interpret the coefficient of determination.

The Relationships among the Coefficient of Correlation, the Coefficient of Determination, and the Standard Error of Estimate

In an earlier section, we discussed the standard error of estimate, which measures how close the actual values are to the regression line. When the standard error is small, it indicates that the two variables are closely related. In the calculation of the standard error, the key term is $\Sigma(Y - Y')^2$. If the value of this term is small, then the standard error will also be small.

The correlation coefficient measures the strength of the linear association between two variables. When the points on the scatter diagram appear close to the line, we note that the correlation coefficient tends to be large. Thus, the standard error of estimate and the coefficient of correlation relate the same information but use a different scale to report the strength of the association. However, both measures involve the term $\Sigma(Y - Y')^2$.

We also noted that the square of the correlation coefficient is the coefficient of determination. The coefficient of determination measures the percent of the variation in Y that is explained by the variation in X.

A convenient vehicle for showing the relationship among these three measures is an ANOVA table. This table is similar to the analysis of variance table developed in Chapter 12. In that chapter, the total variation was divided into two components: that due to the *treatments* and that due to *random error.* The concept is similar in regression

analysis. The total variation, $\Sigma(Y - Y')^2$, is divided into two components: (1) that explained by the *regression* (explained by the independent variable) and (2) the *error,* or unexplained variation. These two categories are identified in the first column of the ANOVA table that follows. The column headed "df" refers to the degrees of freedom associated with each category. The total number of degrees of freedom is $n - 1$. The number of degrees of freedom in the regression is 1, since there is only one independent variable. The number of degrees of freedom associated with the error term is $n - 2$. The term "SS" located in the middle of the ANOVA table refers to the sum of squares—the variation. The terms are computed as follows:

$$\text{Regression} = \text{SSR} = \Sigma(Y' - \bar{Y})^2$$
$$\text{Error variation} = \text{SSE} = \Sigma(Y - Y')^2$$
$$\text{Total variation} = \text{SS total} = \Sigma(Y - \bar{Y})^2$$

The format for the ANOVA table is:

Source	df	SS	MS
Regression	1	SSR	SSR/1
Error	$n - 2$	SSE	SSE/($n - 2$)
Total	$n - 1$	SS total*	

*SS total = SSR + SSE.

The coefficient of determination, r^2, can be obtained directly from the ANOVA table by:

COEFFICIENT OF DETERMINATION

$$r^2 = \frac{\text{SSR}}{\text{SS total}} = 1 - \frac{\text{SSE}}{\text{SS total}} \qquad \textbf{[13–10]}$$

The term "SSR/SS total" is the proportion of the variation in *Y explained* by the independent variable, *X.* Note the effect of the SSE term on r^2. As SSE decreases, r^2 will increase. Conversely, as the standard error decreases, the r^2 term increases.

The standard error of estimate can also be obtained from the ANOVA table using the following equation:

STANDARD ERROR OF ESTIMATE

$$s_{y \cdot x} = \sqrt{\frac{\text{SSE}}{n - 2}} \qquad \textbf{[13–11]}$$

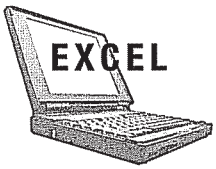
The Copier Sales of America example is used to illustrate the computations of the coefficient of determination and the standard error of estimate from an ANOVA table.

EXAMPLE

In the Copier Sales of America example we studied the relationship between the number of sales calls made and the number of copiers sold. Use a computer software package to determine the least squares regression equation and the ANOVA table. Identify the regression equation, the standard error of estimate, and the coefficient of determination on the computer output. From the ANOVA table on the computer output, determine the coefficient of determination and the standard error of estimate using formulas (13–10) and (13–11).

SOLUTION

The output from Excel follows.



Microsoft Excel - Copier-2

	A	B	C	D	E	F	G	H	I
1	Sales Representative	Calls	Sales			SUMMARY OUTPUT			
2	Tom Keller	20	30						
3	Jeff Hall	40	60		Regression Statistics				
4	Brian Virost	20	40		Multiple R	0.759			
5	Greg Fish	30	60		R Square	0.576			
6	Susan Welch	10	30		Adjusted R Square	0.523			
7	Carlos Ramirez	10	40		Standard Error	9.901			
8	Rich Niles	20	40		Observations	10.000			
9	Mike Keil	20	50						
10	Mark Reynolds	20	30		ANOVA				
11	Soni Jones	30	70			df	SS	MS	F
12					Regression	1	1066	1065.769474	10.87248322
13					Residual	8	784	98.02631579	
14					Total	9	1850		

Sheet1 / Sheet2 / Sheet3

Ready

From formula (13–10) the coefficient of determination is .576, found by

$$r^2 = \frac{\text{SSR}}{\text{SS total}} = \frac{1{,}066}{1{,}850} = .576$$

This is the same value we computed earlier in the chapter, when we found the coefficient of determination by squaring the coefficient of correlation. Again, the interpretation is that the independent variable, *Calls,* explains 57.6 percent of the variation in the number of copiers sold. If we needed the coefficient of correlation, we could find it by taking the square root of the coefficient of determination:

$$r = \sqrt{r^2} = \sqrt{.576} = .759$$

A problem does remain, and that involves the sign for the coefficient of correlation. Recall that the square root of a value could have either a positive or a negative sign. The sign of the coefficient of correlation will always be the same as that of the slope. That is, b and r will always have the same sign. In this case the sign is positive, so the coefficient of correlation is .759.

To find the standard error of estimate, we use formula (13–11):

$$s_{y \cdot x} = \sqrt{\frac{\text{SSE}}{n-2}} = \sqrt{\frac{784.2}{10-2}} = 9.901$$

Again, this is the same value calculated earlier in the chapter. These values are identified on the Excel computer output.

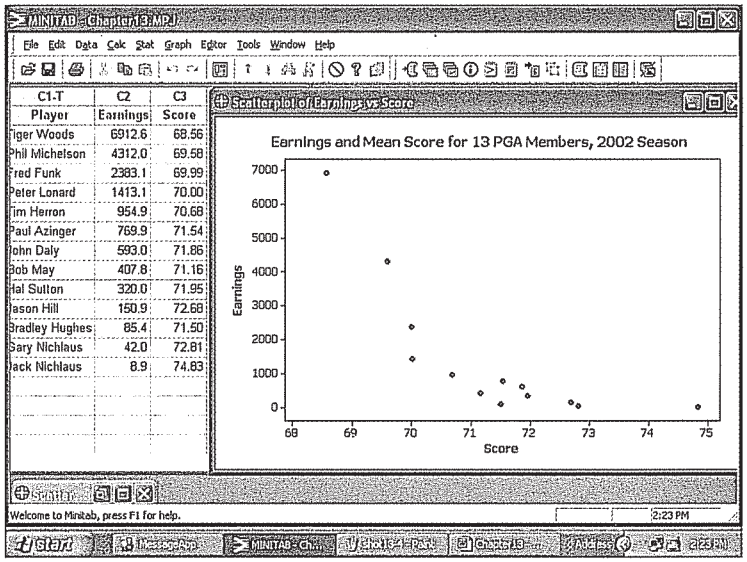
Transforming Data

The coefficient of correlation describes the strength of the *linear* relationship between two variables. It could be that two variables are closely related, but this relationship is not linear. Be cautious when you are interpreting the coefficient of correlation. A value of r may indicate there is no linear relationship, but it could be there is a relationship of some other nonlinear or curvilinear form. To explain, below is a listing of 13 professional golfers, the amount they earned during the 2002 season, and their mean score

per round. (In golf, the objective is to play 18 holes in the least number of strokes. So, lower mean scores are related to the higher earnings.)

Player	Earnings ($000)	Mean score
Tiger Woods	6,912.6	68.56
Phil Michelson	4,312.0	69.58
Fred Funk	2,383.1	69.99
Peter Lonard	1,413.1	70.00
Tim Herron	954.9	70.68
Paul Azinger	769.9	71.54
John Daly	593.0	71.86
Bob May	407.8	71.16
Hal Sutton	320.0	71.95
Jason Hill	150.9	72.68
Bradley Hughes	85.4	71.50
Gary Nicklaus	42.0	72.81
Jack Nicklaus	8.9	74.83

For the above golf data the correlation between the variables, earnings and score, shows a fairly strong negative relationship. The correlation is −0.782, but when we use a scatter diagram to plot the data the relationship appears to be nonlinear. That is, the relationship does not follow a straight line.



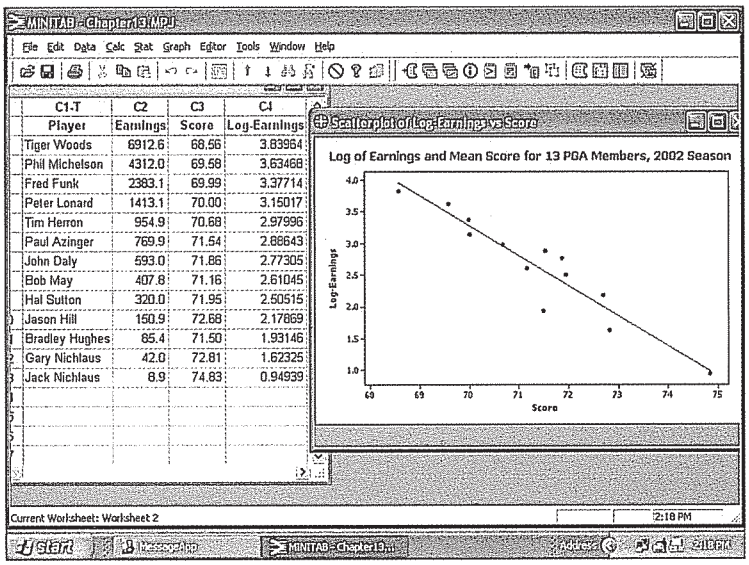
What can we do to explore other (nonlinear) relationships? One possibility is to transform one of the variables. For example, instead of using X as the independent variable we might use its square as the independent variable. Another possibility is to transform the dependent variable.

In the golf-earnings example, changing the scale of the dependent variable is effective. We use MINITAB to determine the log of each golfer's earnings and then find the correlation between the log of earnings and score. The coefficient of correlation increases to −0.943, which means 88.9 percent of the variation in the log of earnings is

accounted for by the independent variable score. Clearly as the mean score increases for a golfer, he can expect his earnings to decrease.

There is no commonly accepted procedure to determine which variable to transform or what transformation to use. So experience and trial and error are your guides. The most common types of transformations are:

- Take the log of one of the variables.
- Square one of the variables.
- Take the square root of one of the variables.
- Take the reciprocal of one of the variables.



	C1-T	C2	C3	C4
	Player	Earnings	Score	Log-Earnings
1	Tiger Woods	6912.6	68.56	3.83964
2	Phil Michelson	4312.0	69.58	3.63468
3	Fred Funk	2383.1	69.99	3.37714
4	Peter Lonard	1413.1	70.00	3.15017
5	Tim Herron	954.9	70.68	2.97996
6	Paul Azinger	769.9	71.54	2.88643
7	John Daly	593.0	71.86	2.77305
8	Bob May	407.8	71.16	2.61045
9	Hal Sutton	320.0	71.95	2.50515
10	Jason Hill	150.9	72.68	2.17869
11	Bradley Hughes	85.4	71.50	1.93146
12	Gary Nichlaus	42.0	72.81	1.62325
13	Jack Nichlaus	8.9	74.83	0.94939

Exercises

31. Given the following ANOVA table:

SOURCE	DF	SS	MS	F
Regression	1	1000.0	1000.00	26.00
Error	13	500.0	38.46	
Total	14	1500.0		

a. Determine the coefficient of determination.
b. Assuming a direct relationship between the variables, what is the coefficient of correlation?
c. Determine the standard error of estimate.

32. On the first statistics exam the coefficient of determination between the hours studied and the grade earned was 80 percent. The standard error of estimate was 10. There were 20 students in the class. Develop an ANOVA table.

33. The information listed below shows the relationship between the interest rate on home mortgages and the number of housing starts for selected periods. Observe that the relationship is inverse; that is, as the interest rate declines the number of housing starts increases.

Rate	Starts
11	9000
10	10000
9	24000
8	40000
7	52000
6	65000
5	80000
4	100000
3	130000
2	135000

a. Plot the above data in a scatter diagram. Can you confirm the inverse relationship?
b. Use statistical software to develop a regression equation. What is the coefficient of determination? What do you conclude about the strength of the relationship between the variables?
c. Estimate the number of housing starts when the interest rate is at 11 or 12 percent. Is this a reasonable conclusion?
d. Transform the data on the number of housing starts to the log of the number of starts. Use this transformed variable to develop a regression equation. How does this transformation affect the coefficient of determination? Is the estimated value of the number of housing starts more reasonable when the interest rate is 11 percent? Give the specific evidence.

34. According to basic economics as the demand for a product increases the price will decrease. Listed below is the number of units demanded and the price.

Demand	Price
2	$120.0
5	90.0
8	80.0
12	70.0
16	50.0
21	45.0
27	31.0
35	30.0
45	25.0
60	21.0

a. Determine the correlation between price and demand. Plot the data in a scatter diagram. Does the relationship seem to be linear?
b. Transform the price to a log to the base 10. Plot the log of the price and the demand. Determine the correlation coefficient. Does this seem to improve the relationship between the variables?

Chapter Outline

I. A scatter diagram is a graphic tool to portray the relationship between two variables.
  A. The dependent variable is scaled on the *Y*-axis and is the variable being estimated.
  B. The independent variable is scaled on the *X*-axis and is the variable used as the estimator.
II. The coefficient of correlation measures the strength of the linear association between two variables.
  A. Both variables must be at least the interval scale of measurement.
  B. The coefficient of correlation can range from −1.00 up to 1.00.
  C. If the correlation between two variables is 0, there is no association between them.
  D. A value of 1.00 indicates perfect positive correlation, and −1.00 perfect negative correlation.
  E. A positive sign means there is a direct relationship between the variables, and a negative sign means there is an inverse relationship.

F. It is designated by the letter r and found by the following equation:

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{(n - 1)\, s_x s_y} \qquad \textbf{[13–1]}$$

G. The following equation is used to determine whether the correlation in the population is different from 0.

$$t = \frac{r\sqrt{n - 2}}{\sqrt{1 - r^2}} \qquad \textbf{[13–2]}$$

III. The coefficient of determination is the fraction of the variation in one variable that is explained by the variation in the other variable.
- **A.** It ranges from 0 to 1.0.
- **B.** It is the square of the coefficient of correlation.

IV. In regression analysis we estimate one variable based on another variable.
- **A.** The relationship between the variables must be linear.
- **B.** Both the independent and the dependent variables must be interval or ratio scale.
- **C.** The least squares criterion is used to determine the regression equation.

V. The least squares regression line is of the form $Y' = a + bX$.
- **A.** Y' is the estimated value of Y for a selected value of X.
- **B.** b is the slope of the fitted line.
  1. It shows the amount of change in Y' for a change of one unit in X.
  2. A positive value for b indicates a direct relationship between the two variables, and a negative value an inverse relationship.
  3. The sign of b and the sign of r, the coefficient of correlation, are always the same.
  4. b is computed using the following equation.

$$b = r\frac{s_y}{s_x} \qquad \textbf{[13–4]}$$

- **C.** a is the constant or intercept.
  1. It is the value of Y' when $X = 0$.
  2. a is computed using the following equation.

$$a = \bar{Y} - b\bar{X} \qquad \textbf{[13–5]}$$

- **D.** X is the value of the independent variable.

VI. The standard error of estimate measures the variation around the regression line.
- **A.** It is in the same units as the dependent variable.
- **B.** It is based on squared deviations from the regression line.
- **C.** Small values indicate that the points cluster closely about the regression line.
- **D.** It is computed using the following formula.

$$s_{y \cdot x} = \sqrt{\frac{\Sigma(Y - Y')^2}{n - 2}} \qquad \textbf{[13–6]}$$

VII. Inference about linear regression is based on the following assumptions.
- **A.** For a given value of X, the values of Y are normally distributed about the line of regression.
- **B.** The standard deviation of each of the normal distributions is the same for all values of X and is estimated by the standard error of estimate.
- **C.** The deviations from the regression line are independent, with no pattern to the size or direction.

VIII. There are two types of interval estimates.
- **A.** In a confidence interval the mean value of Y is estimated for a given value of X.
  1. It is computed from the following formula.

$$Y' \pm t s_{y \cdot x} \sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma(X - \bar{X})^2}} \qquad \textbf{[13–7]}$$

  2. The width of the interval is affected by the level of confidence, the size of the standard error of estimate, and the size of the sample, as well as the value of the independent variable.
- **B.** In a prediction interval the individual value of Y is estimated for a given value of X.

1. It is computed from the following formula.

$$Y' \pm ts_{y \cdot x}\sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma(X - \bar{X})^2}} \qquad [13\text{–}8]$$

2. The difference between formulas (13–7) and (13–8) is the 1 under the radical.
   a. The prediction interval will be wider than the confidence interval.
   b. The prediction interval is also based on the level of confidence, the size of the standard error of estimate, the size of the sample, and the value of the independent variable.

Pronunciation Key

SYMBOL	MEANING	PRONUNCIATION
ΣXY	Sum of the products of X and Y	*Sum X Y*
ρ	Coefficient of correlation in the population	*Rho*
Y'	Estimated value of Y	*Y prime*
$s_{y \cdot x}$	Standard error of estimate	*s sub y dot x*
r^2	Coefficient of determination	*r square*

Chapter Exercises

35. A regional commuter airline selected a random sample of 25 flights and found that the correlation between the number of passengers and the total weight, in pounds, of luggage stored in the luggage compartment is 0.94. Using the .05 significance level, can we conclude that there is a positive association between the two variables?
36. A sociologist claims that the success of students in college (measured by their GPA) is related to their family's income. For a sample of 20 students, the coefficient of correlation is 0.40. Using the 0.01 significance level, can we conclude that there is a positive correlation between the variables?
37. An Environmental Protection Agency study of 12 automobiles revealed a correlation of 0.47 between the engine size and emissions. At the .01 significance level, can we conclude that there is a positive association between these variables? What is the *p*-value? Interpret.
38. A sample of 15 financial executives in the pharmaceutical industry revealed the correlation between the number of fat grams an executive consumed the previous day and that executive's cholesterol level was 0.345.
    a. Does cholesterol seem to increase for those who consumed more fat? How can you tell?
    b. How much of the variation in cholesterol level is accounted for by the number of fat grams consumed?
    c. At the .05 significance level is it reasonable to conclude there is a positive association between fat grams consumed and cholesterol level? What is the *p*-value?
39. A sample of 20 American cities showed the correlation between the population of the city and its unemployment rate was 0.237.
    a. Does unemployment rate seem to increase as the size of the population increases? How can you tell?
    b. How much of the variation in the unemployment rate is accounted for by the variation in the population?
    c. At the .01 significance level, is it reasonable to conclude there is positive association between the unemployment rate and the population?
40. Dr. Megan Boyle wishes to investigate the relationship between stress and job satisfaction. To begin she developed a profile for stress based on assigning points for events such as the death of a spouse, a change in sleeping habits, a change in eating habits, and the addition of a family member. The job satisfaction was also based on assigning points for salary, ability to get along with coworkers, and the job environment. Dr. Boyle sampled 25 workers in the technology sector and found correlation between the stress and job satisfaction was −0.536.
    a. Does job satisfaction seem to increase or decrease as stress increases? How can you tell?
    b. How much of the variation in stress is accounted for by the variation in job satisfaction?
    c. At the .05 significance level, is it reasonable to conclude there is negative association between stress and job satisfaction?

41. What is the relationship between the amount spent per week on food and the size of the family? Do larger families spend more on food? A sample of 10 families in the Chicago area revealed the following figures for family size and the amount spent on food per week.

Family Size	Amount Spent on Food	Family Size	Amount Spent on Food
3	$ 99	3	$111
6	104	4	74
5	151	4	91
6	129	5	119
6	142	3	91

   a. Compute the coefficient of correlation.
   b. Determine the coefficient of determination.
   c. Can we conclude that there is a positive association between the amount spent on food and the family size? Use the .05 significance level.

42. A sample of 12 homes sold last week in St. Paul, Minnesota is selected. Can we conclude that as the size of the home (reported below in thousands of square feet) increases, the selling price (reported in $ thousands) also increases?

Home Size (thousands of square feet)	Selling Price ($ thousands)	Home Size (thousands of square feet)	Selling Price ($ thousands)
1.4	100	1.3	110
1.3	110	0.8	85
1.2	105	1.2	105
1.1	120	0.9	75
1.4	80	1.1	70
1.0	105	1.1	95

   a. Compute the coefficient of correlation.
   b. Determine the coefficient of determination.
   c. Can we conclude that there is a positive association between the size of the home and the selling price? Use the .05 significance level.

43. The manufacturer of Cardio Glide exercise equipment wants to study the relationship between the number of months since the glide was purchased and the length of time the equipment was used last week.

Person	Months Owned	Hours Exercised	Person	Months Owned	Hours Exercised
Rupple	12	4	Massa	2	8
Hall	2	10	Sass	8	3
Bennett	6	8	Karl	4	8
Longnecker	9	5	Malrooney	10	2
Phillips	7	5	Veights	5	5

   a. Plot the information on a scatter diagram. Let hours of exercise be the dependent variable. Comment on the graph.
   b. Determine the coefficient of correlation. Interpret.
   c. At the .01 significance level, can we conclude that there is a negative association between the variables?

44. The following regression equation was computed from a sample of 20 observations:

$$Y' = 15 - 5X$$

SSE was found to be 100 and SS total 400.

a. Determine the standard error of estimate.
b. Determine the coefficient of determination.
c. Determine the coefficient of correlation. (Caution: Watch the sign!)

45. An ANOVA table is:

SOURCE	DF	SS	MS	F
Regression	1	50		
Error				
Total	24	500		

a. Complete the ANOVA table.
b. How large was the sample?
c. Determine the standard error of estimate.
d. Determine the coefficient of determination.

46. Following is a regression equation.

$$Y' = 17.08 + 0.16X$$

This information is also available: $s_{y \cdot x} = 4.05$, $\Sigma(X - \bar{X})^2 = 1030$, and $n = 5$.
a. Estimate the value of Y' when $X = 50$.
b. Develop a 95 percent prediction interval for an individual value of Y for $X = 50$.

47. The National Highway Association is studying the relationship between the number of bidders on a highway project and the winning (lowest) bid for the project. Of particular interest is whether the number of bidders increases or decreases the amount of the winning bid.

Project	Number of Bidders, X	Winning Bid ($ millions), Y	Project	Number of Bidders, X	Winning Bid ($ millions), Y
1	9	5.1	9	6	10.3
2	9	8.0	10	6	8.0
3	3	9.7	11	4	8.8
4	10	7.8	12	7	9.4
5	5	7.7	13	7	8.6
6	10	5.5	14	7	8.1
7	7	8.3	15	6	7.8
8	11	5.5			

a. Determine the regression equation. Interpret the equation. Do more bidders tend to increase or decrease the amount of the winning bid?
b. Estimate the amount of the winning bid if there were seven bidders.
c. A new entrance is to be constructed on the Ohio Turnpike. There are seven bidders on the project. Develop a 95 percent prediction interval for the winning bid.
d. Determine the coefficient of determination. Interpret its value.

48. Mr. William Profit is studying companies going public for the first time. He is particularly interested in the relationship between the size of the offering and the price per share. A sample of 15 companies that recently went public revealed the following information.

Company	Size ($ millions), X	Price per Share, Y	Company	Size ($ millions), X	Price per Share, Y
1	9.0	10.8	9	160.7	11.3
2	94.4	11.3	10	96.5	10.6
3	27.3	11.2	11	83.0	10.5
4	179.2	11.1	12	23.5	10.3
5	71.9	11.1	13	58.7	10.7
6	97.9	11.2	14	93.8	11.0
7	93.5	11.0	15	34.4	10.8
8	70.0	10.7			

a. Determine the regression equation.

b. Determine the coefficient of determination. Do you think Mr. Profit should be satisfied with using the size of the offering as the independent variable?

49. The Bardi Trucking Co., located in Cleveland, Ohio, makes deliveries in the Great Lakes region, the Southeast, and the Northeast. Jim Bardi, the president, is studying the relationship between the distance a shipment must travel and the length of time, in days, it takes the shipment to arrive at its destination. To investigate, Mr. Bardi selected a random sample of 20 shipments made last month. Shipping distance is the independent variable, and shipping time is the dependent variable. The results are as follows:

Shipment	Distance (miles)	Shipping Time (days)	Shipment	Distance (miles)	Shipping Time (days)
1	656	5	11	862	7
2	853	14	12	679	5
3	646	6	13	835	13
4	783	11	14	607	3
5	610	8	15	665	8
6	841	10	16	647	7
7	785	9	17	685	10
8	639	9	18	720	8
9	762	10	19	652	6
10	762	9	20	828	10

a. Draw a scatter diagram. Based on these data, does it appear that there is a relationship between how many miles a shipment has to go and the time it takes to arrive at its destination?

b. Determine the coefficient of correlation. Can we conclude that there is a positive correlation between distance and time? Use the .05 significance level.

c. Determine and interpret the coefficient of determination.

d. Determine the standard error of estimate.

50. Super Markets, Inc. is considering expanding into the Scottsdale, Arizona, area. Ms. Luann Miller, Director of Planning, must present an analysis of the proposed expansion to the operating committee of the board of directors. As a part of her proposal, she needs to include information on the amount people in the region spend per month for grocery items. She would also like to include information on the relationship between the amount spent for grocery items and income. She gathered the following sample information.

Household	Monthly Amount	Monthly Income	Household	Monthly Amount	Monthly Income
1	$555	$4,388	21	$ 913	$6,688
2	489	4,558	22	918	6,752
3	458	4,793	23	710	6,837
4	613	4,856	24	1,083	7,242
5	647	4,856	25	937	7,263
6	661	4,899	26	839	7,540
7	662	4,899	27	1,030	8,009
8	675	5,091	28	1,065	8,094
9	549	5,133	29	1,069	8,264
10	606	5,304	30	1,064	8,392
11	668	5,304	31	1,015	8,414
12	740	5,304	32	1,148	8,882
13	592	5,346	33	1,125	8,925
14	720	5,495	34	1,090	8,989
15	680	5,581	35	1,208	9,053
16	540	5,730	36	1,217	9,138
17	693	5,943	37	1,140	9,329
18	541	5,943	38	1,265	9,649
19	673	6,156	39	1,206	9,862
20	676	6,603	40	1,145	9,883

a. Let the amount spent be the dependent variable and monthly income the independent variable. Create a scatter diagram, using a software package.
b. Determine the regression equation. Interpret the slope value.
c. Determine the coefficient of correlation. Can you conclude that it is greater than 0?

51. Below is information on the price per share and the dividend for a sample of 30 companies.

Company	Price per Share	Dividend	Company	Price per Share	Dividend
1	$20.00	$ 3.14	16	$57.06	$ 9.53
2	22.01	3.36	17	57.40	12.60
3	31.39	0.46	18	58.30	10.43
4	33.57	7.99	19	59.51	7.97
5	35.86	0.77	20	60.60	9.19
6	36.12	8.46	21	64.01	16.50
7	36.16	7.62	22	64.66	16.10
8	37.99	8.03	23	64.74	13.76
9	38.85	6.33	24	64.95	10.54
10	39.65	7.96	25	66.43	21.15
11	43.44	8.95	26	68.18	14.30
12	49.08	9.61	27	69.56	24.42
13	53.73	11.11	28	74.90	11.54
14	54.41	13.28	29	77.91	17.65
15	55.10	10.22	30	80.00	17.36

a. Calculate the regression equation using selling price based on the annual dividend. Interpret the slope value.
b. Determine the coefficient of determination. Interpret its value.
c. Determine the coefficient of correlation. Can you conclude that it is greater than 0 using the .05 significance level?

52. A highway employee performed a regression analysis of the relationship between the number of construction work-zone fatalities and the number of unemployed people in a state. The regression equation is Fatalities = 12.7 + 0.000114 (Unemp). Some additional output is:

```
Predictor          Coef       SE Coef      T        P
Constant         12.726         8.115   1.57    0.134
Unemp       0.00011386   0.00002896   3.93    0.001

Analysis of Variance
Source             DF      SS      MS      F        P
Regression          1   10354   10354   15.46    0.001
Residual Error     18   12054     670
Total              19   22408
```

a. How many states were in the sample?
b. Determine the standard error of estimate.
c. Determine the coefficient of determination.
d. Determine the coefficient of correlation.
e. At the .05 significance level does the evidence suggest there is a positive association between fatalities and the number unemployed?

53. Regression analysis relating the current market value in dollars to the size in square feet of homes in Greene County has been developed. The computer output follows. The regression equation is: Value = −37,186 + 65.0 Size.

```
Predictor       Coef   SE Coef        T       P
Constant     -37186      4629    -8.03   0.000
Size          64.993     3.047    21.33   0.000

Analysis of Variance
Source           DF            SS            MS        F       P
Regression        1   13548662082   13548662082   454.98   0.000
Residual Error   33     982687392      29778406
Total            34   14531349474
```

a. How many homes were in the sample?
b. Compute the standard error of estimate.
c. Compute the coefficient of determination.
d. Compute the coefficient of correlation.
e. At the .05 significance level does the evidence suggest a positive association between the market value of homes and the size of the home in square feet?

54. The following table shows the mean annual percent return on capital (profitability) and the mean annual percentage sales growth for eight aerospace and defense companies.

Company	Profitability	Growth
Alliant Techsystems	23.1	8.0
Boeing	13.2	15.6
General Dynamics	24.2	31.2
Honeywell	11.1	2.5
L-3 Communications	10.1	35.4
Northrop Grunmman	10.8	6.0
Rockwell Collins	27.3	8.7
United Technologies	20.1	3.2

a. Compute the coefficient of correlation. Conduct a test of hypothesis to determine if it is reasonable to conclude that the population correlation is greater than zero. Use the .05 significance level.
b. Develop the regression equation for profitability based on growth. Comment on the slope value.
c. Use a software package to determine the residual for each observation. Which company has the largest residual?

55. The following data shows the retail price for 12 randomly selected laptop computers along with their corresponding processor speeds.

Computers	Speed	Price
1	2.0	$2,689
2	1.6	1,229
3	1.6	1,419
4	1.8	2,589
5	2.0	2,849
6	1.2	1,349
7	2.0	2,929
8	1.6	1,849
9	2.0	2,819
10	1.6	2,669
11	1.0	1,249
12	1.4	1,159

a. Develop a linear equation that can be used to describe how the price depends on the processor speed.
b. Based on your regression equation, is there one machine that seems particularly over- or underpriced?
c. Compute the correlation coefficient between the two variables. At the .05 significance level conduct a test of hypothesis to determine if the population correlation could be greater than zero.

56. A consumer buying cooperative tested the effective heating area of 20 different electric space heaters with different wattages. Here are the results.

Heater	Wattage	Area	Heater	Wattage	Area
1	1,500	205	11	1,250	116
2	750	70	12	500	72
3	1,500	199	13	500	82
4	1,250	151	14	1,500	206
5	1,250	181	15	2,000	245
6	1,250	217	16	1,500	219
7	1,000	94	17	750	63
8	2,000	298	18	1,500	200
9	1,000	135	19	1,250	151
10	1,500	211	20	500	44

a. Compute the correlation between the wattage and heating area. Is there a direct or an indirect relationship?
b. Conduct a test of hypothesis to determine if it is reasonable that the coefficient is greater than zero. Use the .05 significance level.
c. Develop the regression equation for effective heating based on wattage.
d. Which heater looks like the "best buy" based on the size of the residual?

57. A dog trainer is exploring the relationship between the size of the dog (weight) and its daily food consumption (measured in standard cups). Below is the result of a sample of 18 observations.

Dog	Weight	Consumption
1	41	3
2	148	8
3	79	5
4	41	4
5	85	5
6	111	6
7	37	3
8	111	6
9	41	3
10	91	5
11	109	6
12	207	10
13	49	3
14	113	6
15	84	5
16	95	5
17	57	4
18	168	9

a. Compute the correlation coefficient. Is it reasonable to conclude that the correlation in the population is greater than zero? Use the .05 significance level.
b. Develop the regression equation for cups based on the dog's weight. How much does each additional cup change the estimated weight of the dog?
c. Is one of the dogs a big undereater or overeater?

exercises.com

58. Suppose you want to study the association between the literacy rate in a country, the population, and the country's gross domestic product (GDP). Go to the website of *Information Please Almanac* (http://www.infoplease.com). Select the category **World,** and then select **Countries.** A list of 195 countries starting with Afghanistan and ending with Zimbabwe will appear. Randomly select a sample of about 20 countries. It may be convenient to use a systematic sample. In other words, randomly select 1 of the first 10 countries and then select every tenth country thereafter. Click on each country name and scan the information to find the literacy rate, the population, and the GDP. Compute the correlation among the variables. In other words, find the correlation between: literacy and population, literacy and GDP, and population and GDP. *Warning:* Be careful of the units. Sometimes population is reported in millions, other times in thousands. At the .05 significance level, can we conclude that the correlation is different from zero for each pair of variables?
59. Many real estate companies and rental agencies now publish their listings on the Web. One example is the Dunes Realty Company, located in Garden City and Surfside Beaches in South Carolina. Go to the Web site http://www.dunes.com and select **Vacation Rentals,** then **Beach Home Search.** Then indicate 5 bedroom, accommodations for 14 people, second row (this means it is across the street from the beach), and no pool or floating dock; select a week in July or August; indicate that you are willing to spend $8,000 per week; and then click on **Search the Beach Homes.** The output should include details on the cottages that met your criteria.
    a. Determine the correlation between the number of baths in each cottage and the weekly rental price. Can you conclude that the correlation is greater than zero at the .05 significance level? Determine the coefficient of determination.
    b. Determine the regression equation using the number of bathrooms as the independent variable and the price per week as the dependent variable. Interpret the regression equation.
    c. Calculate the correlation between the number of people the cottage will accommodate and the weekly rental price. At the .05 significance level can you conclude that it is different from zero?

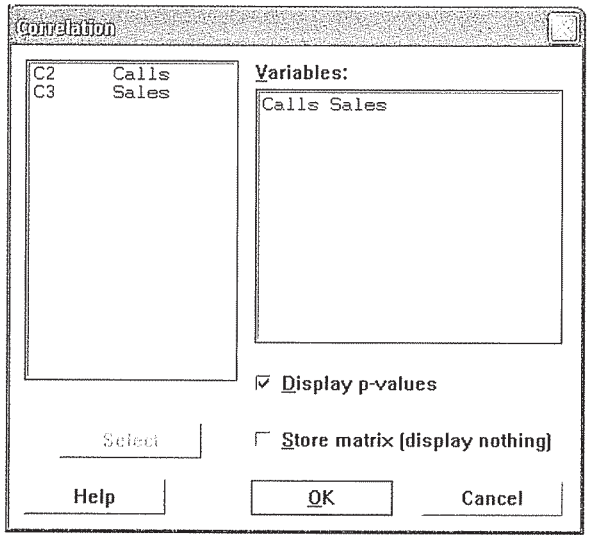
Dataset Exercises

60. Refer to the Real Estate data, which reports information on homes sold in Denver, Colorado, last year.
    a. Let selling price be the dependent variable and size of the home the independent variable. Determine the regression equation. Estimate the selling price for a home with an area of 2,200 square feet. Determine the 95 percent confidence interval and the 95 percent prediction interval for the selling price of a home with 2,200 square feet.
    b. Let selling price be the dependent variable and distance from the center of the city the independent variable. Determine the regression equation. Estimate the selling price of a home 20 miles from the center of the city. Determine the 95 percent confidence interval and the 95 percent prediction interval for homes 20 miles from the center of the city.
    c. Can you conclude that the independent variables "distance from the center of the city" and "selling price" are negatively correlated and that the area of the home and the selling price are positively correlated? Use the .05 significance level. Report the p-value of the test.
61. Refer to the Baseball 2003 data, which reports information on the 2003 Major League Baseball season.
    a. Let the games won be the dependent variable and total team salary, in millions of dollars, be the independent variable. Can you conclude that there is a positive association between the two variables? Determine the regression equation. Interpret the slope, that is the value of b. How many additional wins will an additional $5 million in salary bring?
    b. Determine the correlation between games won and ERA and between games won and team batting average. Which has the stronger correlation? Can we conclude that there is a positive correlation between wins and team batting and a negative correlation between wins and ERA? Use the .05 significance level.
    c. Assume the number of games won is the dependent variable and attendance the independent variable. Can we conclude that the correlation between these two variables is greater than 0? Use the .05 significance level.

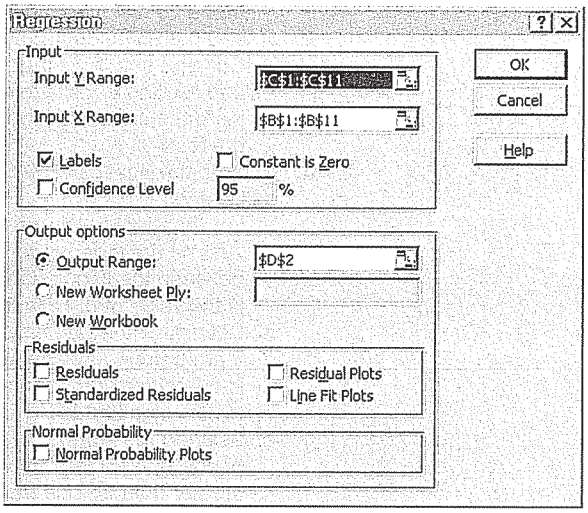
62. Refer to the Wage data, which reports information on annual wages for a sample of 100 workers. Also included are variables relating to industry, years of education, and gender for each worker.
    a. Determine the correlation between the annual wage and the years of education. At the .05 significance level can we conclude there is a positive correlation between the two variables?
    b. Determine the correlation between the annual wage and the years of work experience. At the .05 significance level can we conclude there is a positive correlation between the two variables?
63. Refer to the CIA data, which reports demographic and economic information on 46 countries.
    a. You wish to use the Labor force variable as the independent variable to predict the unemployment rate. Interpret the slope value. Use the appropriate linear regression equation to predict unemployment in the United Arab Emirates.
    b. Find the correlation coefficient between the levels of exports and imports. Use the .05 significance level to test whether there is a positive correlation between these two variables.
    c. Does there appear to be a relationship between the percentage of the population over 65 and the literacy percentage? Support your answer with statistical evidence. Conduct an appropriate test of hypothesis and interpret the result.

Software Commands

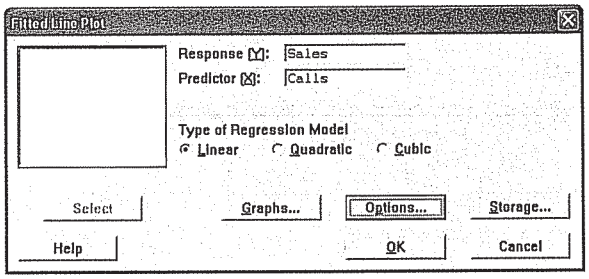
1. The MINITAB commands for the output showing the coefficient of correlation on page 385 are:
    a. Enter the sales representative's name in C1, the number of calls in C2, and the sales in C3.
    b. Select **Stat, Basic Statistics,** and **Correlation.**
    c. Select *Calls* and *Sales* as the variables, click on **Display p-values,** and then click **OK.**

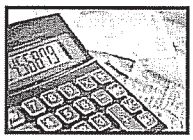


2. The computer commands for the Excel output on page 394 are:
    a. Enter the variable names in row 1 of columns A, B, and C. Enter the data in rows 2 through 11 in the same columns.
    b. Select **Tools, Data Analysis,** and then select **Regression.**
    c. For our spreadsheet we have *Calls* in column B and *Sales* in column C. The **Input Y-Range** is *C1:C11* and the **Input X-Range** is *B1:B11,* click on **Labels,** select *D1* as the **Output Range,** and click **OK.**



3. The MINITAB commands to the confidence intervals and prediction intervals on page 399 are:
   a. Select **Stat, Regression,** and **Fitted line plot.**
   b. In the next dialog box the **Response (Y)** is Sales and **Predictor (X)** is Calls. Select **Linear** for the type of regression model and then click on **Options.**
   c. In the **Options** dialog box click on **Display confidence and prediction bands,** use the **95.0 for confidence level,** and then in the **Title** box type an appropriate heading, then click **OK** and then **OK** again.

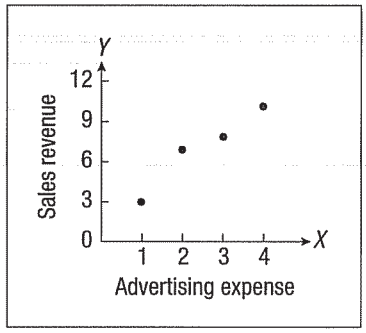




Chapter 13 Answers to Self-Review

13–1 **a.** Advertising expense is the independent variable and sales revenue is the dependent variable.

b.



c.

X	Y	$(X-\bar{X})$	$(X-\bar{X})^2$	$(Y-\bar{Y})$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$
2	7	−0.5	.25	0	0	0
1	3	−1.5	2.25	−4	16	6
3	8	0.5	.25	1	1	0.5
4	10	1.5	2.25	3	9	4.5
10	28		5.00		26	11

$$\bar{X} = \frac{10}{4} = 2.5 \qquad \bar{Y} = \frac{28}{4} = 7$$

$$s_x = \sqrt{\frac{5}{3}} = 1.2909944$$

$$s_y = \sqrt{\frac{26}{3}} = 2.9439203$$

$$r = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{(n-1)\,s_x s_y}$$

$$= \frac{11}{(4-1)(1.2909944)(2.9439203)}$$

$$= 0.9648$$

d. There is a strong correlation between the advertising expense and sales.

e. $r^2 = .93$, 93% of the variation in sales is "explained" by variation in advertising.

13–2 H_0: $\rho \leq 0$, H_1: $\rho > 0$. H_0 is rejected if $t > 1.714$.

$$t = \frac{.43\sqrt{25-2}}{\sqrt{1-(.43)^2}} = 2.284$$

H_0 is rejected. There is a positive correlation between the percent of the vote received and the amount spent on the campaign.

13–3 **a.** See the calculations in Self-Review 13–1, part (c).

$$b = \frac{rs_y}{s_x} = \frac{(0.9648)(2.9439)}{1.2910} = 2.2$$

$$a = \frac{28}{4} - 2.2\left(\frac{10}{4}\right) = 7 - 5.5 = 1.5$$

b. The slope is 2.2. This indicates that an increase of \$1 million in advertising will result in an increase of \$2.2 million in sales. The intercept is 1.5. If there was no expenditure for advertising, sales would be \$1.5 million.

c. $Y' = 1.5 + 2.2(3) = 8.1$

13–4 0.9487, found by:

Y	Y'	$(Y-Y')$	$(Y-Y')^2$
7	5.9	1.1	1.21
3	3.7	−0.7	.49
8	8.1	−0.1	.01
10	10.3	−0.3	.09
			1.80

$$s_{y\cdot x} = \sqrt{\frac{\Sigma(Y-Y')^2}{n-2}}$$

$$= \sqrt{\frac{1.80}{4-2}} = .9487$$

13–5 Since Y' for an X of 3 is 8.1, found by $Y' = 1.5 + 2.2(3) = 8.1$, then $\bar{X} = 2.5$ and $\Sigma(X-\bar{X})^2 = 5$.

t from Appendix F for $4 - 2 = 2$ degrees of freedom at the .10 level is 2.920.

$$Y' \pm t(s_{y\cdot x})\sqrt{\frac{1}{n} + \frac{(X-\bar{X})^2}{\Sigma(X-\bar{X})^2}}$$

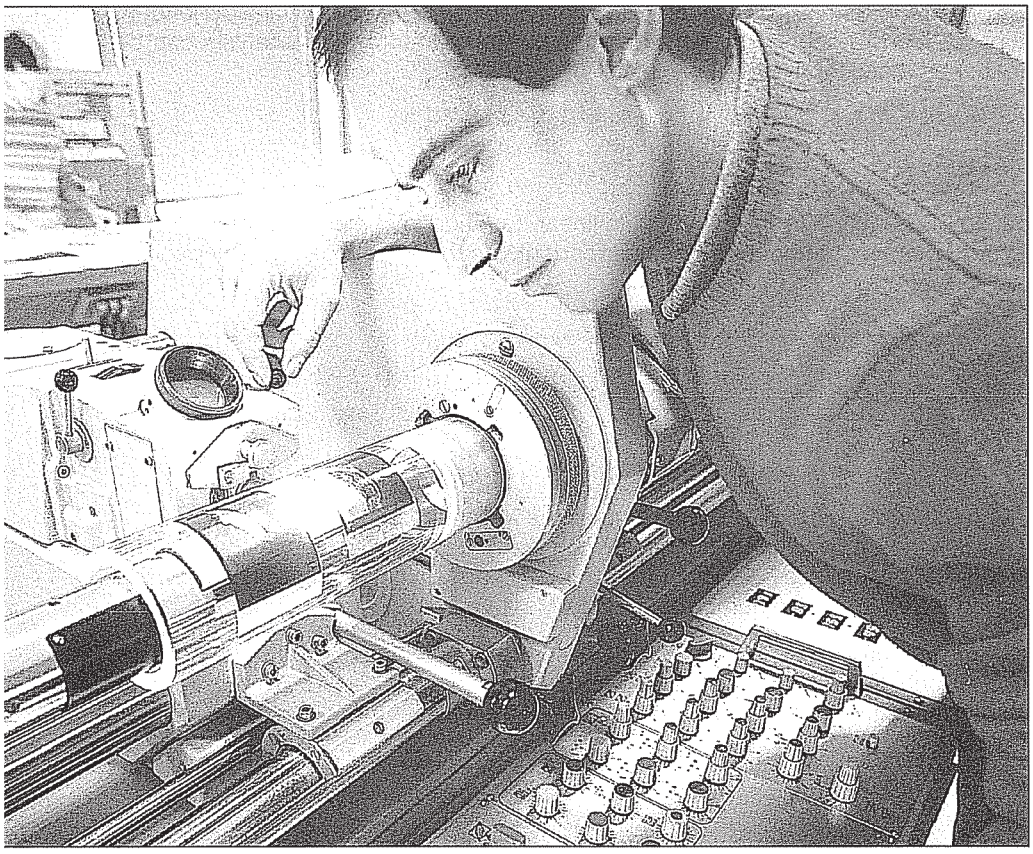
$$= 8.1 \pm 2.920(0.9487)\sqrt{\frac{1}{n} + \frac{(3-2.5)^2}{5}}$$

$$= 8.1 \pm 2.920(0.9487)(0.5477)$$

$$= 6.58 \text{ and } 9.62 \text{ (in \$ millions)}$$

14

Multiple Regression and Correlation Analysis



Thompson Photo Works purchased several new, highly sophisticated machines. The production department needed some guidance with respect to qualifications needed by an operator. In order to explore factors needed to estimate performance on the new machines, they explored four variables. (See Goal 1 and Exercise 2.)

GOALS

When you have completed this chapter, you will be able to:

1. **Describe the relationship between several *independent variables* and a *dependent variable* using a *multiple regression equation.***
2. **Compute and interpret the *multiple standard error of estimate* and the *coefficient of determination.***
3. **Interpret a *correlation matrix.***
4. **Set up and interpret an ANOVA table.**
5. **Conduct a test of hypothesis to determine whether regression coefficients differ from zero.**
6. **Conduct a test of hypothesis on each of the regression coefficients.**

Introduction

In Chapter 13 we described the relationship between a pair of interval- or ratio-scaled measurements. We began the chapter by studying the coefficient of correlation, which measures strength of the relationship. A coefficient near plus or minus 1.00 (−.88 or .78, for example) indicates a very strong linear relationship, whereas a value near 0 (−.12 or .18, for example) means that the relationship is weak. Next we developed a procedure to determine a linear equation to express the relationship between the two variables. We referred to this as a *line of regression.* This line describes the relationship between the variables. It also describes the overall pattern of a dependent variable (Y) to a single independent or explanatory variable (X).

In multiple linear correlation and regression we use additional independent variables (denoted X_1, X_2, . . ., and so on) that help us better explain or predict the dependent variable (Y). Almost all of the ideas we saw in simple linear correlation and regression extend to this more general situation. However, the additional independent variables do lead to some new considerations. Multiple regression analysis can be used either as a descriptive or as an inferential technique.

Multiple Regression Analysis

The general descriptive form of a multiple linear equation is shown in formula (14–1). We use k to represent the number of independent variables. So k can be any positive integer.

GENERAL MULTIPLE REGRESSION EQUATION

$$Y' = a + b_1X_1 + b_2X_2 + b_3X_3 + \cdots + b_kX_k \qquad [14\text{-}1]$$

where:

a is the intercept, the value of Y when all the X's are zero.
b_j is the amount by which Y changes when that particular X_j increases by one unit with all other values held the same. The subscript $_j$ can assume values between 1 and k, which is the number of independent variables.

When there are only two independent variables, this equation can be portrayed graphically as a plane. Chart 14–1 is a graph of the relationship $Y' = a + b_1X_1 + b_2X_2$ used to summarize or "fit" 10 observations.

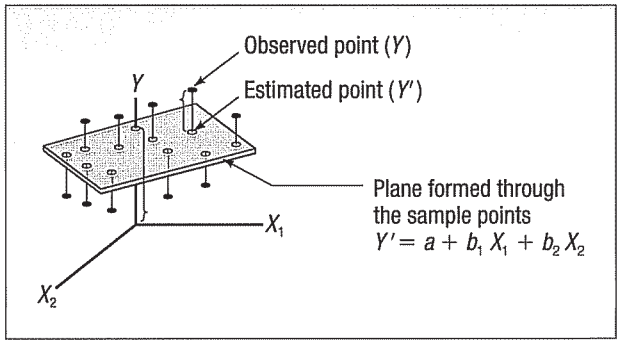


CHART 14–1 Regression Plane with Ten Sample Points

To illustrate the interpretation of the intercept and the two regression coefficients, suppose a vehicle's mileage per gallon of gasoline is directly related to the octane rating of the gasoline being used (X_1) and inversely related to the weight of the automobile (X_2). Assume that the regression equation, calculated using statistical software, is:

$$Y' = 6.3 + 0.2X_1 - 0.001X_2$$

The intercept value of 6.3 indicates the regression equation intersects the Y-axis at 6.3 when both X_1 and X_2 are zero. Of course, this does not make any physical sense to own an automobile that has no (zero) weight and to use gasoline with no octane. It is important to keep in mind that a regression equation is not generally used outside the range of the sample values.

The b_1 of 0.2 indicates that for each increase of 1 in the octane rating of the gasoline, the automobile would travel 2/10 of a mile more per gallon, *regardless of the weight of the vehicle.* That is, the vehicle's weight is held constant. The b_2 value of -0.001 reveals that for each increase of one pound in the vehicle's weight, the number of miles traveled per gallon decreases by 0.001, *regardless of the octane of the gasoline being used.*

As an example, an automobile with 92-octane gasoline in the tank and weighing 2,000 pounds would travel an average 22.7 miles per gallon, found by:

$$Y' = a + b_1X_1 + b_2X_2 = 6.3 + 0.2(92) - 0.001(2{,}000) = 22.7$$

The value of 22.7 is in miles per gallon.

The values for the coefficients in the multiple linear equation are found by using the method of least squares. Recall from the previous chapter that the least squares method makes the sum of the squared differences between the fitted and actual values of Y as small as possible. Because the calculations are very tedious, they are usually performed by a statistical software package, such as Excel or MINITAB. Fortunately, the information reported is fairly standard.

Inferences in Multiple Linear Regression

Thus far, multiple regression analysis has been viewed only as a way to describe the relationship between a dependent variable and several independent variables. However, the least squares method also has the ability to draw inferences or generalizations about the relationship for an entire population. Recall that when you create confidence intervals or perform hypothesis tests as a part of inferential statistics, you view the data as a random sample taken from some population.

In the multiple regression setting, we assume there is an unknown population regression equation that relates the dependent variable to the k independent variables. This is sometimes called a **model** of the relationship. In symbols we write:

$$Y' = \alpha + \beta_1X_1 + \beta_2X_2 + \cdots + \beta_kX_k$$

This equation is analogous to formula (14–1) except the coefficients are now reported as Greek letters. We use the Greek letters to denote *population parameters*. Then under a certain set of assumptions, which will be discussed shortly, the computed values of a and b_j are sample statistics. These sample statistics are point estimates of the corresponding population parameters α and β_j. These point estimates have normally distributed sampling distributions. These sampling distributions are each centered at their respective parameter values. To put it another way, the means of the sampling distributions are equal to the parameter values to be estimated. Thus, by using the properties of the sampling distributions of these statistics, we can make inferences about the population parameters.

We begin the discussion of multiple regression by describing a situation involving three independent variables.

EXAMPLE

Salsberry Realty sells homes along the east coast of the United States. One of the questions most frequently asked by prospective buyers is: If we purchase this home, how much can we expect to pay to heat it during the winter? The research department at Salsberry has been asked to develop some guidelines regarding heating costs for single-family homes. Three variables are thought to relate to the heating costs: (1) the mean daily outside temperature, (2) the number of inches of insulation in the attic, and (3) the age of the furnace. To investigate, Salsberry's research department selected a random sample of 20 recently sold homes. They determined the cost to heat the home last January, as well as the January outside temperature in the region, the number of inches of insulation in the attic, and the age of the furnace. The sample information is reported in Table 14–1.

TABLE 14–1 Factors in January Heating Cost for a Sample of 20 Homes

Home	Heating Cost ($)	Mean Outside Temperature (°F)	Attic Insulation (inches)	Age of Furnace (years)
1	$250	35	3	6
2	360	29	4	10
3	165	36	7	3
4	43	60	6	9
5	92	65	5	6
6	200	30	5	5
7	355	10	6	7
8	290	7	10	10
9	230	21	9	11
10	120	55	2	5
11	73	54	12	4
12	205	48	5	1
13	400	20	5	15
14	320	39	4	7
15	72	60	8	6
16	272	20	5	8
17	94	58	7	3
18	190	40	8	11
19	235	27	9	8
20	139	30	7	5

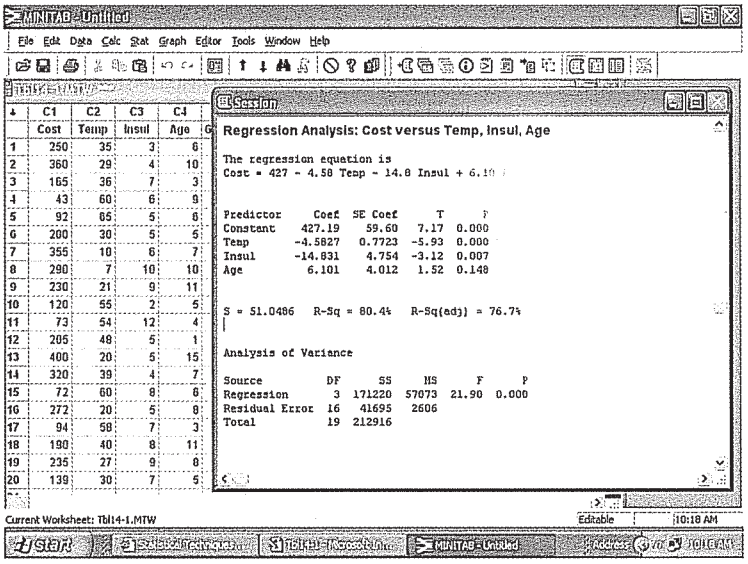
Determine the multiple regression equation. Which variables are the independent variables? Which variable is the dependent variable? Discuss the regression coefficients. What does it indicate if some coefficients are positive and some coefficients are negative? What is the intercept value? What is the estimated heating cost for a home if the mean outside temperature is 30 degrees, there are 5 inches of insulation in the attic, and the furnace is 10 years old?

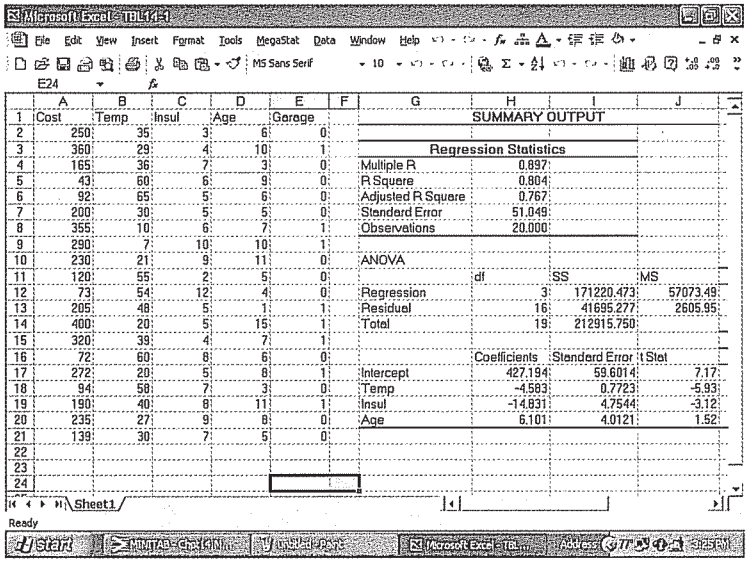
Statistics in Action

Many studies indicate a woman will earn about 70 percent of what a man would for the same work. Researchers at the University of Michigan Institute for Social Research found that about one-third of the difference can be explained by such social factors as differences in education, seniority, and work interruptions. The remaining two-thirds is not explained by these social factors.

SOLUTION

The MINITAB and Excel statistical software systems generate the outputs shown below.





The dependent variable is the January heating cost. There are three independent variables, the mean outside temperature, the number of inches of insulation in the attic, and the age of the furnace.

The general form of a multiple regression equation with three independent variables is:

$$Y' = a + b_1X_1 + b_2X_2 + b_3X_3$$

In this case the estimated multiple regression equation is $Y' = 427 - 4.58X_1 - 14.8X_2 + 6.10X_3$. The intercept value is 427. This is the point where the regression equation crosses the Y-axis. The regression coefficients for the mean outside temperature and the amount of attic insulation are both negative. This is not surprising. As the outside

temperature increases, the cost to heat the home will go down. Hence, we would expect an inverse relationship. For each degree the mean temperature increases, we expect the heating cost to decrease $4.58 per month. So if the mean temperature in Boston is 25 degrees and it is 35 degrees in Philadelphia, all other things being the same, we expect the heating cost would be $45.80 less in Philadelphia.

The variable "attic insulation" also shows an inverse relationship: the more insulation in the attic, the less the cost to heat the home. So the negative sign for this coefficient is logical. For each additional inch of insulation, we expect the cost to heat the home to decline $14.80 per month, regardless of the outside temperature or the age of the furnace.

The age of the furnace variable shows a direct relationship. With an older furnace, the cost to heat the home increases. Specifically, for each additional year older the furnace is, we expect the cost to increase $6.10 per month.

The estimated heating cost for the month is $276.60 if the mean outside temperature for the month is 30 degrees, there are 5 inches of insulation in the attic, and the furnace is 10 years old.

$$Y' = a + b_1X_1 + b_2X_2 + b_3X_3 = 427 - 4.58(30) - 14.8(5) + 6.10(10) = 276.60$$

Self-Review 14–1

The quality control engineer at Palmer Industries is interested in estimating the tensile strength of steel wire based on its outside diameter and the amount of molybdenum in the steel. As an experiment, she selected 25 pieces of wire, measured the outside diameters, and determined the molybdenum content. Then she measured the tensile strength of each piece. The results of the first four were:

Piece	Tensile Strength (psi), Y	Outside Diameter (mm), X_1	Amount of Molybdenum (units), X_2
A	11	.3	6
B	9	.2	5
C	16	.4	8
D	12	.3	7

Using a statistical software package, the QC engineer determined the multiple regression equation to be $Y' = -0.5 + 20X_1 + 1X_2$.

(a) From the equation, what is the estimated tensile strength of a steel wire having an outside diameter of .35 mm and 6.4 units of molybdenum?
(b) Interpret the value of b_1 in the equation.

Exercises

1. The director of marketing at Reeves Wholesale Products is studying monthly sales. Three independent variables were selected as estimators of sales: regional population, per-capita income, and regional unemployment rate. The regression equation was computed to be (in dollars):

$$Y' = 64{,}100 + 0.394X_1 + 9.6X_1 - 11{,}600X_3$$

a. What is the full name of the equation?
b. Interpret the number 64,100.
c. What are the estimated monthly sales for a particular region with a population of 796,000, per-capita income of $6,940, and an unemployment rate of 6.0 percent?

2. Thompson Photo Works purchased several new, highly sophisticated processing machines. The production department needed some guidance with respect to qualifications needed by an operator. Is age a factor? Is the length of service as an operator important? In order to explore further the factors needed to estimate performance on the new processing machines, four variables were listed:

   X_1 = Length of time an employee was in the industry. X_3 = Prior on-the-job rating.
   X_2 = Mechanical aptitude test score. X_4 = Age.

   Performance on the new machine is designated Y.

   Thirty employees were selected at random. Data were collected for each, and their performances on the new machines were recorded. A few results are:

Name	Performance on New Machine, Y	Length of Time in Industry, X_1	Mechanical Aptitude Score, X_2	Prior On-the-Job Performance, X_3	Age, X_4
Andy Kosin	112	12	312	121	52
Sue Annis	113	2	380	123	27

   The equation is:

$$Y' = 11.6 + 0.4X_1 + 0.286X_2 + 0.112X_3 + 0.002X_4$$

   a. What is the full designation of the equation?
   b. How many dependent variables are there? Independent variables?
   c. What is the number 0.286 called?
   d. As age increases by one year, how much does estimated performance on the new machine increase?
   e. Carl Knox applied for a job at Photo Works. He has been in the business for six years, and scored 280 on the mechanical aptitude test. Carl's prior on-the-job performance rating is 97, and he is 35 years old. Estimate Carl's performance on the new machine.

3. A sample of General Mills employees was studied to determine their degree of satisfaction with their present life. A special index, called the index of satisfaction, was used to measure satisfaction. Six factors were studied, namely, age at the time of first marriage (X_1), annual income (X_2), number of children living (X_3), value of all assets (X_4), status of health in the form of an index (X_5), and the average number of social activities per week—such as bowling and dancing (X_6). Suppose the multiple regression equation is:

$$Y' = 16.24 + 0.017X_1 + 0.0028X_2 + 42X_3 + 0.0012X_4 + 0.19X_5 + 26.8X_6$$

   a. What is the estimated index of satisfaction for a person who first married at 18, has an annual income of $26,500, has three children living, has assets of $156,000, has an index of health status of 141, and has 2.5 social activities a week on the average?
   b. Which would add more to satisfaction, an additional income of $10,000 a year or two more social activities a week?

4. Cellulon, a manufacturer of home insulation, wants to develop guidelines for builders and consumers regarding the effects (1) of the thickness of the insulation in the attic of a home and (2) of the outdoor temperature on natural gas consumption. In the laboratory they varied the insulation thickness and temperature. A few of the findings are:

Monthly Natural Gas Consumption (cubic feet), Y	Thickness of Insulation (inches), X_1	Outdoor Temperature (°F), X_2
30.3	6	40
26.9	12	40
22.1	8	49

   On the basis of the sample results, the regression equation is:

$$Y' = 62.65 - 1.86X_1 - 0.52X_2$$

a. How much natural gas can homeowners expect to use per month if they install 6 inches of insulation and the outdoor temperature is 40 degrees F?
b. What effect would installing 7 inches of insulation instead of 6 have on the monthly natural gas consumption (assuming the outdoor temperature remains at 40 degrees F)?
c. Why are the regression coefficients b_1 and b_2 negative? Is this logical?

Multiple Standard Error of Estimate

In the Salsberry Realty example we estimated the cost to heat a home during the month of January when the mean outside temperature was 30 degrees, there were 5 inches of attic insulation, and the furnace was 10 years old to be $276.60. We would expect to find some random error in this estimate. Sometimes a home with these statistics would cost more than $276.60 to heat and other times less. The error in this estimate is measured by the **multiple standard error of estimate.** The standard error, as it is usually called, is denoted $s_{y \cdot 123}$. The subscripts indicate that three independent variables are being used to estimate the value of Y.

Recall from Chapter 13 the standard error of estimate described the variation around the regression line. A small standard error indicates the points are close to the regression line, whereas a large value indicates the points are scattered about the regression line. The same concept is true in multiple regression. If we have two independent variables, then we can think of the variation around a regression plane. See Chart 14–1 on page 422. If there are more than two independent variables, we do not have a geometric interpretation of the equation, but the standard error is still a measure of the "error" or variability in the prediction.

The formula to compute the standard error is similar to that used in the previous chapter. See formula (13–6) on page 393. The numerator is the sum of the squared differences between the estimated and the actual values of the dependent variable. In the denominator, we adjust for the fact that we are considering several, that is, k, independent variables.

MULTIPLE STANDARD ERROR OF ESTIMATE

$$s_{y \cdot 12 \cdots k} = \sqrt{\frac{\Sigma (Y - Y')^2}{n - (k + 1)}} \qquad [14\text{–}2]$$

where:

Y is the observation.
Y' is the value estimated from the regression equation.
n is the number of observations in the sample.
k is the number of independent variables.

In the Salsberry Realty example, $k = 3$.

Again, we use the Salsberry Realty problem to illustrate. The first home had a mean outside temperature of 35 degrees, 3 inches of attic insulation, and a 6-year-old furnace. Substituting these values into the regression equation, the estimated heating cost is $258.90, determined by $427 - 4.58(35) - 14.80(3) + 6.10(6)$. The Y' values for the other homes are found similarly and are reported in Table 14–2.

The actual heating cost for the first home is $250, in contrast to the estimated cost of $258.90. That is, the error in the prediction is −$8.90, found by ($250 − $258.90). This difference between the actual heating cost and the estimated heating cost is called the **residual.** To find the multiple standard error of estimate, we determine the residual for each of the sampled homes, square the residual, and then total the squared residuals. The total is reported in the lower right corner of Table 14–2.

In this example $n = 20$ and $k = 3$ (three independent variables), so the multiple standard error of estimate is:

$$s_{y \cdot 123} = \sqrt{\frac{\Sigma (Y - Y')^2}{n - (k + 1)}} = \sqrt{\frac{41{,}695.58}{20 - (3 + 1)}} = 51.05$$

TABLE 14–2 Calculations Needed for the Multiple Standard Error of Estimate

Home	Temperature (°F)	Insulation (inches)	Age (years)	Cost, Y	Y'	$(Y - Y')$	$(Y - Y')^2$
1	35	3	6	$250	258.90	−8.90	79.21
2	29	4	10	360	295.98	64.02	4,098.56
3	36	7	3	165	176.82	−11.82	139.71
4	60	6	9	43	118.30	−75.30	5,670.09
5	65	5	6	92	91.90	0.10	0.01
6	30	5	5	200	246.10	−46.10	2,125.21
7	10	6	7	355	335.10	19.90	396.01
8	7	10	10	290	307.94	−17.94	321.84
9	21	9	11	230	264.72	−34.72	1,205.48
10	55	2	5	120	176.00	−56.00	3,136.00
11	54	12	4	73	26.48	46.52	2,164.11
12	48	5	1	205	139.26	65.74	4,321.75
13	20	5	15	400	352.90	47.10	2,218.41
14	39	4	7	320	231.88	88.12	7,765.13
15	60	8	6	72	70.40	1.60	2.56
16	20	5	8	272	310.20	−38.20	1,459.24
17	58	7	3	94	76.06	17.94	321.84
18	40	8	11	190	192.50	−2.50	6.25
19	27	9	8	235	218.94	16.06	257.92
20	30	7	5	139	216.50	−77.50	6,006.25
Total							41,695.58

How do we interpret the 51.05? It is the typical "error" we make when we use this equation to predict the cost. First, the units are the same as the dependent variable, so the standard error is in dollars. Second, if the errors are normally distributed, about 68 percent of the residuals should be between ±51.05 and about 95 percent should be less than ±2(51.05) or ±102.10. Refer to the second column from the right in Table 14–2, the column headed $(Y - Y')$. Of the 20 residuals reported in this column, 14 are less than ±51.05 and all are less than ±102.10, which is quite close to the guidelines of 68 percent and 95 percent.

In Chapter 13 we used the standard error of estimate to construct confidence intervals and prediction intervals. We will not detail these procedures for multiple regression, but they are available on statistical software systems, such as MINITAB.

Assumptions about Multiple Regression and Correlation

Before continuing our discussion, we list the assumptions underlying both multiple regression and multiple correlation. As noted in several previous chapters, we identify the assumptions because if they are not fully met, the results might be biased. For instance, in selecting a sample, we assume that all the items in the population have a chance of being selected. If our research involves surveying all those who ski, but we ignore those over 40 because we believe they are "too old," we would be biasing the responses toward the younger skiers. It should be mentioned, however, that in practice strict adherence to the following assumptions is not always possible in multiple regression and correlation problems involving the ever-changing business climate. But the statistical techniques discussed in this chapter appear to work well even when one or more of the

assumptions are violated. Even if the values in the multiple regression equation are "off" slightly, our estimates based on the equation will be closer than any that could otherwise be made.

Each of the following assumptions will be discussed in more detail as we progress through the chapter.

1. The independent variables and the dependent variable have a linear relationship.
2. The dependent variable is continuous and at least interval scale.
3. The variation in the difference between the actual and the predicted values is the same for all fitted values of Y. That is, $(Y - Y')$ must be approximately the same for all values of Y'. When this is the case, differences exhibit **homoscedasticity.**

Homoscedasticity

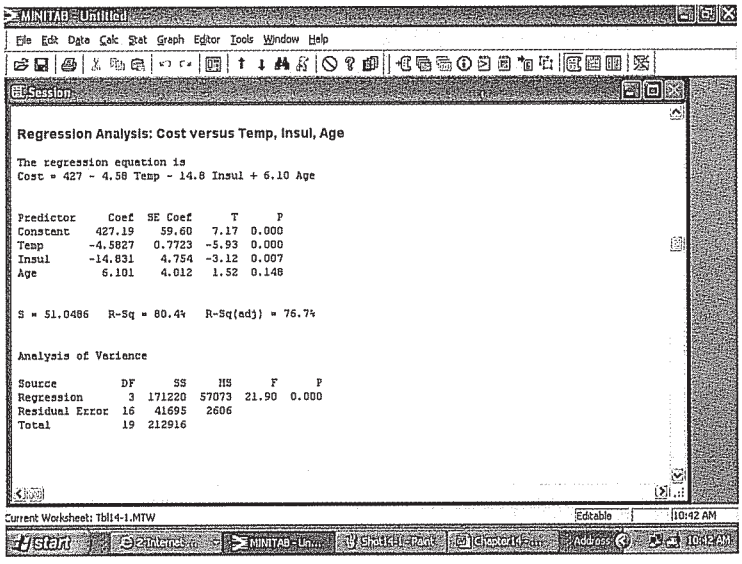
4. The residuals, computed by $Y - Y'$, are normally distributed with a mean of 0.
5. Successive observations of the dependent variable are uncorrelated. Violation of this assumption is called **autocorrelation.** Autocorrelation often happens when data are collected successively over periods of time.

Autocorrelation

Statistical tests are available to detect homoscedasticity and autocorrelation. For those interested, these tests are covered in more advanced textbooks such as *Applied Linear Regression Models* by Kutner, Nachtscheim, and Neter (4th ed., 2004, published by McGraw-Hill/Irwin).

The ANOVA Table

As mentioned previously, the multiple regression calculations are lengthy. Fortunately, many software systems are available to perform the calculations. Most of the systems report the results in a fairly standard format. The outputs from MINITAB and Excel shown on page 425 is typical. It includes the regression equation, the standard error of estimate, the coefficient of determination, as well as an analysis of variance table. We have already described the meaning of the regression coefficients in the equation $Y' = 427 - 4.58X_1 - 14.8X_2 + 6.10X_3$. We will discuss the "Coef," "StDev," and "T" (i.e., t ratio) columns later in the chapter. A portion of the output from MINITAB is repeated here.



First, let's focus on the analysis of variance table. It is similar to the ANOVA table described in Chapter 12. In that chapter the variation was divided into two components: variance due to the *treatments* and variance due to random *error.* Here the total

variance is also divided into two components: variance explained by the **regression,** that is, the independent variables, and the **error variance, residual error,** or unexplained variation. These two categories are identified in the "Source" column of the analysis of variance table. In the example there are 20 observations, so $n = 20$. The *total* number of degrees of freedom is $n - 1$, or $20 - 1 = 19$. The number of degrees of freedom in the "Regression" row is the number of independent variables. We let k represent the number of independent variables, so $k = 3$. The number of degrees of freedom in the "Residual Error" row is $n - (k + 1) = 20 - (3 + 1) = 16$ degrees of freedom.

The heading "SS" in the middle of the ANOVA table refers to the sum of squares, or the variation.

$$\text{Total variation} = \text{SS total} = \Sigma(Y - \bar{Y})^2 = 212{,}916$$
$$\text{Residual error} = \text{SSE} = \Sigma(Y - Y')^2 = 41{,}695$$
$$\text{Regression variation} = \text{SSR} = \Sigma(Y' - \bar{Y})^2 = \text{SS total} - \text{SSE}$$
$$= 212{,}916 - 41{,}695 = 171{,}220$$

The column headed "MS" (mean square) is determined by dividing the SS term by the *df* term. Thus, MSR, the mean square regression, is equal to SSR/k, and MSE equals SSE/$[n - (k + 1)]$. The general format of the ANOVA table is:

Source	*df*	SS	MS	*F*
Regression	k	SSR	MSR = SSR/k	MSR/MSE
Error	$n - (k + 1)$	SSE	MSE = SSE/$[n - (k + 1)]$	
Total	$n - 1$	SS total		

The **coefficient of multiple determination,** written as R^2, is the percent of the total variation explained by the regression. It is the sum of squares due to the regression, divided by the sum of squares total.

COEFFICIENT OF MULTIPLE DETERMINATION

$$R^2 = \frac{\text{SSR}}{\text{SS total}} \qquad [14\text{–}3]$$

$$R^2 = \frac{\text{SSR}}{\text{SS total}} = \frac{171{,}220}{212{,}916} = .804$$

The multiple standard error of estimate may also be found directly from the ANOVA table.

$$s_{y\cdot123} = \sqrt{\frac{\text{SSE}}{n - (k + 1)}} = \sqrt{\frac{41{,}695}{[20-(3+1)]}} = 51.05$$

These values, $R^2 = .804$ and $s_{y\cdot123} = 51.05$, are included in the MINITAB output.

Self-Review 14–2

Refer to the following ANOVA table.

SOURCE	DF	SS	MS	F
Regression	4	10	2.50	10.0
Error	20	5	0.25	
Total	24	15		

(a) How large was the sample?
(b) How many independent variables are there?
(c) Compute the coefficient of multiple determination.
(d) Compute the multiple standard error of estimate.

Exercises

5. Refer to the following ANOVA table.

SOURCE	DF	SS	MS	F
Regression	3	21	7.0	2.33
Error	15	45	3.0	
Total	18	66		

   a. How large was the sample?
   b. How many independent variables are there?
   c. Compute the coefficient of multiple determination.
   d. Compute the multiple standard error of estimate.

6. Refer to the following ANOVA table.

SOURCE	DF	SS	MS	F
Regression	5	60	12	1.714
Error	20	140	7	
Total	25	200		

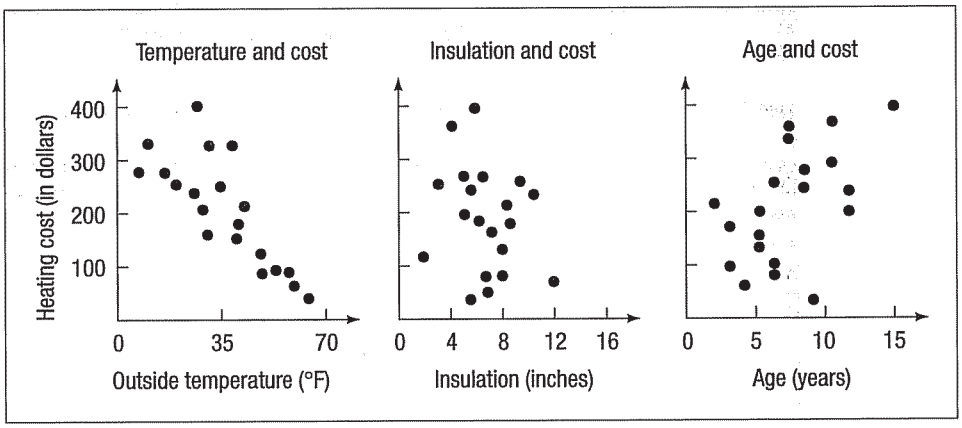
   a. How large was the sample?
   b. How many independent variables are there?
   c. Compute the coefficient of multiple determination.
   d. Compute the multiple standard error of estimate.

Evaluating the Regression Equation

Earlier in the chapter we described an example in which Salsberry Realty developed, using multiple regression techniques, an equation to express the cost to heat a home during the month of January based on the mean outside temperature, the number of inches of attic insulation, and the age of the furnace. The equation seemed reasonable, but we may wish to verify that the multiple coefficient of determination is significantly larger than zero, evaluate the regression coefficients to see which are not equal to zero, and verify that the regression assumptions are met.

Using a Scatter Diagram

There are three independent variables, designated X_1, X_2, and X_3. The dependent variable, the heating cost, is designated Y. In order to visualize the relationships between the dependent variable and each of the independent variables, we can draw the following scatter diagrams.



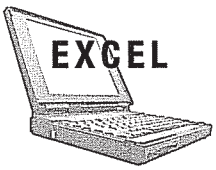
Of the three independent variables, the strongest association is between heating cost and the mean outside temperature. The relationships between cost and temperature and cost and insulation both are inverse. That is, as the independent variable increases, the dependent variable decreases. The relationship between the heating cost and the age of the furnace is direct. As the furnace gets older, it costs more to heat the home.

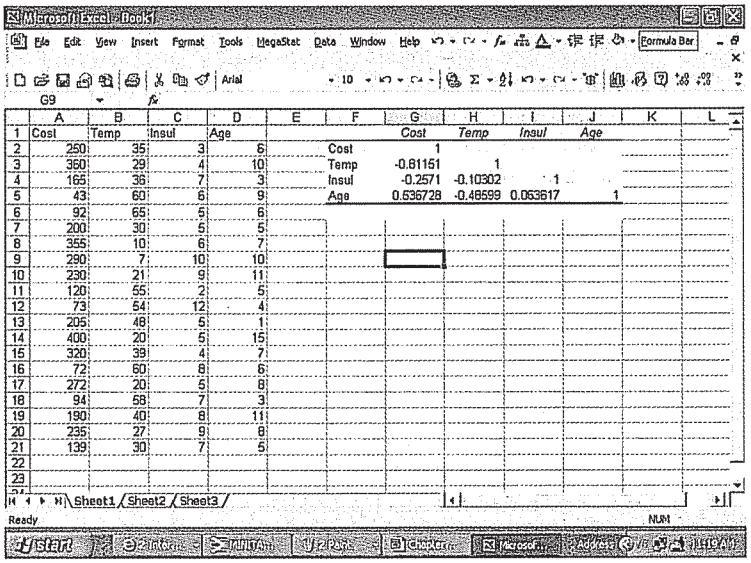
Correlation Matrix

A correlation matrix is also useful in analyzing the factors involved in the cost to heat a home.

CORRELATION MATRIX A matrix showing the coefficients of correlation between all pairs of variables.

The correlation matrix of the Salsberry Realty example follows. The matrix, which appears on the right-hand side of the output was developed using the Excel software.





	A	B	C	D	E	F	G	H	I	J	K	L
1	Cost	Temp	Insul	Age			Cost	Temp	Insul	Age		
2	250	35	3	6		Cost	1					
3	360	29	4	10		Temp	-0.81151	1				
4	165	36	7	3		Insul	-0.2571	-0.10302	1			
5	43	60	6	9		Age	0.536728	-0.46599	0.063617	1		
6	92	65	5	6								
7	200	30	5	5								
8	355	10	6	7								
9	290	7	10	10								
10	230	21	9	11								
11	120	55	2	5								
12	73	54	12	4								
13	205	48	5	1								
14	400	20	5	15								
15	320	39	4	7								
16	72	60	8	6								
17	272	20	5	8								
18	94	58	7	3								
19	190	40	8	11								
20	235	27	9	8								
21	139	30	7	5								
22												
23												

Cost is the dependent variable, *Y*. We are particularly interested in independent variables that have a strong correlation with the dependent variable. We may wish to develop a simpler multiple regression equation using fewer independent variables and the correlation matrix helps us identify which variables may be relatively more important. As indicated in the output, temperature has the strongest correlation with cost, −0.81151. The negative sign indicates the inverse relationship we were expecting. Age has a stronger correlation with cost than insulation and, again as we expected, the correlation between cost and the age of the furnace is direct. It is 0.53673.

A second use of the correlation matrix is to check for **multicollinearity.**

MULTICOLLINEARITY Correlation among the independent variables.

Multicollinearity can distort the standard error of estimate and may, therefore, lead to incorrect conclusions as to which independent variables are statistically significant. In this case, the correlation between the age of the furnace and the temperature is the strongest, but it is not large enough to cause a problem. A common rule of thumb is that correlations among the independent variables between $-.70$ and .70 do not cause difficulties. The usual remedy for multicollinearity is to drop one of the independent variables that are strongly correlated and recompute the regression equation.

Global Test: Testing the Multiple Regression Model

We can test the ability of the independent variables $X_1, X_2, \ldots, X_k$ to explain the behavior of the dependent variable Y. To put this in question form: Can the dependent variable be estimated without relying on the independent variables? The test used is referred to as the **global test.** Basically, it investigates whether it is possible all the independent variables have zero net regression coefficients. To put it another way, could the amount of explained variation, R^2, occur by chance?

To relate this question to the heating cost example, we will test whether the independent variables (amount of insulation in the attic, mean daily outside temperature, and age of furnace) are capable of effectively estimating home heating costs.

Recall that in testing a hypothesis, we first state the null hypothesis and the alternate hypothesis. In the heating cost example, there are three independent variables. Recall that b_1, b_2, and b_3 are sample net regression coefficients. The corresponding coefficients in the population are given the symbols β_1, β_2, and β_3. We now test whether the net regression coefficients in the population are all zero. The null hypothesis is:

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0$$

The alternate hypothesis is:

H_1: Not all the βs are 0.

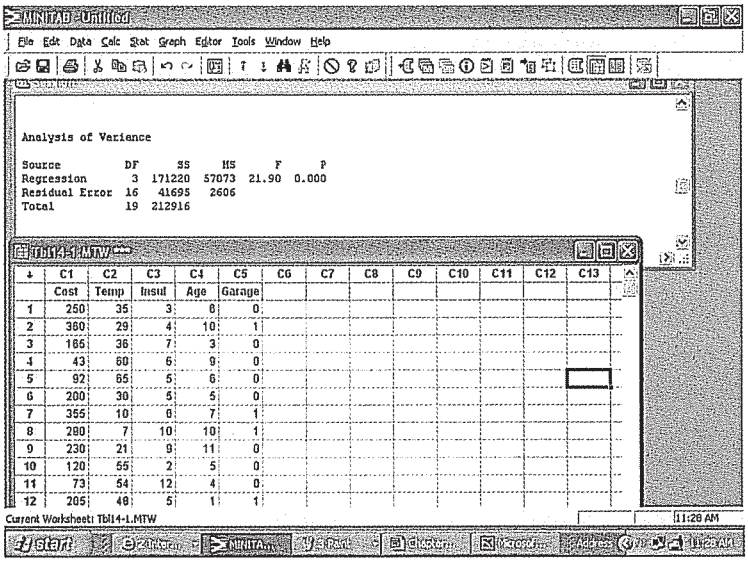
If the null hypothesis is true, it implies the regression coefficients are all zero and, logically, are of no use in estimating the dependent variable (heating cost). Should that be the case, we would have to search for some other independent variables—or take a different approach—to predict home heating costs.

To test the null hypothesis that the multiple regression coefficients are all zero, we employ the F distribution introduced in Chapter 12. We will use the .05 level of significance. Recall these characteristics of the F distribution:

Characteristics of the F distribution

1. It is positively skewed, with the critical value located in the right tail. The critical value is the point that separates the region where H_0 is not rejected from the region of rejection.
2. It is constructed by knowing the number of degrees of freedom in the numerator and the number of degrees of freedom in the denominator.

The degrees of freedom for the numerator and the denominator may be found in the analysis of variance table. That portion of the table is included on the next page. The top number in the column marked "DF" is 3, indicating that there are 3 degrees of freedom in the numerator. The middle number in the "DF" column (16) indicates that there are 16 degrees of freedom in the denominator. The number 16 is found by $n - (k + 1) = 20 - (3 + 1) = 16$. The number 3 corresponds to the number of independent variables.



The value of F is found from the following equation.

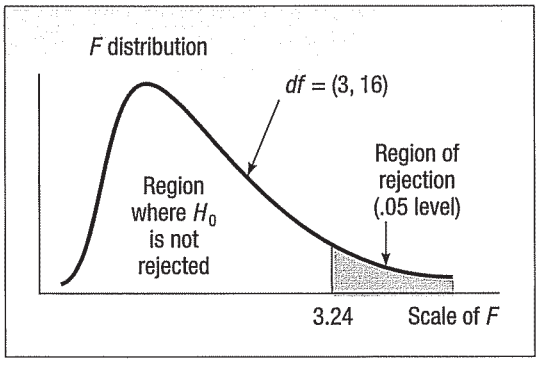
GLOBAL TEST

$$F = \frac{\text{SSR}/k}{\text{SSE}/[n - (k + 1)]} \qquad [14\text{–}4]$$

SSR is the sum of the squares "explained by" the regression, SSE the sum of squares error, n the number of observations, and k the number of independent variables. Inserting these values in formula (14–4) gives:

$$F = \frac{\text{SSR}/k}{\text{SSE}/[n - (k + 1)]} = \frac{171{,}220/3}{41{,}695/[20 - (3 + 1)]} = 21.90$$

The critical value of F is found in Appendix G. Using the table for the .05 significance level, move horizontally to 3 degrees of freedom in the numerator, then down to 16 degrees of freedom in the denominator, and read the critical value. It is 3.24. The region where H_0 is not rejected and the region where H_0 is rejected are shown in the following diagram.



Continuing with the global test, the decision rule is: Do not reject the null hypothesis that all the regression coefficients are 0 if the computed value of F is less than or equal to 3.24. If the computed F is greater than 3.24, reject H_0 and accept the alternate hypothesis, H_1.

The computed value of F is 21.90, which is in the rejection region. The null hypothesis that all the multiple regression coefficients are zero is therefore rejected. The p-value is 0.000 from the above analysis of variance table, so it is quite unlikely that H_0 is true. The null hypothesis is rejected, indicating that not all the regression coefficients are zero. From a practical standpoint, this means that some of the independent variables (amount of insulation, etc.) do have the ability to explain the variation in the dependent variable (heating cost). We expected this decision. Logically, the outside temperature, the amount of insulation, and age of the furnace have a great bearing on heating costs. The global test assures us that they do.

Evaluating Individual Regression Coefficients

So far we have shown that some, but not necessarily all, of the regression coefficients are not equal to zero and thus useful for predictions. The next step is to test the variables *individually* to determine which regression coefficients may be 0 and which are not.

Why is it important to find whether it is possible that any of the βs equal 0? If a β could equal 0, it implies that this particular independent variable is of no value in explaining any variation in the dependent value. If there are coefficients for which H_0 cannot be rejected, we may want to eliminate them from the regression equation.

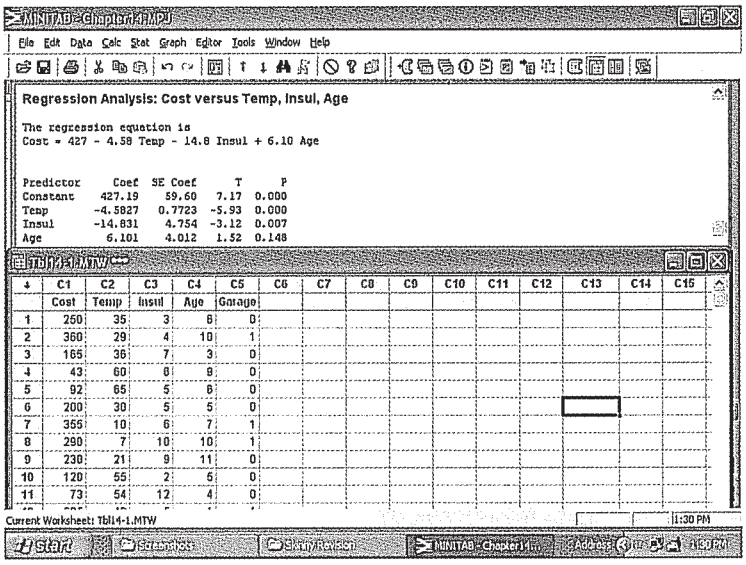
We will now conduct three separate tests of hypothesis—for temperature, for insulation, and for the age of the furnace.

For temperature:	For insulation:	For furnace age:
H_0: $\beta_1 = 0$	H_0: $\beta_2 = 0$	H_0: $\beta_3 = 0$
H_1: $\beta_1 \neq 0$	H_1: $\beta_2 \neq 0$	H_1: $\beta_3 \neq 0$

We will test the hypotheses at the .05 level. The way the alternate hypothesis is stated indicates that the test is two-tailed.

The test statistic follows the Student t distribution with $n - (k + 1)$ degrees of freedom. The number of sample observations is n. There are 20 homes in the study, so $n = 20$. The number of independent variables is k, which is 3. Thus, there are $n - (k + 1) = 20 - (3 + 1) = 16$ degrees of freedom.

The critical value for t is in Appendix F. For a two-tailed test with 16 degrees of freedom using the .05 significance level, H_0 is rejected if t is less than -2.120 or greater than 2.120. The MINITAB software produced the following output.



File Edit Data Calc Stat Graph Editor Tools Window Help

Regression Analysis: Cost versus Temp, Insul, Age

The regression equation is
Cost = 427 - 4.58 Temp - 14.8 Insul + 6.10 Age

Predictor	Coef	SE Coef	T	P
Constant	427.19	59.60	7.17	0.000
Temp	-4.5827	0.7723	-5.93	0.000
Insul	-14.831	4.754	-3.12	0.007
Age	6.101	4.012	1.52	0.148

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
	Cost	Temp	Insul	Age	Garage										
1	250	35	3	6	0										
2	360	29	4	10	1										
3	165	36	7	3	0										
4	43	60	6	6	0										
5	92	65	5	6	0										
6	200	30	5	5	0										
7	355	10	6	7	1										
8	290	7	10	10	1										
9	230	21	9	11	0										
10	120	55	2	5	0										
11	73	54	12	4	0										

Current Worksheet: Tbl14-1.MTW

The column headed "Coef" shows the regression coefficients for the multiple regression equation:

$$Y' = 427.19 - 4.5827X_1 - 14.831X_2 + 6.101X_3$$

Interpreting the term $-4.5827X_1$ in the equation: For each degree the temperature increases, it is expected that the heating cost will decrease about $4.58, holding the two other variables constant.

The column on the MINITAB output labeled "SE Coef" indicates the standard error of the sample regression coefficient. Recall that Salsberry Realty selected a sample of 20 homes along the east coast of the United States. If they were to select a second sample at random and compute the regression coefficients of that sample, the values would not be exactly the same. If they repeated the sampling process many times, however, we could design a sampling distribution of these regression coefficients. The column labeled "SE Coef" estimates the variability of these regression coefficients. The sampling distribution of Coef/SE Coef follows the t distribution with $n - (k + 1)$ degrees of freedom. Hence, we are able to test the independent variables individually to determine whether the net regression coefficients differ from zero. The computed t ratio is -5.93 for temperature and -3.12 for insulation. Both of these t values are in the rejection region to the left of -2.120. Thus, we conclude that the regression coefficients for the temperature and insulation variables are *not* zero. The computed t for age of the furnace is 1.52, so we conclude that β_3 could equal 0. The independent variable "age of the furnace" is not a significant predictor of heating cost. It can be dropped from the analysis. We can test individual regression coefficients using the t distribution. The formula is:

TESTING INDIVIDUAL REGRESSION COEFFICIENTS

$$t = \frac{b_i - 0}{s_{b_i}} \qquad [14\text{–}5]$$

The b_i refers to any one of the net regression coefficients and s_{b_i} refers to standard deviation of the net regression coefficient. We include 0 in the equation because the null hypothesis is $\beta_i = 0$.

To illustrate this formula, refer to the test of the regression coefficient for the independent variable Temperature. We let b_1 refer to the net regression coefficient. From the computer output on page 436 it is -4.5827. s_{b_i} is the standard deviation of the sampling distribution of the net regression coefficient for the independent variable Temperature. Again, from the computer output on page 436, it is 0.7723. Inserting these values in formula (14–5):

$$t = \frac{b_1 - 0}{s_{b_1}} = \frac{-4.5827 - 0}{0.7723} = -5.93$$

This is the value found in the "T" column of the output.

In Self-Review 14–3, we run the multiple regression example again using MINITAB, but only two variables—"temperature" and "insulation"—are included. These two variables explained 77.6 percent of the variation in heating cost. Using all three variables—temperature, insulation, and furnace age—a total of 80.4 percent of the variation is explained. The additional variable increased R^2 by only 2.8 percent—a rather small increase for the addition of an independent variable.

At this point we should also develop a strategy for deleting independent variables. In the Salsberry Realty case there were three independent variables and one (age) had a regression coefficient that did not differ from 0. It is clear that we should drop that variable. So we delete that variable and rerun the regression equation. However, in some instances it may not be as clear-cut which variable to delete.

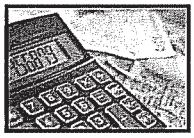
To explain, suppose we developed a multiple regression equation based on five independent variables. We conducted the global test and found that some of the

regression coefficients were different from zero. Next, we tested the regression coefficients individually and found that three were significant and two were not. The preferred procedure is to drop the single independent variable with the *smallest absolute* t *value* or *largest* p-*value* and rerun the regression equation with the four remaining variables. Then, on the new regression equation with four independent variables, conduct the individual tests. If there are still regression coefficients that are not significant, again drop the variable with the smallest absolute *t* value. To describe the process in another way, we should delete only one variable at a time. Each time we delete a variable, we need to rerun the regression equation and check the remaining variables.

This process of selecting variables to include in a regression model can be automated, using Excel, MINITAB, Megastat, or other statistical software. Most of the software systems include methods to sequentially remove and/or add independent variables and at the same time provide estimates of the percentage of variation explained (the *R*-square term). Two of the common methods are **stepwise regression** and **best subset regression.** It may take a long time, but in the extreme you could compute every regression between the dependent variable and any possible subset of the independent variables.

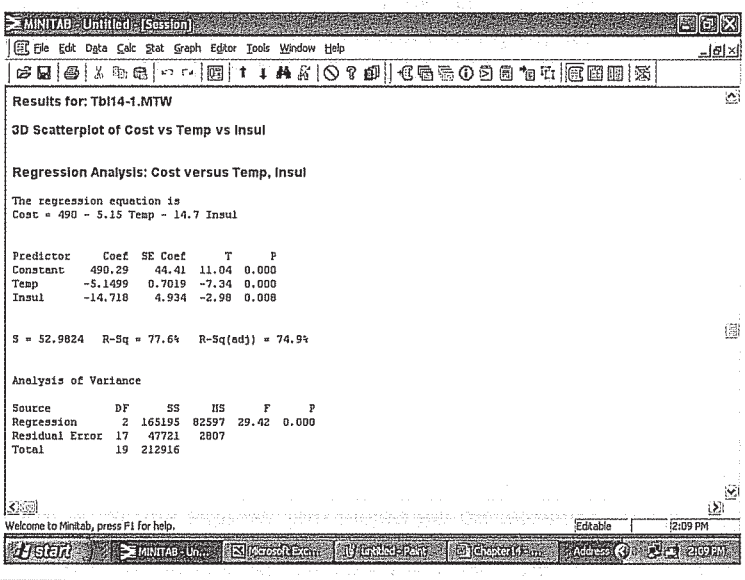
Unfortunately, on occasion, the software may work "too hard" to find an equation that fits all the quirks of your particular data set. The resultant equation may not represent the relationship in the population. You will need to use your judgment to choose among the equations presented. Consider whether the results are logical. They should have a simple interpretation and be consistent with your knowledge of the application under study.

Self-Review 14–3



The multiple regression and correlation data for the preceding heating cost example were rerun using only the first two significant independent variables—temperature and insulation. (See the following MINITAB output.)

(a) What is the new multiple regression equation? (Temperature is X_1 and insulation X_2.)
(b) What is the coefficient of multiple determination? Interpret.
(c) How can you tell that these two independent variables are of value in predicting heating costs?
(d) What is the *p*-value of insulation? Interpret.



```
MINITAB - Untitled - [Session]
File Edit Data Calc Stat Graph Editor Tools Window Help

Results for: Tbl14-1.MTW

3D Scatterplot of Cost vs Temp vs Insul

Regression Analysis: Cost versus Temp, Insul

The regression equation is
Cost = 490 - 5.15 Temp - 14.7 Insul

Predictor      Coef  SE Coef      T      P
Constant     490.29    44.41  11.04  0.000
Temp        -5.1499   0.7019  -7.34  0.000
Insul       -14.718    4.934  -2.98  0.008

S = 52.9824   R-Sq = 77.6%   R-Sq(adj) = 74.9%

Analysis of Variance

Source          DF      SS     MS      F      P
Regression       2  165195  82597  29.42  0.000
Residual Error  17   47721   2807
Total           19  212916

Welcome to Minitab, press F1 for help.                    Editable   2:09 PM
start  MINITAB - Un...  Microsoft Exc...  Untitled - Paint  Chapter 14 ...  Address  2:09 PM
```