



CHAPTER 10

Correlation and Regression

Objectives

After completing this chapter, you should be able to

- 1 Draw a scatter plot for a set of ordered pairs.
- 2 Compute the correlation coefficient.
- 3 Test the hypothesis $H_0: \rho = 0$.
- 4 Compute the equation of the regression line.
- 5 Compute the coefficient of determination.
- 6 Compute the standard error of the estimate.
- 7 Find a prediction interval.
- 8 Be familiar with the concept of multiple regression.

Outline

- 10-1 Introduction
- 10-2 Scatter Plots
- 10-3 Correlation
- 10-4 Regression
- 10-5 Coefficient of Determination and Standard Error of the Estimate
- 10-6 Multiple Regression (Optional)
- 10-7 Summary



Statistics Today

Do Dust Storms Affect Respiratory Health?

Southeast Washington state has a long history of seasonal dust storms. Several researchers decided to see what effect, if any, these storms had on the respiratory health of the people living in the area. They undertook (among other things) to see if there was a relationship between the amount of dust and sand particles in the air when the storms occur and the number of hospital emergency room visits for respiratory disorders at three community hospitals in southeast Washington. Using methods of correlation and regression, which are explained in this chapter, they were able to determine the effect of these dust storms on local residents. See *Statistics Today—Revisited*.

Source: B. Hefflin, B. Jalaludin, N. Cobb, C. Johnson, L. Jecha, and R. Etzel, "Surveillance for Dust Storms and Respiratory Diseases in Washington State, 1991," *Archives of Environmental Health* 49, no. 3 (May–June 1994), pp. 170–74. Reprinted with permission of the Helen Dwight Reid Education Foundation. Published by Heldref Publications, 1319 18th St. N.W., Washington, D.C. 20036-1802. Copyright 1994.

10-1

Introduction

In Chapters 7 and 8, two areas of inferential statistics—confidence intervals and hypothesis testing—were explained. Another area of inferential statistics involves determining whether a relationship between two or more numerical or quantitative variables exists. For example, a businessperson may want to know whether the volume of sales for a given month is related to the amount of advertising the firm does that month. Educators are interested in determining whether the number of hours a student studies is related to the student's score on a particular exam. Medical researchers are interested in questions such as, Is caffeine related to heart damage? or Is there a relationship between a person's age and his or her blood pressure? A zoologist may want to know whether the birth weight of a certain animal is related to its life span. These are only a few of the many questions that can be answered by using the techniques of correlation and regression analysis. **Correlation** is a statistical method used to determine whether a relationship between

variables exists. **Regression** is a statistical method used to describe the nature of the relationship between variables, that is, positive or negative, linear or nonlinear.

The purpose of this chapter is to answer these questions statistically:

1. Are two or more variables related?
2. If so, what is the strength of the relationship?
3. What type of relationship exists?
4. What kind of predictions can be made from the relationship?

Unusual Stat

A person walks on average 100,000 miles in his or her lifetime. This is about 3.4 miles per day.

To answer the first two questions, statisticians use a numerical measure to determine whether two or more variables are related and to determine the strength of the relationship between or among the variables. This measure is called a *correlation coefficient*. For example, there are many variables that contribute to heart disease, among them lack of exercise, smoking, heredity, age, stress, and diet. Of these variables, some are more important than others; therefore, a physician who wants to help a patient must know which factors are most important.

To answer the third question, one must ascertain what type of relationship exists. There are two types of relationships: *simple* and *multiple*. In a **simple relationship**, there are two variables—an **independent variable**, also called an explanatory variable or a predictor variable, and a **dependent variable**, also called a response variable. A simple relationship analysis is called *simple regression*, and there is one independent variable that is used to predict the dependent variable. For example, a manager may wish to see whether the number of years the salespeople have been working for the company has anything to do with the amount of sales they make. This type of study involves a simple relationship, since there are only two variables—years of experience and amount of sales.

In a **multiple relationship**, called *multiple regression*, two or more independent variables are used to predict one dependent variable. For example, an educator may wish to investigate the relationship between a student's success in college and factors such as the number of hours devoted to studying, the student's GPA, and the student's high school background. This type of study involves several variables.

Simple relationships can also be positive or negative. A **positive relationship** exists when both variables increase or decrease at the same time. For instance, a person's height and weight are related; and the relationship is positive, since the taller a person is, generally, the more the person weighs. In a **negative relationship**, as one variable increases, the other variable decreases, and vice versa. For example, if one measures the strength of people over 60 years of age, one will find that as age increases, strength generally decreases. The word *generally* is used here because there are exceptions.

Finally, the fourth question asks what type of predictions can be made. Predictions are made in all areas and daily. Examples include weather forecasting, stock market analyses, sales predictions, crop predictions, gasoline predictions, and sports predictions. Some predictions are more accurate than others, due to the strength of the relationship. That is, the stronger the relationship is between variables, the more accurate the prediction is.

10–2

Scatter Plots

Objective 1

Draw a scatter plot for a set of ordered pairs.

In simple correlation and regression studies, the researcher collects data on two numerical or quantitative variables to see whether a relationship exists between the variables. For example, if a researcher wishes to see whether there is a relationship between number of hours of study and test scores on an exam, she must select a random sample of

students, determine the hours each studied, and obtain their grades on the exam. A table can be made for the data, as shown here.

Student	Hours of study x	Grade y (%)
A	6	82
B	2	63
C	1	57
D	5	88
E	2	68
F	3	75

As stated previously, the two variables for this study are called the independent variable and the dependent variable. The independent variable is the variable in regression that can be controlled or manipulated. In this case, “number of hours of study” is the independent variable and is designated as the x variable. The dependent variable is the variable in regression that cannot be controlled or manipulated. The grade the student received on the exam is the dependent variable, designated as the y variable. The reason for this distinction between the variables is that one assumes that the grade the student earns *depends* on the number of hours the student studied. Also, one assumes that, to some extent, the student can regulate or *control* the number of hours he or she studies for the exam.

The determination of the x and y variables is not always clear-cut and is sometimes an arbitrary decision. For example, if a researcher studies the effects of age on a person’s blood pressure, the researcher can generally assume that age affects blood pressure. Hence, the variable *age* can be called the *independent variable*, and the variable *blood pressure* can be called the *dependent variable*. On the other hand, if a researcher is studying the attitudes of husbands on a certain issue and the attitudes of their wives on the same issue, it is difficult to say which variable is the independent variable and which is the dependent variable. In this study, the researcher can arbitrarily designate the variables as independent and dependent.

The independent and dependent variables can be plotted on a graph called a *scatter plot*. The independent variable x is plotted on the horizontal axis, and the dependent variable y is plotted on the vertical axis.

A **scatter plot** is a graph of the ordered pairs (x, y) of numbers consisting of the independent variable x and the dependent variable y .

The scatter plot is a visual way to describe the nature of the relationship between the independent and dependent variables. The scales of the variables can be different, and the coordinates of the axes are determined by the smallest and largest data values of the variables.

The procedure for drawing a scatter plot is shown in Examples 10–1 through 10–3.

Example 10–1



Construct a scatter plot for the data obtained in a study of age and systolic blood pressure of six randomly selected subjects. The data are shown in the table.

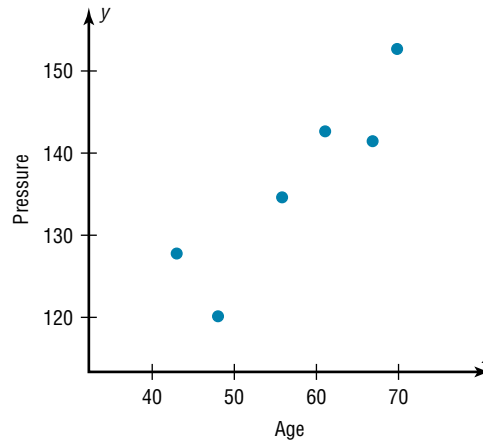
Subject	Age x	Pressure y
A	43	128
B	48	120
C	56	135
D	61	143
E	67	141
F	70	152

Solution

Step 1 Draw and label the x and y axes.

Step 2 Plot each point on the graph, as shown in Figure 10-1.

Figure 10-1
Scatter Plot for
Example 10-1



Example 10-2



Construct a scatter plot for the data obtained in a study on the number of absences and the final grades of seven randomly selected students from a statistics class. The data are shown here.

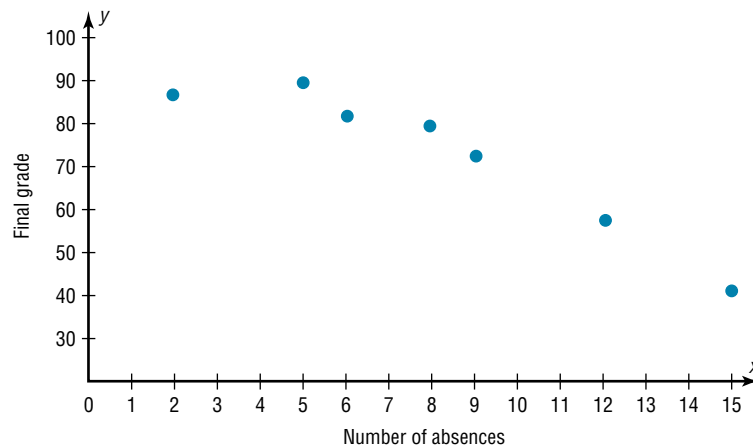
Student	Number of absences x	Final grade y (%)
A	6	82
B	2	86
C	15	43
D	9	74
E	12	58
F	5	90
G	8	78

Solution

Step 1 Draw and label the x and y axes.

Step 2 Plot each point on the graph, as shown in Figure 10-2.

Figure 10-2
Scatter Plot for
Example 10-2



Example 10–3

Construct a scatter plot for the data obtained in a study on the number of hours that nine people exercise each week and the amount of milk (in ounces) each person consumes per week. The data are shown.

Subject	Hours x	Amount y
A	3	48
B	0	8
C	2	32
D	5	64
E	8	10
F	5	32
G	10	56
H	2	72
I	1	48

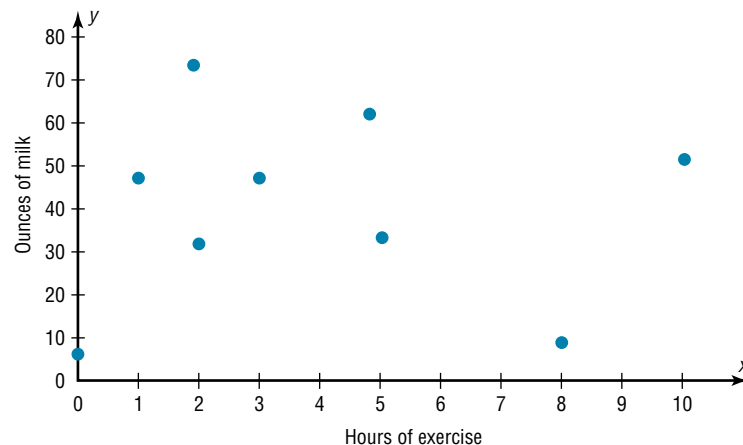
Solution

Step 1 Draw and label the x and y axes.

Step 2 Plot each point on the graph, as shown in Figure 10–3.

Figure 10–3

Scatter Plot for Example 10–3

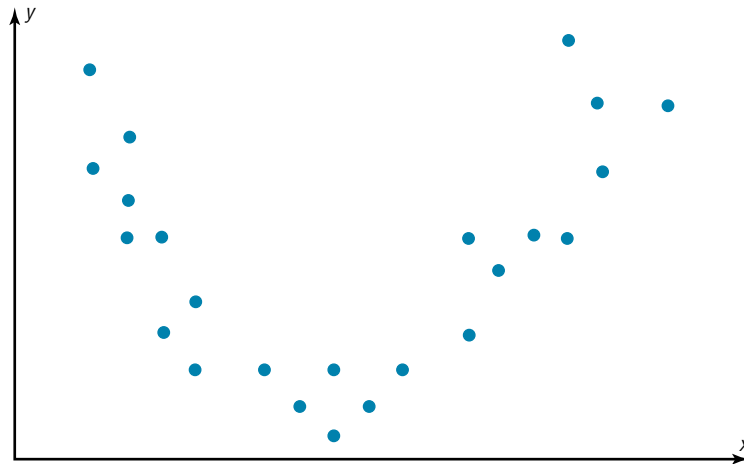


After the plot is drawn, it should be analyzed to determine which type of relationship, if any, exists. For example, the plot shown in Figure 10–1 suggests a positive relationship, since as a person's age increases, blood pressure tends to increase also. The plot of the data shown in Figure 10–2 suggests a negative relationship, since as the number of absences increases, the final grade decreases. Finally, the plot of the data shown in Figure 10–3 shows no specific type of relationship, since no pattern is discernible.

Note that the data shown in Figures 10–1 and 10–2 also suggest a linear relationship, since the points seem to fit a straight line, although not perfectly. Sometimes a scatter plot, such as the one in Figure 10–4, shows a curvilinear relationship between the data. In this situation, the methods shown in this section and in Section 10–3 cannot be used. Methods for curvilinear relationships are beyond the scope of this book.

Figure 10–4

Scatter Plot Suggesting a Curvilinear Relationship



10–3

Correlation

Correlation Coefficient

Objective 2

Compute the correlation coefficient.

As stated in Section 10–1, statisticians use a measure called the *correlation coefficient* to determine the strength of the relationship between two variables. There are several types of correlation coefficients. The one explained in this section is called the **Pearson product moment correlation coefficient** (PPMC), named after statistician Karl Pearson, who pioneered the research in this area.

The **correlation coefficient** computed from the sample data measures the strength and direction of a linear relationship between two variables. The symbol for the sample correlation coefficient is r . The symbol for the population correlation coefficient is ρ (Greek letter rho).

The *range of the correlation coefficient* is from -1 to $+1$. If there is a *strong positive linear relationship* between the variables, the value of r will be close to $+1$. If there is a *strong negative linear relationship* between the variables, the value of r will be close to -1 . When there is no linear relationship between the variables or only a weak relationship, the value of r will be close to 0 . See Figure 10–5.

The graphs in Figure 10–6 show the relationship between the correlation coefficients and their corresponding scatter plots. Notice that as the value of the correlation coefficient increases from 0 to $+1$ (parts *a*, *b*, and *c*), data values become closer to an increasingly stronger relationship. As the value of the correlation coefficient decreases from 0 to -1 (parts *d*, *e*, and *f*), the data values also become closer to a straight line. Again this suggests a stronger relationship.

Figure 10–5

Range of Values for the Correlation Coefficient

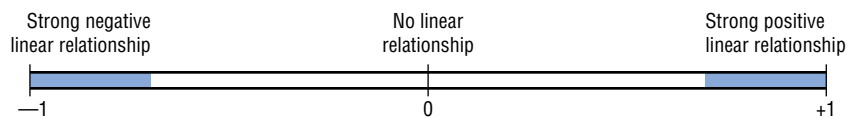
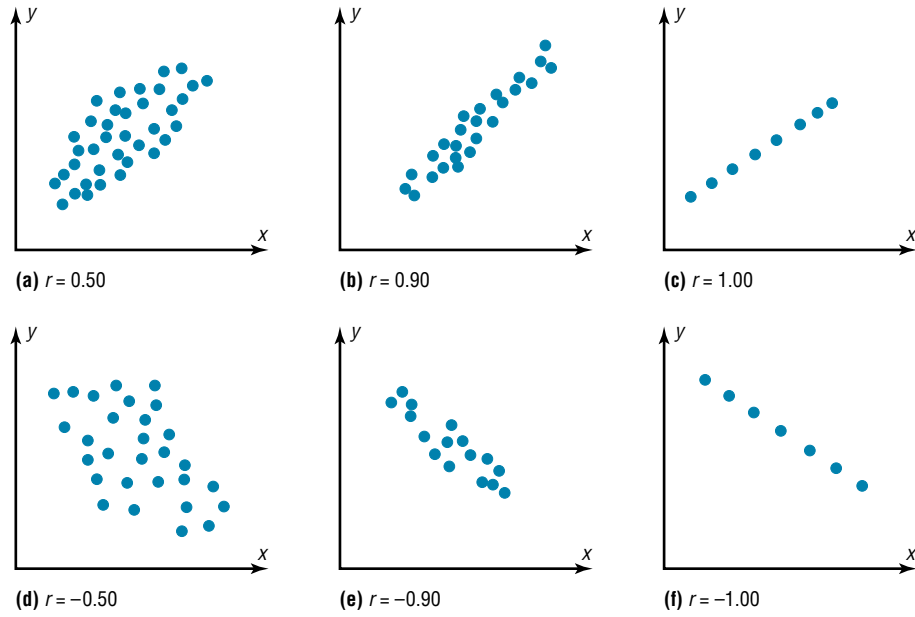


Figure 10–6
Relationship Between the Correlation Coefficient and the Scatter Plot



There are several ways to compute the value of the correlation coefficient. One method is to use the formula shown here.

Formula for the Correlation Coefficient r

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

where n is the number of data pairs.

Rounding Rule for the Correlation Coefficient Round the value of r to three decimal places.

The formula looks somewhat complicated, but using a table to compute the values, as shown in Example 10–4, makes it somewhat easier to determine the value of r .

Example 10–4

Compute the value of the correlation coefficient for the data obtained in the study of age and blood pressure given in Example 10–1.

Solution

Step 1 Make a table, as shown here.

Subject	Age x	Pressure y	xy	x^2	y^2
A	43	128			
B	48	120			
C	56	135			
D	61	143			
E	67	141			
F	70	152			

Step 2 Find the values of xy , x^2 , and y^2 and place these values in the corresponding columns of the table.

The completed table is shown.

Subject	Age x	Pressure y	xy	x^2	y^2
A	43	128	5,504	1,849	16,384
B	48	120	5,760	2,304	14,400
C	56	135	7,560	3,136	18,225
D	61	143	8,723	3,721	20,449
E	67	141	9,447	4,489	19,881
F	70	152	10,640	4,900	23,104
	$\Sigma x = 345$	$\Sigma y = 819$	$\Sigma xy = 47,634$	$\Sigma x^2 = 20,399$	$\Sigma y^2 = 112,443$

Step 3 Substitute in the formula and solve for r .

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}}$$

$$= \frac{(6)(47,634) - (345)(819)}{\sqrt{[(6)(20,399) - (345)^2][(6)(112,443) - (819)^2]}} = 0.897$$

The correlation coefficient suggests a strong positive relationship between age and blood pressure.

Example 10-5

Compute the value of the correlation coefficient for the data obtained in the study of the number of absences and the final grade of the seven students in the statistics class given in Example 10-2.

Solution

Step 1 Make a table.

Step 2 Find the values of xy , x^2 , and y^2 and place these values in the corresponding columns of the table.

Student	Number of absences x	Final grade y (%)	xy	x^2	y^2
A	6	82	492	36	6,724
B	2	86	172	4	7,396
C	15	43	645	225	1,849
D	9	74	666	81	5,476
E	12	58	696	144	3,364
F	5	90	450	25	8,100
G	8	78	624	64	6,084
	$\Sigma x = 57$	$\Sigma y = 511$	$\Sigma xy = 3745$	$\Sigma x^2 = 579$	$\Sigma y^2 = 38,993$

Step 3 Substitute in the formula and solve for r .

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}}$$

$$= \frac{(7)(3745) - (57)(511)}{\sqrt{[(7)(579) - (57)^2][(7)(38,993) - (511)^2]}} = -0.944$$

The value of r suggests a strong negative relationship between a student's final grade and the number of absences a student has. That is, the more absences a student has, the lower is his or her grade.

Example 10–6

Compute the value of the correlation coefficient for the data given in Example 10–3 for the number of hours a person exercises and the amount of milk a person consumes per week.

Solution

Step 1 Make a table.

Step 2 Find the values of xy , x^2 , and y^2 and place these values in the corresponding columns of the table.

Subject	Hours x	Amount y	xy	x^2	y^2
A	3	48	144	9	2,304
B	0	8	0	0	64
C	2	32	64	4	1,024
D	5	64	320	25	4,096
E	8	10	80	64	100
F	5	32	160	25	1,024
G	10	56	560	100	3,136
H	2	72	144	4	5,184
I	1	48	48	1	2,304
	$\Sigma x = 36$	$\Sigma y = 370$	$\Sigma xy = 1,520$	$\Sigma x^2 = 232$	$\Sigma y^2 = 19,236$

Step 3 Substitute in the formula and solve for r .

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}}$$

$$= \frac{(9)(1520) - (36)(370)}{\sqrt{[(9)(232) - (36)^2][(9)(19,236) - (370)^2]}} = 0.067$$

The value of r indicates a very weak positive relationship between the variables.

In Example 10–4, the value of r was high (close to 1.00); in Example 10–6, the value of r was much lower (close to 0). This question then arises: When is the value of r due to chance, and when does it suggest a significant linear relationship between the variables? This question will be answered next.

Objective 3

Test the hypothesis $H_0: \rho = 0$.

The Significance of the Correlation Coefficient

As stated before, the range of the correlation coefficient is between -1 and $+1$. When the value of r is near $+1$ or -1 , there is a strong linear relationship. When the value of r is near 0 , the linear relationship is weak or nonexistent. Since the value of r is computed from data obtained from samples, there are two possibilities when r is not equal to zero: either the value of r is high enough to conclude that there is a significant linear relationship between the variables, or the value of r is due to chance.

To make this decision, one uses a hypothesis-testing procedure. The traditional method is similar to the one used in previous chapters.

- Step 1** State the hypotheses.
Step 2 Find the critical values.
Step 3 Compute the test value.
Step 4 Make the decision.
Step 5 Summarize the results.

The population correlation coefficient is computed from taking all possible (x, y) pairs; it is designated by the Greek letter ρ (rho). The sample correlation coefficient can then be used as an estimator of ρ if the following assumptions are valid.

1. The variables x and y are *linearly* related.
2. The variables are *random* variables.
3. The two variables have a *bivariate normal distribution*.

This means for any given value of x , the y variable is normally distributed.

Formally defined, the **population correlation coefficient** ρ is the correlation computed by using all possible pairs of data values (x, y) taken from a population.

Interesting Fact

Scientists think that a person is never more than 3 feet away from a spider at any given time!

In hypothesis testing, one of these is true:

- $H_0: \rho = 0$ This null hypothesis means that there is no correlation between the x and y variables in the population.
 $H_1: \rho \neq 0$ This alternative hypothesis means that there is a significant correlation between the variables in the population.

When the null hypothesis is rejected at a specific level, it means that there is a significant difference between the value of r and 0. When the null hypothesis is not rejected, it means that the value of r is not significantly different from 0 (zero) and is probably due to chance.

Several methods can be used to test the significance of the correlation coefficient. Three methods will be shown in this section. The first uses the t test.

Historical Notes

A mathematician named Karl Pearson (1857–1936) became interested in Francis Galton's work and saw that the correlation and regression theory could be applied to other areas besides heredity. Pearson developed the correlation coefficient that bears his name.

Formula for the t Test for the Correlation Coefficient

$$t = r \sqrt{\frac{n-2}{1-r^2}}$$

with degrees of freedom equal to $n - 2$.

Although hypothesis tests can be one-tailed, most hypotheses involving the correlation coefficient are two-tailed. Recall that ρ represents the population correlation coefficient. Also, if there is no linear relationship, the value of the correlation coefficient will be 0. Hence, the hypotheses will be

$$H_0: \rho = 0 \quad \text{and} \quad H_1: \rho \neq 0$$

One does not have to identify the claim here, since the question will always be whether there is a significant linear relationship between the variables.

The two-tailed critical values are used. These values are found in Table F in Appendix C. Also, when one is testing the significance of a correlation coefficient, both variables x and y must come from normally distributed populations.

Example 10–7

Test the significance of the correlation coefficient found in Example 10–4. Use $\alpha = 0.05$ and $r = 0.897$.

Solution

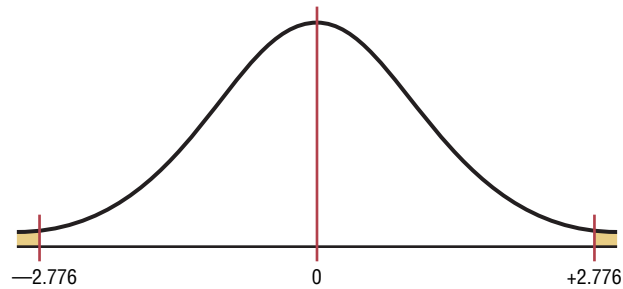
Step 1 State the hypotheses.

$$H_0: \rho = 0 \quad \text{and} \quad H_1: \rho \neq 0$$

Step 2 Find the critical values. Since $\alpha = 0.05$ and there are $6 - 2 = 4$ degrees of freedom, the critical values obtained from Table F are ± 2.776 , as shown in Figure 10–7.

Figure 10–7

Critical Values for Example 10–7



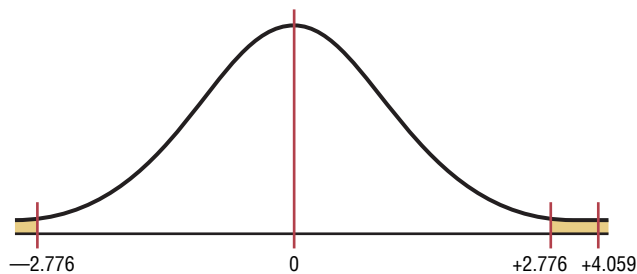
Step 3 Compute the test value.

$$t = r\sqrt{\frac{n-2}{1-r^2}} = (0.897)\sqrt{\frac{6-2}{1-(0.897)^2}} = 4.059$$

Step 4 Make the decision. Reject the null hypothesis, since the test value falls in the critical region, as shown in Figure 10–8.

Figure 10–8

Test Value for Example 10–7



Step 5 Summarize the results. There is a significant relationship between the variables of age and blood pressure.

The second method that can be used to test the significance of r is the P -value method. The method is the same as that shown in Chapters 8 and 9. It uses the following steps.

Step 1 State the hypotheses.

Step 2 Find the test value. (In this case, use the t test.)

Figure 10–9
Finding the Critical Value from Table I

d.f.	$\alpha = 0.05$	$\alpha = 0.01$
1		
2		
3		
4		
5		
6		
7	0.666	

Step 3 Find the *P*-value. (In this case, use Table F.)

Step 4 Make the decision.

Step 5 Summarize the results.

Referring to Example 10–7, we see that the *t* value obtained in step 3 is 4.059 and d.f. = 4. Using Table F with d.f. = 4 and the row Two tails, the value 4.059 falls between 3.747 and 4.604; hence, $0.01 < P\text{-value} < 0.02$. (The *P*-value obtained from a calculator is 0.015.) That is, the *P*-value falls between 0.01 and 0.02. The decision then is to reject the null hypothesis since $P\text{-value} < 0.05$.

The third method of testing the significance of *r* is to use Table I in Appendix C. This table shows the values of the correlation coefficient that are significant for a specific α level and a specific number of degrees of freedom. For example, for 7 degrees of freedom and $\alpha = 0.05$, the table gives a critical value of 0.666. Any value of *r* greater than +0.666 or less than –0.666 will be significant, and the null hypothesis will be rejected. See Figure 10–9. When Table I is used, one need not compute the *t* test value. Table I is for two-tailed tests only.

Example 10–8

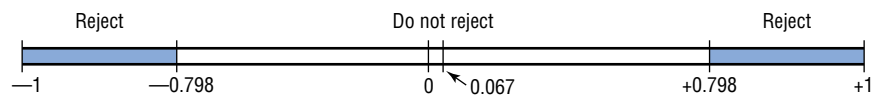
Using Table I, test the significance of the correlation coefficient $r = 0.067$, obtained in Example 10–6, at $\alpha = 0.01$.

Solution

$$H_0: \rho = 0 \quad \text{and} \quad H_1: \rho \neq 0$$

Since the sample size is 9, there are 7 degrees of freedom. When $\alpha = 0.01$ and with 7 degrees of freedom, the value obtained from Table I is 0.798. For a significant relationship, a value of *r* greater than +0.798 or less than –0.798 is needed. Since $r = 0.067$, the null hypothesis is not rejected. Hence, there is not enough evidence to say that there is a significant linear relationship between the variables. See Figure 10–10.

Figure 10–10
Rejection and Nonrejection Regions for Example 10–8





Correlation and Causation

Researchers must understand the nature of the linear relationship between the independent variable x and the dependent variable y . When a hypothesis test indicates that a significant linear relationship exists between the variables, researchers must consider the possibilities outlined next.

Possible Relationships Between Variables

When the null hypothesis has been rejected for a specific α value, any of the following five possibilities can exist.

1. *There is a direct cause-and-effect relationship between the variables.* That is, x causes y . For example, water causes plants to grow, poison causes death, and heat causes ice to melt.
2. *There is a reverse cause-and-effect relationship between the variables.* That is, y causes x . For example, suppose a researcher believes excessive coffee consumption causes nervousness, but the researcher fails to consider that the reverse situation may occur. That is, it may be that an extremely nervous person craves coffee to calm his or her nerves.
3. *The relationship between the variables may be caused by a third variable.* For example, if a statistician correlated the number of deaths due to drowning and the number of cans of soft drink consumed daily during the summer, he or she would probably find a significant relationship. However, the soft drink is not necessarily responsible for the deaths, since both variables may be related to heat and humidity.
4. *There may be a complexity of interrelationships among many variables.* For example, a researcher may find a significant relationship between students' high school grades and college grades. But there probably are many other variables involved, such as IQ, hours of study, influence of parents, motivation, age, and instructors.
5. *The relationship may be coincidental.* For example, a researcher may be able to find a significant relationship between the increase in the number of people who are exercising and the increase in the number of people who are committing crimes. But common sense dictates that any relationship between these two values must be due to coincidence.

When two variables are highly correlated, item 3 in the box states that there exists a possibility that the correlation is due to a third variable. If this is the case and the third variable is unknown to the researcher or not accounted for in the study, it is called a **lurking variable**. An attempt should be made by the researcher to identify such variables and to use methods to control their influence.

Also, one should be cautious when the data for one or both of the variables involve averages rather than individual data. It is not wrong to use averages, but the results cannot be generalized to individuals since averaging tends to smooth out the variability among individual data values. The result could be a higher correlation than actually exists.

Thus, when the null hypothesis is rejected, the researcher must consider all possibilities and select the appropriate one as determined by the study. Remember, correlation does not necessarily imply causation.

Applying the Concepts 10–3

Stopping Distances

In a study on speed control, it was found that the main reasons for regulations were to make traffic flow more efficient and to minimize the risk of danger. An area that was focused on in the study was the distance required to completely stop a vehicle at various speeds. Use the following table to answer the questions.

MPH	Braking distance (feet)
20	20
30	45
40	81
50	133
60	205
80	411

Assume MPH is going to be used to predict stopping distance.

- Which of the two variables is the independent variable?
- Which is the dependent variable?
- What type of variable is the independent variable?
- What type of variable is the dependent variable?
- Construct a scatter plot for the data.
- Is there a linear relationship between the two variables?
- Redraw the scatter plot, and change the distances between the independent-variable numbers. Does the relationship look different?
- Is the relationship positive or negative?
- Can braking distance be accurately predicted from MPH?
- List some other variables that affect braking distance.
- Compute the value of r .
- Is r significant at $\alpha = 0.05$?

See page 580 for the answers.

Exercises 10–3

- What is meant by the statement that two variables are related?
- How is a linear relationship between two variables measured in statistics? Explain.
- What is the symbol for the sample correlation coefficient? The population correlation coefficient?
- What is the range of values for the correlation coefficient?
- What is meant when the relationship between the two variables is positive? Negative?
- Give examples of two variables that are positively correlated and two that are negatively correlated.

Speaking of Statistics

In correlation and regression studies, it is difficult to control all variables. This study shows some of the consequences when researchers overlook certain aspects in studies. Suggest ways that the extraneous variables might be controlled in future studies.

Coffee Not Disease Culprit, Study Says

NEW YORK (AP)—Two new studies suggest that coffee drinking, even up to 5½ cups per day, does not increase the risk of heart disease, and other studies that claim to have found increased risks might have missed the true culprits, a researcher says.

“It might not be the coffee cup in one hand, it might be the cigarette or coffee roll in the other,” said Dr. Peter W. F. Wilson, the author of one of the new studies.

He noted in a telephone interview Thursday that many coffee drinkers, particularly heavy coffee drinkers, are smokers. And one of the new studies found that coffee drinkers had excess fat in their diets.

The findings of the new studies conflict sharply with a study reported in November 1985 by Johns Hopkins University scientists in Baltimore.

The Hopkins scientists found that coffee drinkers who consumed five or more cups of coffee per day had three times the heart-disease risk of non-coffee drinkers.

The reason for the discrepancy appears to be that many of the coffee drinkers in the Hopkins study also smoked—and it was the

smoking that increased their heart-disease risk, said Wilson.

Wilson, director of laboratories for the Framingham Heart Study in Framingham, Mass., said Thursday at a conference sponsored by the American Heart Association in Charleston, S.C., that he had examined the coffee intake of 3,937 participants in the Framingham study during 1956–66 and an additional 2,277 during the years 1972–1982.

In contrast to the subjects in the Hopkins study, most of these coffee drinkers consumed two or three cups per day, Wilson said. Only 10 percent drank six or more cups per day.

He then looked at blood cholesterol levels and heart and blood vessel disease in the two groups. “We ran these analyses for coronary heart disease, heart attack, sudden death and stroke and in absolutely every analysis, we found no link with coffee,” Wilson said.

He found that coffee consumption was linked to a significant decrease in total blood cholesterol in men, and to a moderate increase in total cholesterol in women.

Source: Reprinted with permission of the Associated Press.

7. Give an example of a correlation study, and identify the independent and dependent variables.
8. What is the diagram of the independent and dependent variables called? Why is drawing this diagram important?
9. What is the name of the correlation coefficient used in this section?
10. What statistical test is used to test the significance of the correlation coefficient?
11. When two variables are correlated, can the researcher be sure that one variable causes the other? Why or why not?
 - d. Test the significance of the correlation coefficient at $\alpha = 0.05$, using Table I.
 - e. Give a brief explanation of the type of relationship.
12. A medical researcher wishes to see if there is a relationship between prescription drug prices for identical drugs and identical dosages that are prescribed for humans and for animals. The prices are shown. Is there a relationship between the prices?

Prices for humans x	0.67	0.64	1.20	0.51	0.87	0.74	0.50	1.22
Prices for animals y	0.13	0.18	0.42	0.25	0.57	0.57	0.49	1.28


Source: House Committee on Government Reform.

(The information in this exercise will be used for Exercise 12 in Section 10–4.)

13. A researcher wishes to determine if a person's age is related to the number of hours he or she exercises per week. The data for the sample are shown here.

(The information in this exercise will be used for Exercises 13 and 36 in Section 10-4 and Exercise 15 in Section 10-5.)


Age x	18	26	32	38	52	59
Hours y	10	5	2	3	1.5	1

-  **14.** An environmentalist wants to determine the relationships between the numbers (in thousands) of forest fires over the year and the number (in hundred thousands) of acres burned. The data for 8 recent years are shown. Describe the relationship.


Number of fires x	72	69	58	47	84	62	57	45
Number of acres burned y	62	42	19	26	51	15	30	15

Source: National Interagency Fire Center.


(The information in this exercise will be used for Exercises 14 and 36 in Section 10-4 and Exercises 16 and 20 in Section 10-5.)

-  **15.** The director of an alumni association for a small college wants to determine whether there is any type of relationship between the amount of an alumnus's contribution (in dollars) and the years the alumnus has been out of school. The data follow. (The information in this exercise will be used for Exercises 15, 36, and 37 in Section 10-4 and Exercise 17 in Section 10-5.)

Years x	1	5	3	10	7	6
Contribution y	500	100	300	50	75	80


-  **16.** A store manager wishes to find out whether there is a relationship between the age of her employees and the number of sick days they take each year. The data for the sample are shown. (The information in this exercise will be used for Exercises 16 and 37 in Section 10-4 and Exercise 18 in Section 10-5.)

Age x	18	26	39	48	53	58
Days y	16	12	9	5	6	2

-  **17.** A criminology student wishes to see if there is a relationship between the number of larceny crimes and the number of vandalism crimes on college campuses in southwestern Pennsylvania. The data are shown. Is there a relationship between the two types of crimes?


Number of larceny crimes x	24	6	16	64	10	25	35
Number of vandalism crimes y	21	3	6	15	21	61	20

(The information in this exercise will be used for Exercise 17 of Section 10-4.)

-  **18.** A football fan wishes to see how the number of pass attempts (not completions) relates to the number of yards gained for quarterbacks in past NFL season playoff games. The data are shown for five quarterbacks. Describe the relationships.

Pass attempts x	116	90	82	108	92
Yards gained y	1001	823	851	873	839


(The information in this exercise will be used for Exercises 18 and 38 in Section 10-4.)

-  **19.** A meteorologist wants to see if there is a relationship between the number of tornadoes that occur each year and the number of deaths attributed to the tornadoes. The data are shown for 10 recent years. Is there a relationship?


No. of tornadoes x	1133	1132	1297	1173	1082
No. of deaths y	53	39	39	33	69

No. of tornadoes x	1234	1148	1424	1342	898
No. of deaths y	30	67	130	94	40

(The information in this exercise will be used for Exercise 19 of Section 10-4.)


-  **20.** An emergency service wishes to see whether a relationship exists between the outside temperature and the number of emergency calls it receives for a 7-hour period. The data are shown. (The information in this exercise will be used for Exercises 20 and 38 in Section 10-4.)

Temperature x	68	74	82	88	93	99	101
No. of calls y	7	4	8	10	11	9	13

-  **21.** A random sample of U.S. cities is selected to determine if there is a relationship between the population (in thousands) of people under 5 years of age and the population (in thousands) of those 65 years of age and older. The data for the sample are shown here. (The information in this exercise will be used for Exercises 21 and 36 in Section 10-4.)


Under 5 x	178	27	878	314	322	143
65 and over y	361	72	1496	501	585	207

Source: *N.Y. Times Almanac*.

-  **22.** The results of a survey of the average monthly rents (in dollars) for one-bedroom apartments and two-bedroom apartments in randomly selected metropolitan areas are shown. Determine if there is a relationship between the rents. (The information in this exercise will be used for Exercise 22 in Section 10-4.)


One-bedroom x	782	486	451	529	618	520	845
Two-bedroom y	1223	902	739	954	1055	875	1455

Source: *N.Y. Times Almanac*.

-  **23.** The average normal daily temperature (in degrees Fahrenheit) and the corresponding average monthly precipitation (in inches) for the month of June are shown here for seven randomly selected cities in the United States. Determine if there is a relationship between the two variables. (The information in this exercise will be used for Exercise 23 in Section 10-4.)

Avg. daily temp. x	86	81	83	89	80	74	64
Avg. mo. precip. y	3.4	1.8	3.5	3.6	3.7	1.5	0.2


Source: *N.Y. Times Almanac*.

-  **24.** A random sample of Hall of Fame pitchers' career wins and their total number of strikeouts is shown next. Is there a relationship between the variables? (The information in this exercise will be used for Exercise 24 in Section 10-4.)

Wins x	329	150	236	300	284	207
Strikeouts y	4136	1155	1956	2266	3192	1277

Wins x	247	314	273	324
Strikeouts y	1068	3534	1987	3574

Source: *N.Y. Times Almanac*.

-  **25.** The number of calories and the number of milligrams of cholesterol for a random sample of fast-food chicken sandwiches from seven restaurants are shown here. Is there a relationship between the variables? (The information in this exercise will be used in Exercise 25 in Section 10-4.)

Calories x	390	535	720	300	430	500	440
Cholesterol y	43	45	80	50	55	52	60

Source: *The Doctor's Pocket Calorie, Fat, and Carbohydrate Counter*.

- 26.** An architect wants to determine the relationship between the heights (in feet) of a building and the number of stories in the building. The data for a sample of 10 buildings in Pittsburgh are shown. Explain the relationship.

Stories x	64	54	40	31	45	38	42	41	37	40
Height y	841	725	635	616	615	582	535	520	511	485

Source: *World Almanac Book of Facts*.

(The information in this exercise will be used for Exercise 26 of Section 10-4.)

- 27.** A hospital administrator wants to see if there is a relationship between the number of licensed beds and the number of staffed beds in local hospitals. The data for a specific day are shown. Describe the relationship.

Licensed beds x	144	32	175	185	208	100	169
Staffed beds y	112	32	162	141	103	80	118

Source: *Pittsburgh Tribune-Review*.


(The information in this exercise will be used for Exercise 28 of this section and Exercise 27 in Section 10-4.)

Extending the Concepts


- 28.** One of the formulas for computing r is

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{(n - 1)(s_x)(s_y)}$$

Using the data in Exercise 27, compute r with this formula. Compare the results.

-  **29.** Compute r for the data set shown. Explain the reason for this value of r . Now, interchange the values of x and y and compute r again. Compare this value with the previous one. Explain the results of the comparison.

x	1	2	3	4	5
y	3	5	7	9	11

-  **30.** Compute r for the following data and test the hypothesis $H_0: \rho = 0$. Draw the scatter plot; then explain the results.

x	-3	-2	-1	0	1	2	3
y	9	4	1	0	1	4	9

10-4

Regression

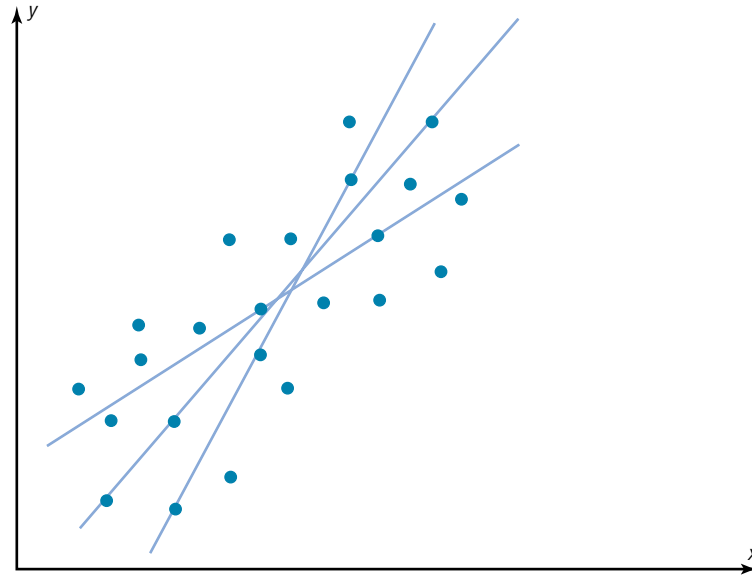
Objective 4

Compute the equation of the regression line.

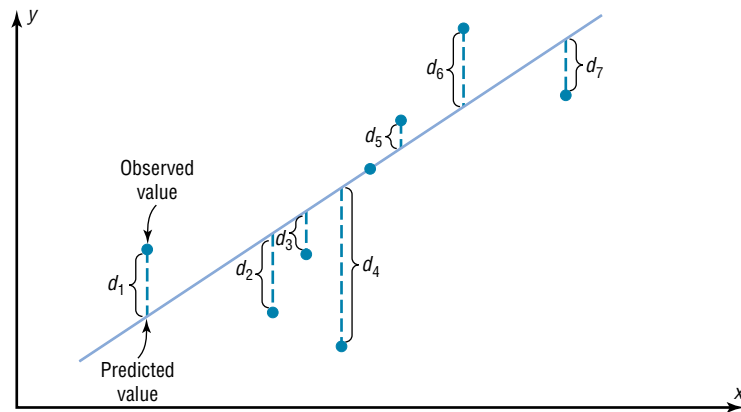
In studying relationships between two variables, collect the data and then construct a scatter plot. The purpose of the scatter plot, as indicated previously, is to determine the nature of the relationship. The possibilities include a positive linear relationship, a negative linear relationship, a curvilinear relationship, or no discernible relationship. After the scatter plot is drawn, the next steps are to compute the value of the correlation coefficient and to test the significance of the relationship. If the value of the correlation coefficient is significant, the next step is to determine the equation of the **regression line**, which is the data's line of best fit. (*Note:* Determining the regression line when r is not significant and then making predictions using the regression line are meaningless.) The purpose of

Figure 10–11

Scatter Plot with Three Lines Fit to the Data

**Figure 10–12**

Line of Best Fit for a Set of Data Points



Historical Notes

Francis Galton drew the line of best fit visually. An assistant of Karl Pearson's named G. Yule devised the mathematical solution using the least-squares method, employing a mathematical technique developed by Adrien-Marie Legendre about 100 years earlier.

the regression line is to enable the researcher to see the trend and make predictions on the basis of the data.

Line of Best Fit

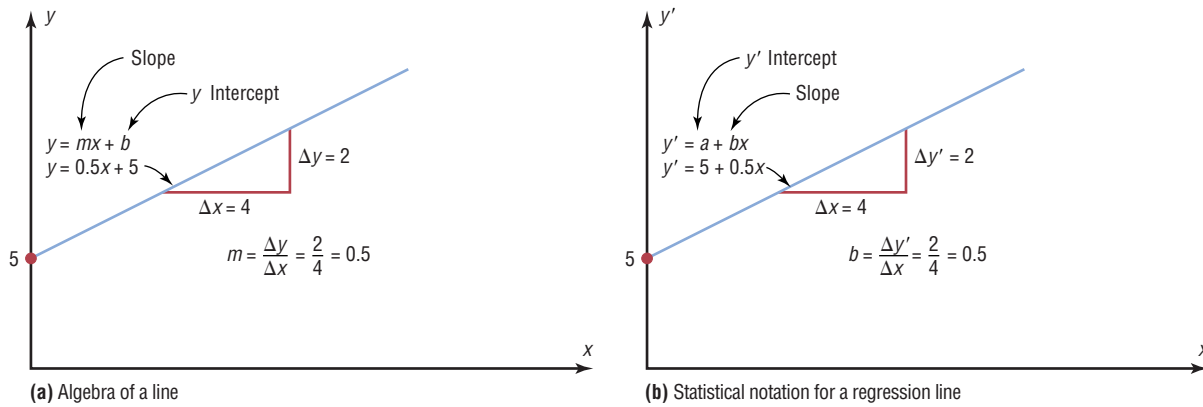
Figure 10–11 shows a scatter plot for the data of two variables. It shows that several lines can be drawn on the graph near the points. Given a scatter plot, one must be able to draw the *line of best fit*. *Best fit* means that the sum of the squares of the vertical distances from each point to the line is at a minimum. The reason one needs a line of best fit is that the values of y will be predicted from the values of x ; hence, the closer the points are to the line, the better the fit and the prediction will be. See Figure 10–12. When r is positive, the line slopes upward and to the right. When r is negative, the line slopes downward from left to right.

Determination of the Regression Line Equation

In algebra, the equation of a line is usually given as $y = mx + b$, where m is the slope of the line and b is the y intercept. (Students who need an algebraic review of the properties of a line should refer to Appendix A, Section A–3, before studying this section.) In

Figure 10–13

A Line as Represented in Algebra and in Statistics



statistics, the equation of the regression line is written as $y' = a + bx$, where a is the y' intercept and b is the slope of the line. See Figure 10–13.

There are several methods for finding the equation of the regression line. Two formulas are given here. *These formulas use the same values that are used in computing the value of the correlation coefficient.* The mathematical development of these formulas is beyond the scope of this book.

Formulas for the Regression Line $y' = a + bx$

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

where a is the y' intercept and b is the slope of the line.

Rounding Rule for the Intercept and Slope Round the values of a and b to three decimal places.

Example 10–9

Find the equation of the regression line for the data in Example 10–4, and graph the line on the scatter plot of the data.

Solution

The values needed for the equation are $n = 6$, $\Sigma x = 345$, $\Sigma y = 819$, $\Sigma xy = 47,634$, and $\Sigma x^2 = 20,399$. Substituting in the formulas, one gets

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{(819)(20,399) - (345)(47,634)}{(6)(20,399) - (345)^2} = 81.048$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{(6)(47,634) - (345)(819)}{(6)(20,399) - (345)^2} = 0.964$$

Hence, the equation of the regression line $y' = a + bx$ is

$$y' = 81.048 + 0.964x$$

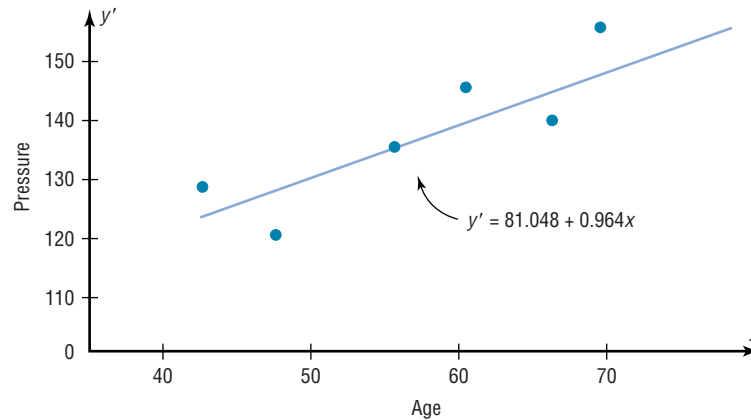
The graph of the line is shown in Figure 10-14.

Figure 10-14

Regression Line for Example 10-9

Historical Note

In 1795, Adrien-Marie Legendre (1752–1833) measured the meridian arc on the earth's surface from Barcelona, Spain, to Dunkirk, England. This measure was used as the basis for the measure of the meter. Legendre developed the least-squares method around the year 1805.



Note: When one is drawing the scatter plot and the regression line, it is sometimes desirable to *truncate* the graph (see Chapter 2). The motive is to show the line drawn in the range of the independent and dependent variables. For example, the regression line in Figure 10-14 is drawn between the x values of approximately 43 and 82 and the y' values of approximately 120 and 152. The range of the x values in the original data shown in Example 10-4 is $70 - 43 = 27$, and the range of the y' values is $152 - 120 = 32$. Notice that the x axis has been truncated; the distance between 0 and 40 is not shown in the proper scale compared to the distance between 40 and 50, 50 and 60, etc. The y' axis has been similarly truncated.

The important thing to remember is that when the x axis and sometimes the y' axis have been truncated, do not use the y' intercept value a to graph the line. To be on the safe side when graphing the regression line, use a value for x selected from the range of x values.

Example 10-10

Find the equation of the regression line for the data in Example 10-5, and graph the line on the scatter plot.

Solution

The values needed for the equation are $n = 7$, $\Sigma x = 57$, $\Sigma y = 511$, $\Sigma xy = 3745$, and $\Sigma x^2 = 579$. Substituting in the formulas, one gets

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{(511)(579) - (57)(3745)}{(7)(579) - (57)^2} = 102.493$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{(7)(3745) - (57)(511)}{(7)(579) - (57)^2} = -3.622$$

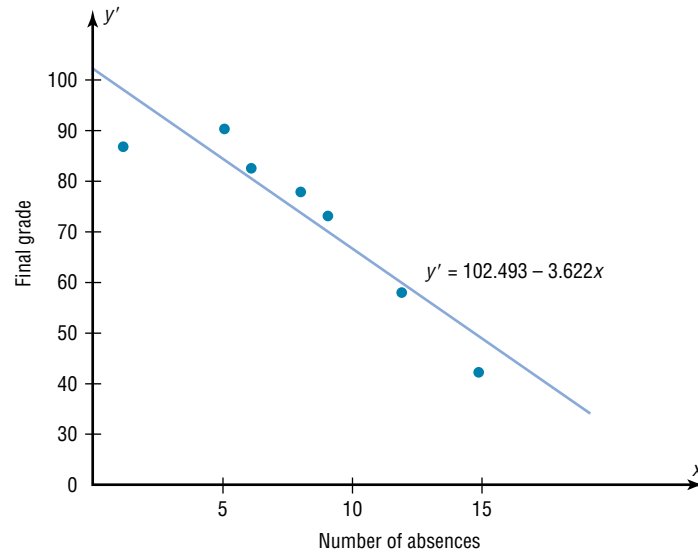
Hence, the equation of the regression line $y' = a + bx$ is

$$y' = 102.493 - 3.622x$$

The graph of the line is shown in Figure 10-15.

Figure 10–15

Regression Line for Example 10–10



The sign of the correlation coefficient and the sign of the slope of the regression line will always be the same. That is, if r is positive, then b will be positive; if r is negative, then b will be negative. The reason is that the numerators of the formulas are the same and determine the signs of r and b , and the denominators are always positive. The regression line will always pass through the point whose x coordinate is the mean of the x values and whose y coordinate is the mean of the y values, that is, (\bar{x}, \bar{y}) .

The regression line can be used to make predictions for the dependent variable. The method for making predictions is shown in Example 10–11.

Example 10–11

Using the equation of the regression line found in Example 10–9, predict the blood pressure for a person who is 50 years old.

Solution

Substituting 50 for x in the regression line $y' = 81.048 + 0.964x$ gives

$$y' = 81.048 + (0.964)(50) = 129.248 \text{ (rounded to 129)}$$

In other words, the predicted systolic blood pressure for a 50-year-old person is 129.

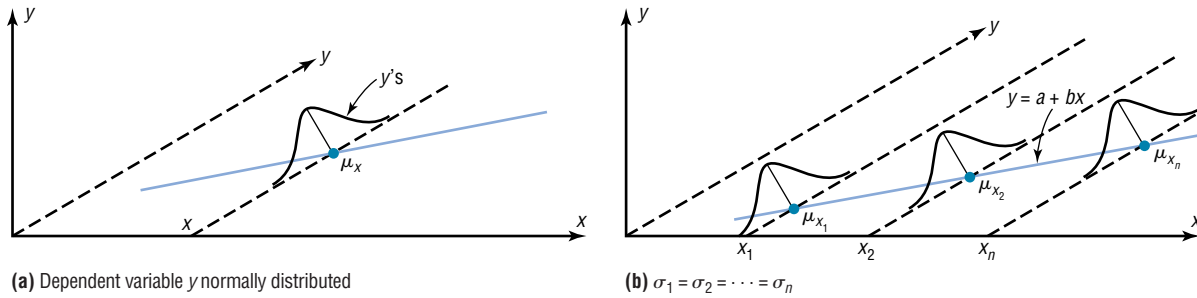
The value obtained in Example 10–11 is a point prediction, and with point predictions, no degree of accuracy or confidence can be determined. More information on prediction is given in Section 10–5.

The magnitude of the change in one variable when the other variable changes exactly 1 unit is called a **marginal change**. The value of slope b of the regression line equation represents the marginal change. For example, the slope of the regression line equation in Example 10–9 is 0.964. This means that for each increase of 1 year, the value of y (systolic blood pressure reading) changes by 0.964 unit on average. In other words, for each year a person ages, her or his blood pressure rises about 1 unit.

When r is not significantly different from 0, the best predictor of y is the mean of the data values of y . For valid predictions, the value of the correlation coefficient must be significant. Also, two other assumptions must be met.

Figure 10-16

Assumptions for Predictions



Assumptions for Valid Predictions in Regression

1. For any specific value of the independent variable x , the value of the dependent variable y must be normally distributed about the regression line. See Figure 10-16(a).
2. The standard deviation of each of the dependent variables must be the same for each value of the independent variable. See Figure 10-16(b).

Extrapolation, or making predictions beyond the bounds of the data, must be interpreted cautiously. For example, in 1979, some experts predicted that the United States would run out of oil by the year 2003. This prediction was based on the current consumption and on known oil reserves at that time. However, since then, the automobile industry has produced many new fuel-efficient vehicles. Also, there are many as yet undiscovered oil fields. Finally, science may someday discover a way to run a car on something as unlikely but as common as peanut oil. In addition, the price of a gallon of gasoline was predicted to reach \$10 a few years later. Fortunately this has not come to pass. *Remember that when predictions are made, they are based on present conditions or on the premise that present trends will continue.* This assumption may or may not prove true in the future.

The steps for finding the value of the correlation coefficient and the regression line equation are summarized in this Procedure Table:

Interesting Fact

It is estimated that wearing a motorcycle helmet reduces the risk of a fatal accident by 30%.

Procedure Table

Finding the Correlation Coefficient and the Regression Line Equation

- Step 1** Make a table, as shown in step 2.
- Step 2** Find the values of xy , x^2 , and y^2 . Place them in the appropriate columns and sum each column.

x	y	xy	x^2	y^2
·	·	·	·	·
·	·	·	·	·
·	·	·	·	·
$\Sigma x =$ <u> </u>	$\Sigma y =$ <u> </u>	$\Sigma xy =$ <u> </u>	$\Sigma x^2 =$ <u> </u>	$\Sigma y^2 =$ <u> </u>

Procedure Table (Continued)

Step 3 Substitute in the formula to find the value of r .

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

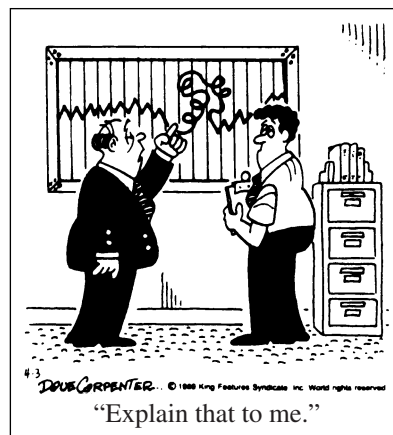
Step 4 When r is significant, substitute in the formulas to find the values of a and b for the regression line equation $y' = a + bx$.

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2} \quad b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

A scatter plot should be checked for outliers. An outlier is a point that seems out of place when compared with the other points (see Chapter 3). Some of these points can affect the equation of the regression line. When this happens, the points are called **influential points** or **influential observations**.

When a point on the scatter plot appears to be an outlier, it should be checked to see if it is an influential point. An influential point tends to “pull” the regression line toward the point itself. To check for an influential point, the regression line should be graphed with the point included in the data set. Then a second regression line should be graphed that excludes the point from the data set. If the position of the second line is changed considerably, the point is said to be an influential point. Points that are outliers in the x direction tend to be influential points.

Researchers should use their judgment as to whether to include influential observations in the final analysis of the data. If the researcher feels that the observation is not necessary, then it should be excluded so that it does not influence the results of the study. However, if the researcher feels that it is necessary, then he or she may want to obtain additional data values whose x values are near the x value of the influential point and then include them in the study.



Source: Reprinted with special permission of King Features Syndicate.

Applying the Concepts 10-4

Stopping Distances Revisited

In a study on speed and braking distance, researchers looked for a method to estimate how fast a person was traveling before an accident by measuring the length of their skid marks. An area that was focused on in the study was the distance required to completely stop a vehicle at various speeds. Use the following table to answer the questions.

MPH	Braking distance (feet)
20	20
30	45
40	81
50	133
60	205
80	411

Assume MPH is going to be used to predict stopping distance.

- Find the linear regression equation.
- What does the slope tell you about MPH and the braking distance? How about the y intercept?
- Find the braking distance when MPH = 45.
- Find the braking distance when MPH = 100.
- Comment on predicting beyond the given data values.

See page 580 for the answers.

Exercises 10-4

- What two things should be done before one performs a regression analysis?
 - What are the assumptions for regression analysis?
 - What is the general form for the regression line used in statistics?
 - What is the symbol for the slope? For the y intercept?
 - What is meant by the *line of best fit*?
 - When all the points fall on the regression line, what is the value of the correlation coefficient?
 - What is the relationship between the sign of the correlation coefficient and the sign of the slope of the regression line?
 - As the value of the correlation coefficient increases from 0 to 1, or decreases from 0 to -1 , how do the points of the scatter plot fit the regression line?
 - How is the value of the correlation coefficient related to the accuracy of the predicted value for a specific value of x ?
 - If the value of r is not significant, what can be said about the regression line?
 - When the value of r is not significant, what value should be used to predict y ?
- For Exercises 12 through 27, use the same data as for the corresponding exercises in Section 10-3. For each exercise, find the equation of the regression line and find the y' value for the specified x value. Remember that no regression should be done when r is not significant.**
- Price of drugs for humans and animals

Humans x	0.67	0.64	1.20	0.51	0.87	0.74	0.50	1.22
Animals y	0.13	0.18	0.42	0.25	0.57	0.57	0.49	1.28

 Find y' when $x = 0.75$.
 - Ages and exercise

Age x	18	26	32	38	52	59
Hours y	10	5	2	3	1.5	1

 Find y' when $x = 35$ years.
 - Number of fires and number of acres burned

Fires x	72	69	58	47	84	62	57	45
Acres y	62	41	19	26	51	15	30	15

 Find y' when $x = 60$.

15. Years and contribution

Years x	1	5	3	10	7	6
Contribution y, \$	500	100	300	50	75	80

Find y' when $x = 4$ years.

16. Age and sick days

Age x	18	26	39	48	53	58
Days y	16	12	9	5	6	2

Find y' when $x = 47$ years.

17. Larceny crimes and vandalism crimes

Larceny x	24	6	16	64	10	25	35
Vandalism y	21	3	6	15	21	61	20

Find y' when $x = 40$.

18. Pass attempts and yards gained

Attempts x	116	90	82	108	92
Yards y	1001	823	851	873	837

Find y' when $x = 95$.

19. Tornadoes and deaths

Tornadoes x	1113	1132	1297	1173	1082
Deaths y	53	39	39	33	69
Tornadoes x	1234	1148	1424	1342	898
Deaths y	30	67	130	94	40

Find y' when $x = 1000$.

20. Temperature in degrees Fahrenheit and number of emergency calls

Temperature x	68	74	82	88	93	99	101
No. of calls y	7	4	8	10	11	9	13

Find y' when $x = 80^\circ\text{F}$.

21. Number (in thousands) of people under 5 years old and people 65 and over living in six randomly selected cities in the United States

Under 5 x	178	27	878	314	322	143
65 and older y	361	72	1496	501	585	207

Find y' when $x = 200$ thousand.

22. Rents for one-bedroom and two-bedroom apartments

One-bedroom x, \$	782	486	451	529	618	520	845
Two-bedroom y, \$	1223	902	739	954	1055	875	1455

Find y' when $x = \$700$.

23. Temperatures (in degrees Fahrenheit) and precipitation (in inches)

Avg. daily temp. x	86	81	83	89	80	74	64
Avg. mo. precip. y	3.4	1.8	3.5	3.6	3.7	1.5	0.2

Find y' when $x = 70^\circ\text{F}$.

24. Wins and strikeouts for Hall of Fame pitchers

Wins x	329	150	236	300	284	207
Strikeouts y	4136	1155	1956	2266	3192	1277

Wins x	247	314	273	324
Strikeouts y	1068	3534	1987	3574

Find y' when $x = 260$ wins.

25. Calories and cholesterol

Calories x	390	535	720	300	430	500	440
Cholesterol y	43	45	80	50	55	52	60

Find y' when $x = 600$ calories.

26. Stories and heights of buildings

Stories x	64	54	40	31	45	38	42	41	37	40
Heights y	841	725	635	616	615	582	535	520	511	485

Find y' when $x = 44$.


27. Licensed beds and staffed beds

Licensed beds x	144	32	175	185	208	100	169
Staffed beds y	112	32	162	141	103	80	118

Find y' when $x = 44$.


For Exercises 28 through 33, do a complete regression analysis by performing these steps.

- Draw a scatter plot.
- Compute the correlation coefficient.
- State the hypotheses.
- Test the hypotheses at $\alpha = 0.05$. Use Table I.
- Determine the regression line equation.
- Plot the regression line on the scatter plot.
- Summarize the results.


 28. These data were obtained for the years 1993 through 1998 and indicate the number of fireworks (in millions) used and the related injuries. Predict the number of injuries if 100 million fireworks are used during a given year.

Fireworks in use x	67.6	87.1	117	115	118	113
Related injuries y	12,100	12,600	12,500	10,900	7800	7000

Source: National Council of Fireworks Safety, American Pyrotechnic Assoc.


 29. These data were obtained from a survey of the number of years people smoked and the percentage of lung damage they sustained. Predict the percentage of lung damage for a person who has smoked for 30 years.

Years x	22	14	31	36	9	41	19
Damage y	20	14	54	63	17	71	23

 30. A medical researcher wishes to describe the relationship between the prescription cost of a brand

name drug and its generic equivalent. The data (in dollars) are shown. Describe the relationship.

Brand name x	96	93	59	80	44	47	15	56
Generic y	42	31	17	16	8	12	6	22

-  **31.** These data were obtained from a sample of counties in southwestern Pennsylvania and indicate the number (in thousands) of tons of bituminous coal produced in each county and the number of employees working in coal production in each county. Predict the number of employees needed to produce 500 thousand tons of coal. The data are given here.

Tons x	227	5410	5328	147	729
No. of employees y	110	731	1031	20	118
Tons x	8095	635	6157		
No. of employees y	1162	103	752		

- 32.** A television executive selects 10 television shows and compares the average number of viewers the show had last year with the average number of viewers this year. The data (in millions) are shown. Describe the relationship.


Viewers last year x	26.6	17.85	20.3	16.8	20.8
Viewers this year y	28.9	19.2	26.4	13.7	20.2
Viewers last year x	16.7	19.1	18.9	16.0	15.8
Viewers this year y	18.8	25.0	21.0	16.8	15.3

Source: Nielson Media Research.

- 33.** An educator wants to see how the number of absences for a student in her class affects the student's final grade. The data obtained from a sample are shown.

No. of absences x	10	12	2	0	8	5
Final grade y	70	65	96	94	75	82

For Exercises 34 and 35, do a complete regression analysis and test the significance of r at $\alpha = 0.05$, using the P -value method.

-  **34.** A physician wishes to know whether there is a relationship between a father's weight (in pounds) and his newborn son's weight (in pounds). The data are given here.

Father's weight x	176	160	187	210	196	142	205	215
Son's weight y	6.6	8.2	9.2	7.1	8.8	9.3	7.4	8.6

- 35.** Is a person's age related to his or her net worth? A sample of 10 billionaires is selected, and the person's age and net worth are compared. The data are given here.

Age x	56	39	42	60	84	37	68	66	73	55
Net worth (in billions) y	18	14	12	14	11	10	10	7	7	5

Source: The Associated Press.

Extending the Concepts

- 36.** For Exercises 13, 15, and 21 in Section 10-3, find the mean of the x and y variables. Then substitute the mean of the x variable into the corresponding regression line equations found in Exercises 12, 13, and 14 in this section and find y' . Compare the value of y' with \bar{y} for each exercise. Generalize the results.
- 37.** The y intercept value a can also be found by using the equation

$$a = \bar{y} - b\bar{x}$$

Verify this result by using the data in Exercises 15 and 16 of Sections 10-3 and 10-4.

- 38.** The value of the correlation coefficient can also be found by using the formula

$$r = \frac{bs_x}{s_y}$$

where s_x is the standard deviation of the x value and s_y is the standard deviation of the y values. Verify this result for Exercises 18 and 20 of Section 10-3.

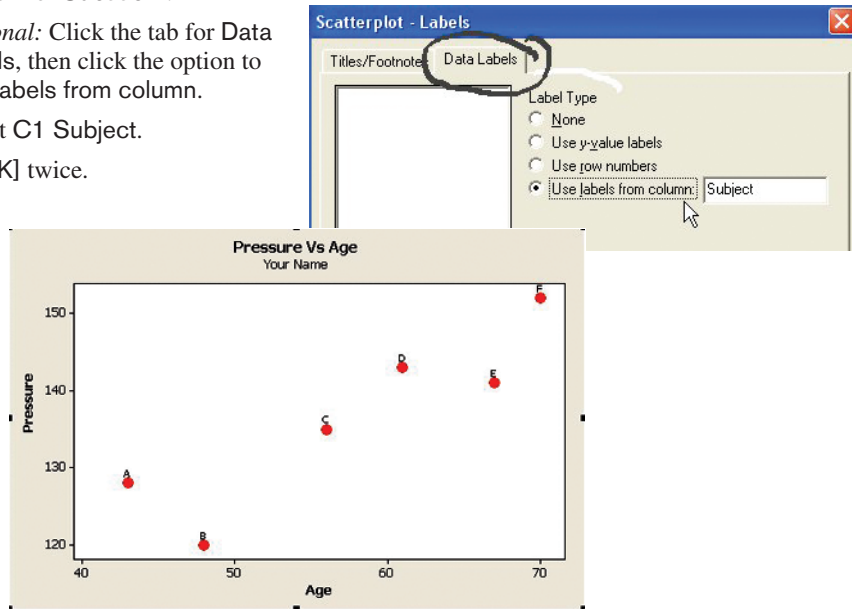
Technology Step by Step

MINITAB Step by Step

Create a Scatter Plot

- These instructions use Example 10-1. Enter the data into three columns. The subject column is optional (see step 6b).
- Name the columns **C1 Subject**, **C2 Age**, and **C3 Pressure**.
- Select **Graph>Scatterplot**, then select Simple and click [OK].
- Double-click on **C3 Pressure** for the [Y] variable and **C2 Age** for the predictor [X] variable.

5. Click [Data View]. The Data Display should be Symbols. If not, click the option box to select it. Click [OK].
6. Click [Labels].
 - a) Type **Pressure vs. Age** in the text box for Titles/Footnotes, then type **Your Name** in the box for Subtitle 1.
 - b) *Optional:* Click the tab for Data Labels, then click the option to Use labels from column.
 - c) Select C1 Subject.
7. Click [OK] twice.



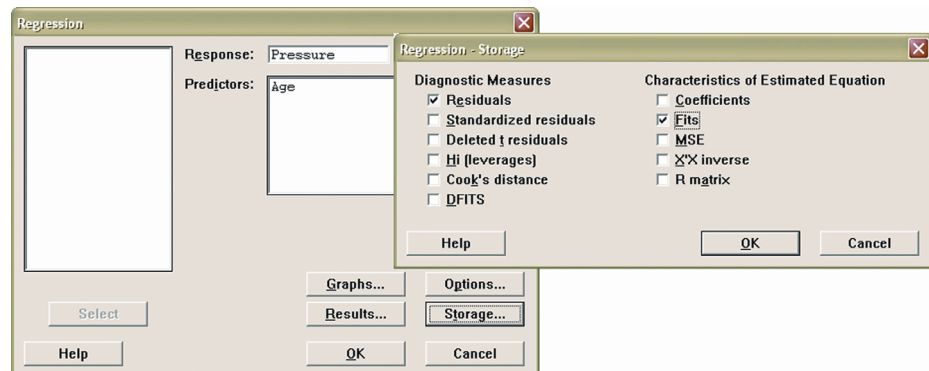
Calculate the Correlation Coefficient

8. Select **Stat>Basic Statistics>Correlation.**
9. Double-click C3 Pressure, then double-click C2 Age. The box for Display p-values should be checked.
10. Click [OK]. The correlation coefficient will be displayed in the session window, $r = +0.897$ with a P -value of 0.015.

Determine the Equation of the Least-Squares Regression Line

11. Select **Stat>Regression>Regression.**
12. Double-click Pressure in the variable list to select it for the Response variable Y.
13. Double-click C2 Age in the variable list to select it for the Predictors variable X.
14. Click on [Storage], then check the boxes for Residuals and Fits.
15. Click [OK] twice.

The session window will contain the regression analysis as shown.



In the worksheet two new columns will be added with the fitted values and residuals. Summary: The scatter plot and correlation coefficient confirm a strong positive linear correlation between pressure and age. The null hypothesis would be rejected at a significance level of 0.015. The equation of the regression equation is $\text{pressure} = 81.0 + 0.964(\text{age})$.

↓	C1-T	C2	C3	C4	C5
	Subject	Age	Pressure	RES1	FITS1
1	A	43	128	5.48353	122.516
2	B	48	120	-7.33838	127.338
3	C	56	135	-0.05343	135.053
4	D	61	143	3.12467	139.875
5	E	67	141	-4.66162	145.662
6	F	70	152	3.44524	148.555

Regression Analysis: Pressure versus Age

The regression equation is
 $\text{Pressure} = 81.0 + 0.964 \text{ Age}$

Predictor	Coef	SE Coef	T	P
Constant	81.05	13.88	5.84	0.004
Age	0.9644	0.2381	4.05	0.015
S = 5.641		R-Sq = 80.4%		R-Sq (adj) = 75.5%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	522.21	522.21	16.41	0.015
Residual Error	4	127.29	31.82		
Total	5	649.50			

TI-83 Plus or TI-84 Plus Step by Step

Correlation and Regression

To graph a scatter plot:

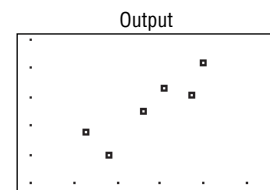
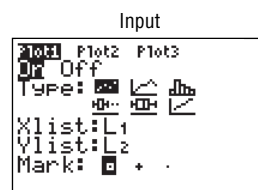
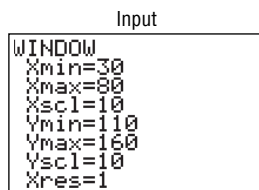
1. Enter the x values in L_1 and the y values in L_2 .
2. Make sure the Window values are appropriate. Select an X_{\min} slightly less than the smallest x data value and an X_{\max} slightly larger than the largest x data value. Do the same for Y_{\min} and Y_{\max} . Also, you may need to change the X_{scl} and Y_{scl} values, depending on the data.
3. Press **2nd** [STAT PLOT] **1** for Plot 1. The other y functions should be turned off.
4. Move the cursor to **On** and press **ENTER** on the Plot 1 menu.
5. Move the cursor to the graphic that looks like a scatter plot next to **Type** (first graph), and press **ENTER**. Make sure the X list is L_1 , and the Y list is L_2 .
6. Press **GRAPH**.

Example TI10-1

Draw a scatter plot for the data from Example 10-1.

x	43	48	56	61	67	70
y	128	120	135	143	141	152

The input and output screens are shown.



To find the equation of the regression line:

1. Press **STAT** and move the cursor to **Calc**.
2. Press **8** for **LinReg(a+bx)** then **ENTER**. The values for a and b will be displayed.

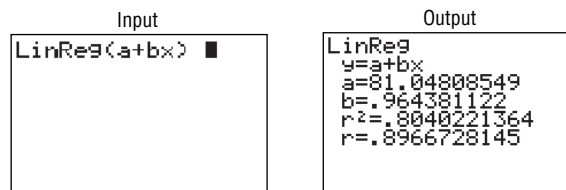
In order to have the calculator compute and display the correlation coefficient and coefficient of determination as well as the equation of the line, you must set the diagnostics display mode to on. Follow these steps:

1. Press **2nd** [CATALOG].
2. Use the arrow keys to scroll down to DiagnosticOn.
3. Press **ENTER** to copy the command to the home screen.
4. Press **ENTER** to execute the command.

You will have to do this only once. Diagnostic display mode will remain on until you perform a similar set of steps to turn it off.

Example TI10-2

Find the equation of the regression line for the data in Example TI10-1, as shown in Example 10-9. The input and output screens are shown.



The equation of the regression line is $y' = 81.04808549 + 0.964381122x$.

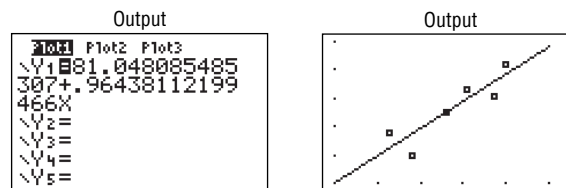
To plot the regression line on the scatter plot:

1. Press **Y=** and **CLEAR** to clear any previous equations.
2. Press **VARS** and then **5** for Statistics.
3. Move the cursor to EQ and press **1** for RegEQ. The line will be in the Y= screen.
4. Press **GRAPH**.

Example TI10-3

Draw the regression line found in Example TI10-2 on the scatter plot.

The output screens are shown.



To test the significance of b and ρ :

1. Press **STAT** and move the cursor to TESTS.
2. Press **E** (ALPHA SIN) for LinRegTTest. Make sure the Xlist is L_1 , the Ylist is L_2 , and the Freq is 1.
3. Select the appropriate alternative hypothesis.
4. Move the cursor to Calculate and press **ENTER**.

Example TI10-4

Test the hypothesis from Examples 10-4 and 10-7, $H_0: \rho = 0$ for the data in Example 10-1. Use $\alpha = 0.05$.

Input	Output	Output
<pre>LinRegTTest Xlist:L1 Ylist:L2 Freq:1 B & P: \neq <0 >0 RegEQ: Calculate</pre>	<pre>LinRegTTest y=a+bx B\neq0 and P\neq0 t=4.050983638 P=.0154631742 df=4 a=81.04808549</pre>	<pre>LinRegTTest y=a+bx B\neq0 and P\neq0 b=.964381122 s=5.641090817 r²=.8040221364 r=.8966728145</pre>

In this case, the t test value is 4.050983638. The P -value is 0.0154631742, which is significant. The decision is to reject the null hypothesis at $\alpha = 0.05$, since $0.0154631742 < 0.05$; $r = 0.8966728145$, $r^2 = 0.8040221364$.

There are two other ways to store the equation for the regression line in Y_1 for graphing.

1. Type Y_1 after the LinReg(a+bx) command.
2. Type Y_1 in the RegEQ: spot in the LinRegTTest.

To get Y_1 do this:

Press **VARS** for variables, move cursor to Y-VARS, press **1** for Function, press **1** for Y_1 .

Excel Step by Step

Scatter Plots

Creating scatter plots in Excel is straightforward when one uses the Chart Wizard.

1. Click on the Chart Wizard icon (it looks like a colorful histogram).
2. Select chart type XY (Scatter) under the Standard Types tab. Click on [Next >].
3. Enter the data range, and specify whether the data for each variable are stored in columns (as we have done in our examples) or rows. Click on [Next >].
4. The next dialog box enables you to set various options for displaying the plot. In most cases, the defaults will be okay. After entering the desired options (note that there are several tabs for this screen), click on [Next >].
5. Use this final dialog box to specify where the chart will be located. Click on [Finish].

Correlation Coefficient

The CORREL function returns the correlation coefficient.

1. Enter the data in columns A and B.
2. Select a blank cell, then click on the f_x button.
3. Under Function category, select Statistical. From the Function name list, select CORREL.
4. Enter the data range (**A1:AN**, where N is the number of sample data pairs) for the first variable in Array1. Enter the data range for the second variable in Array2. The correlation coefficient will be displayed in the selected cell.

Correlation and Regression

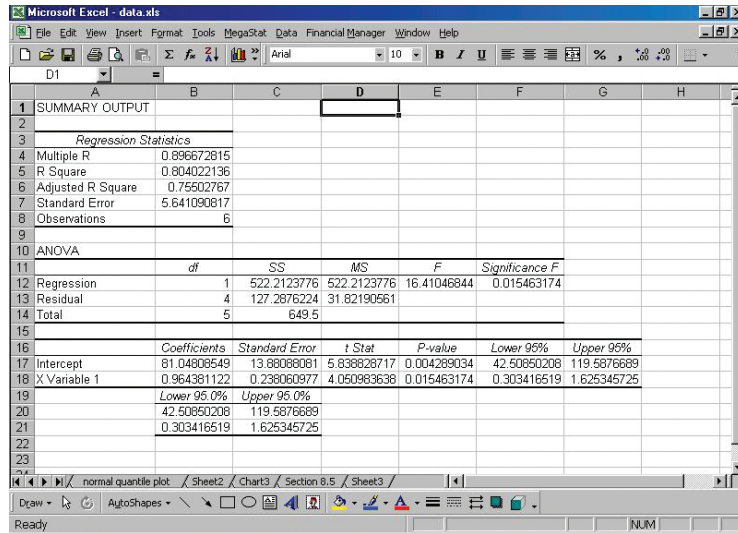
This procedure will allow you to calculate the Pearson product moment correlation coefficient without performing a regression analysis.

1. Enter the data from Example 10-2 in a new worksheet. Enter the seven values for the numbers of absences in column A and the corresponding final grades in column B.
2. Select **Tools>Data Analysis>Correlation**. Click the icon next to the Input Range box. This will minimize the dialog box. Use the mouse to highlight the data from columns A and B. Once the data are selected, click the icon to restore the box.

3. Make sure that the data are grouped by columns, and select New Worksheet Ply, then click [OK].

This procedure will allow you to conduct a regression analysis and compute the correlation coefficient.

1. Enter the seven values for the number of absences in column A and the corresponding final grades in column B.
2. Select **Tools>Data Analysis>Regression**.
3. Enter **B1:B7** for the Input Y Range, and then enter **A1:A7** for the Input X Range.
4. Click [OK].



10-5

Coefficient of Determination and Standard Error of the Estimate

The previous sections stated that if the correlation coefficient is significant, the equation of the regression line can be determined. Also, for various values of the independent variable x , the corresponding values of the dependent variable y can be predicted. Several other measures are associated with the correlation and regression techniques. They include the coefficient of determination, the standard error of estimate, and the prediction interval. But before these concepts can be explained, the different types of variation associated with the regression model must be defined.

Types of Variation for the Regression Model



Consider the following hypothetical regression model.

x	1	2	3	4	5
y	10	8	12	16	20

The equation of the regression line is $y' = 4.8 + 2.8x$, and $r = 0.919$. The sample y values are 10, 8, 12, 16, and 20. The predicted values, designated by y' , for each x can be found by substituting each x value into the regression equation and finding y' . For example, when $x = 1$,

$$y' = 4.8 + 2.8x = 4.8 + (2.8)(1) = 7.6$$

Now, for each x , there is an observed y value and a predicted y' value, for example, when $x = 1$, $y = 10$, and $y' = 7.6$. Recall that the closer the observed values are to the predicted values, the better the fit is and the closer r is to $+1$ or -1 .

The *total variation* $\Sigma(y - \bar{y})^2$ is the sum of the squares of the vertical distances each point is from the mean. The total variation can be divided into two parts: that which is attributed to the relationship of x and y and that which is due to chance. The variation obtained from the relationship (i.e., from the predicted y' values) is $\Sigma(y' - \bar{y})^2$ and is called the *explained variation*. Most of the variations can be explained by the relationship. The closer the value r is to $+1$ or -1 , the better the points fit the line and the closer $\Sigma(y' - \bar{y})^2$ is to $\Sigma(y - \bar{y})^2$. In fact, if all points fall on the regression line, $\Sigma(y' - \bar{y})^2$ will equal $\Sigma(y - \bar{y})^2$, since y' would be equal to y in each case.

On the other hand, the variation due to chance, found by $\Sigma(y - y')^2$, is called the *unexplained variation*. This variation cannot be attributed to the relationship. When the unexplained variation is small, the value of r is close to $+1$ or -1 . If all points fall on the regression line, the unexplained variation $\Sigma(y - y')^2$ will be 0. Hence, the *total variation* is equal to the sum of the explained variation and the unexplained variation. That is,

$$\Sigma(y - \bar{y})^2 = \Sigma(y' - \bar{y})^2 + \Sigma(y - y')^2$$

These values are shown in Figure 10-17. For a single point, the differences are called *deviations*. For the hypothetical regression model given earlier, for $x = 1$ and $y = 10$, one gets $y' = 7.6$ and $\bar{y} = 13.2$.

The procedure for finding the three types of variation is illustrated next.

Step 1 Find the predicted y' values.

For $x = 1$ $y' = 4.8 + 2.8x = 4.8 + (2.8)(1) = 7.6$

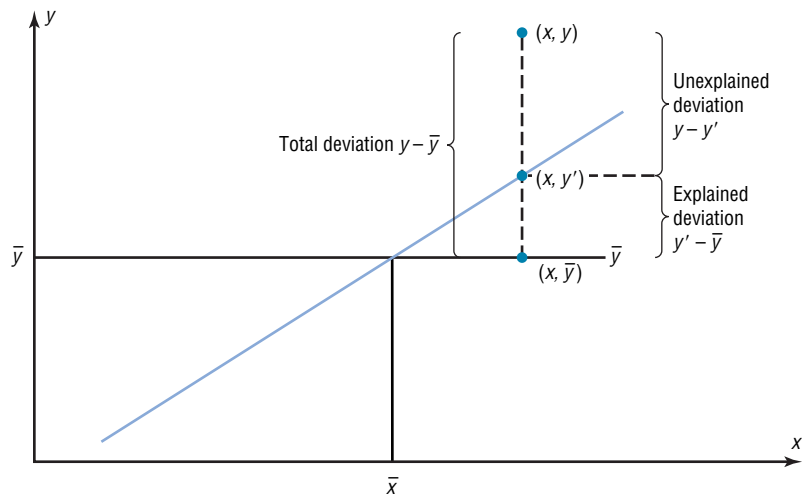
For $x = 2$ $y' = 4.8 + (2.8)(2) = 10.4$

For $x = 3$ $y' = 4.8 + (2.8)(3) = 13.2$

For $x = 4$ $y' = 4.8 + (2.8)(4) = 16.0$

For $x = 5$ $y' = 4.8 + (2.8)(5) = 18.8$

Figure 10-17
Deviations for the Regression Equation



Unusual Stat

There are 1,929,770, 126,028,800 different color combinations for Rubik's cube and only one correct solution in which all the colors of the squares on each face are the same.

Hence, the values for this example are as follows:

x	y	y'
1	10	7.6
2	8	10.4
3	12	13.2
4	16	16.0
5	20	18.8

Step 2 Find the mean of the y values.

$$\bar{y} = \frac{10 + 8 + 12 + 16 + 20}{5} = 13.2$$

Step 3 Find the total variation $\Sigma(y - \bar{y})^2$.

$$(10 - 13.2)^2 = 10.24$$

$$(8 - 13.2)^2 = 27.04$$

$$(12 - 13.2)^2 = 1.44$$

$$(16 - 13.2)^2 = 7.84$$

$$(20 - 13.2)^2 = 46.24$$

$$\Sigma(y - \bar{y})^2 = 92.8$$

Step 4 Find the explained variation $\Sigma(y' - \bar{y})^2$.

$$(7.6 - 13.2)^2 = 31.36$$

$$(10.4 - 13.2)^2 = 7.84$$

$$(13.2 - 13.2)^2 = 0.00$$

$$(16 - 13.2)^2 = 7.84$$

$$(18.8 - 13.2)^2 = 31.36$$

$$\Sigma(y' - \bar{y})^2 = 78.4$$

Step 5 Find the unexplained variation $\Sigma(y - y')^2$.

$$(10 - 7.6)^2 = 5.76$$

$$(8 - 10.4)^2 = 5.76$$

$$(12 - 13.2)^2 = 1.44$$

$$(16 - 16)^2 = 0.00$$

$$(20 - 18.8)^2 = 1.44$$

$$\Sigma(y - y')^2 = 14.4$$

Notice that

Total variation = Explained variation + Unexplained variation

$$92.8 = 78.4 + 14.4$$

Note: The values $(y - y')$ are called *residuals*. A **residual** is the difference between the actual value of y and the predicted value y' for a given x value. The mean of the residuals is always zero. As stated previously, the regression line determined by the formulas in Section 10-4 is the line that best fits the points of the scatter plot. The sum of the squares of the residuals computed by using the regression line is the smallest possible value. For this reason, a regression line is also called a **least-squares line**.

Objective 5

Compute the coefficient of determination.

Historical Note

Karl Pearson recommended in 1897 that the French government close all its casinos and turn the gambling devices over to the academic community to use in the study of probability.

Coefficient of Determination

The *coefficient of determination* is the ratio of the explained variation to the total variation and is denoted by r^2 . That is,

$$r^2 = \frac{\text{explained variation}}{\text{total variation}}$$

For the example, $r^2 = 78.4/92.8 = 0.845$. The term r^2 is usually expressed as a percentage. So in this case, 84.5% of the total variation is explained by the regression line using the independent variable.

Another way to arrive at the value for r^2 is to square the correlation coefficient. In this case, $r = 0.919$ and $r^2 = 0.845$, which is the same value found by using the variation ratio.

The **coefficient of determination** is a measure of the variation of the dependent variable that is explained by the regression line and the independent variable. The symbol for the coefficient of determination is r^2 .

Of course, it is usually easier to find the coefficient of determination by squaring r and converting it to a percentage. Therefore, if $r = 0.90$, then $r^2 = 0.81$, which is equivalent to 81%. This result means that 81% of the variation in the dependent variable is accounted for by the variations in the independent variable. The rest of the variation, 0.19, or 19%, is unexplained. This value is called the *coefficient of nondetermination* and is found by subtracting the coefficient of determination from 1. As the value of r approaches 0, r^2 decreases more rapidly. For example, if $r = 0.6$, then $r^2 = 0.36$, which means that only 36% of the variation in the dependent variable can be attributed to the variation in the independent variable.

Coefficient of Nondetermination

$$1.00 - r^2$$

Objective 6

Compute the standard error of the estimate.

Standard Error of the Estimate

When a y' value is predicted for a specific x value, the prediction is a point prediction. However, a prediction interval about the y' value can be constructed, just as a confidence interval was constructed for an estimate of the population mean. The prediction interval uses a statistic called the *standard error of the estimate*.

The **standard error of the estimate**, denoted by s_{est} , is the standard deviation of the observed y values about the predicted y' values. The formula for the standard error of estimate is

$$s_{\text{est}} = \sqrt{\frac{\sum(y - y')^2}{n - 2}}$$

The standard error of the estimate is similar to the standard deviation, but the mean is not used. As can be seen from the formula, the standard error of the estimate is the square root of the unexplained variation—that is, the variation due to the difference of the observed values and the expected values—divided by $n - 2$. So the closer the observed values are to the predicted values, the smaller the standard error of the estimate will be.

Example 10-12 shows how to compute the standard error of the estimate.

Example 10–12

A researcher collects the following data and determines that there is a significant relationship between the age of a copy machine and its monthly maintenance cost. The regression equation is $y' = 55.57 + 8.13x$. Find the standard error of the estimate.

Machine	Age x (years)	Monthly cost y
A	1	\$ 62
B	2	78
C	3	70
D	4	90
E	4	93
F	6	103

Solution

Step 1 Make a table, as shown.

x	y	y'	$y - y'$	$(y - y')^2$
1	62			
2	78			
3	70			
4	90			
4	93			
6	103			

Step 2 Using the regression line equation $y' = 55.57 + 8.13x$, compute the predicted values y' for each x and place the results in the column labeled y' .

$$x = 1 \quad y' = 55.57 + (8.13)(1) = 63.70$$

$$x = 2 \quad y' = 55.57 + (8.13)(2) = 71.83$$

$$x = 3 \quad y' = 55.57 + (8.13)(3) = 79.96$$

$$x = 4 \quad y' = 55.57 + (8.13)(4) = 88.09$$

$$x = 6 \quad y' = 55.57 + (8.13)(6) = 104.35$$

Step 3 For each y , subtract y' and place the answer in the column labeled $y - y'$.

$$62 - 63.70 = -1.70 \quad 90 - 88.09 = 1.91$$

$$78 - 71.83 = 6.17 \quad 93 - 88.09 = 4.91$$

$$70 - 79.96 = -9.96 \quad 103 - 104.35 = -1.35$$

Step 4 Square the numbers found in step 3 and place the squares in the column labeled $(y - y')^2$.

Step 5 Find the sum of the numbers in the last column. The completed table is shown.

x	y	y'	$y - y'$	$(y - y')^2$
1	62	63.70	-1.70	2.89
2	78	71.83	6.17	38.0689
3	70	79.96	-9.96	99.2016
4	90	88.09	1.91	3.6481
4	93	88.09	4.91	24.1081
6	103	104.35	-1.35	1.8225
				169.7392

Step 6 Substitute in the formula and find s_{est} .

$$s_{\text{est}} = \sqrt{\frac{\sum(y - y')^2}{n - 2}} = \sqrt{\frac{169.7392}{6 - 2}} = 6.51$$

In this case, the standard deviation of observed values about the predicted values is 6.51.

The standard error of the estimate can also be found by using the formula

$$s_{\text{est}} = \sqrt{\frac{\sum y^2 - a \sum y - b \sum xy}{n - 2}}$$

Example 10-13

Find the standard error of the estimate for the data for Example 10-12 by using the preceding formula. The equation of the regression line is $y' = 55.57 + 8.13x$.

Solution

Step 1 Make a table.

Step 2 Find the product of x and y values, and place the results in the third column.

Step 3 Square the y values, and place the results in the fourth column.

Step 4 Find the sums of the second, third, and fourth columns. The completed table is shown here.

x	y	xy	y^2
1	62	62	3,844
2	78	156	6,084
3	70	210	4,900
4	90	360	8,100
4	93	372	8,649
6	103	618	10,609
	$\Sigma y = 496$	$\Sigma xy = 1,778$	$\Sigma y^2 = 42,186$

Step 5 From the regression equation $y' = 55.57 + 8.13x$, $a = 55.57$ and $b = 8.13$.

Step 6 Substitute in the formula and solve for s_{est} .

$$\begin{aligned} s_{\text{est}} &= \sqrt{\frac{\sum y^2 - a \sum y - b \sum xy}{n - 2}} \\ &= \sqrt{\frac{42,186 - (55.57)(496) - (8.13)(1,778)}{6 - 2}} = 6.48 \end{aligned}$$

This value is close to the value found in Example 10-12. The difference is due to rounding.

Objective 7

Find a prediction interval.

Prediction Interval

The standard error of estimate can be used for constructing a **prediction interval** (similar to a confidence interval) about a y' value.

When a specific value x is substituted into the regression equation, one gets y' , which is a point estimate for y . For example, if the regression line equation for the age of a machine and the monthly maintenance cost is $y' = 55.57 + 8.13x$ (Example 10-12), then

the predicted maintenance cost for a 3-year-old machine would be $y' = 55.57 + 8.13(3)$, or \$79.96. Since this is a point estimate, one has no idea how accurate it is. But one can construct a prediction interval about the estimate. By selecting an α value, one can achieve a $(1 - \alpha) \cdot 100\%$ confidence that the interval contains the actual mean of the y values that correspond to the given value of x .

The reason is that there are possible sources of prediction errors in finding the regression line equation. One source occurs when finding the standard error of the estimate s_{est} . Two others are errors made in estimating the slope and the y' intercept, since the equation of the regression line will change somewhat if different random samples are used when calculating the equation.

Formula for the Prediction Interval about a Value y'

$$y' - t_{\alpha/2}s_{\text{est}}\sqrt{1 + \frac{1}{n} + \frac{n(x - \bar{X})^2}{n \sum x^2 - (\sum x)^2}} < y < y' + t_{\alpha/2}s_{\text{est}}\sqrt{1 + \frac{1}{n} + \frac{n(x - \bar{X})^2}{n \sum x^2 - (\sum x)^2}}$$

with d.f. = $n - 2$.

Example 10-14

For the data in Example 10-12, find the 95% prediction interval for the monthly maintenance cost of a machine that is 3 years old.

Solution

Step 1 Find $\sum x$, $\sum x^2$, and \bar{X} .

$$\sum x = 20 \quad \sum x^2 = 82 \quad \bar{X} = \frac{20}{6} = 3.3$$

Step 2 Find y' for $x = 3$.

$$y' = 55.57 + 8.13x$$

$$y' = 55.57 + 8.13(3) = 79.96$$

Step 3 Find s_{est} .

$$s_{\text{est}} = 6.48$$

as shown in Example 10-13.

Step 4 Substitute in the formula and solve: $t_{\alpha/2} = 2.776$, d.f. = $6 - 2 = 4$ for 95%.

$$y' - t_{\alpha/2}s_{\text{est}}\sqrt{1 + \frac{1}{n} + \frac{n(x - \bar{X})^2}{n \sum x^2 - (\sum x)^2}} < y < y'$$

$$+ t_{\alpha/2}s_{\text{est}}\sqrt{1 + \frac{1}{n} + \frac{n(x - \bar{X})^2}{n \sum x^2 - (\sum x)^2}}$$

$$79.96 - (2.776)(6.48)\sqrt{1 + \frac{1}{6} + \frac{6(3 - 3.3)^2}{6(82) - (20)^2}} < y < 79.96$$

$$+ (2.776)(6.48)\sqrt{1 + \frac{1}{6} + \frac{6(3 - 3.3)^2}{6(82) - (20)^2}}$$

$$79.96 - (2.776)(6.48)(1.08) < y < 79.96 + (2.776)(6.48)(1.08)$$

$$79.96 - 19.43 < y < 79.96 + 19.43$$

$$60.53 < y < 99.39$$

Hence, one can be 95% confident that the interval $60.53 < y < 99.39$ contains the actual value of y .

Applying the Concepts 10-5

Interpreting Simple Linear Regression

Answer the questions about the following computer-generated information.

Linear correlation coefficient $r = 0.794556$

Coefficient of determination = 0.631319

Standard error of estimate = 12.9668

Explained variation = 5182.41

Unexplained variation = 3026.49

Total variation = 8208.90

Equation of regression line $y' = 0.725983X + 16.5523$

Level of significance = 0.1

Test statistic = 0.794556

Critical value = 0.378419

- Are both variables moving in the same direction?
- Which number measures the distances from the prediction line to the actual values?
- Which number is the slope of the regression line?
- Which number is the y intercept of the regression line?
- Which number can be found in a table?
- Which number is the allowable risk of making a type I error?
- Which number measures the variation explained by the regression?
- Which number measures the scatter of points about the regression line?
- What is the null hypothesis?
- Which number is compared to the critical value to see if the null hypothesis should be rejected?
- Should the null hypothesis be rejected?

See page 581 for the answers.

Exercises 10-5

- What is meant by the *explained variation*? How is it computed?
- What is meant by the *unexplained variation*? How is it computed?
- What is meant by the *total variation*? How is it computed?
- Define the coefficient of determination.
- How is the coefficient of determination found?
- Define the coefficient of nondetermination.
- How is the coefficient of nondetermination found?
- $r = 0.81$
- $r = 0.70$
- $r = 0.45$
- $r = 0.37$
- $r = 0.15$
- $r = 0.05$
- Define the standard error of the estimate for regression. When can the standard error of the estimate be used to construct a prediction interval about a value y' ?
- Compute the standard error of the estimate for Exercise 13 in Section 10-3. The regression line equation was found in Exercise 13 in Section 10-4.
- Compute the standard error of the estimate for Exercise 14 in Section 10-3. The regression line equation was found in Exercise 14 in Section 10-4.
- Compute the standard error of the estimate for Exercise 15 in Section 10-3. The regression line equation was found in Exercise 15 in Section 10-4.

For Exercises 8 through 13, find the coefficients of determination and nondetermination and explain the meaning of each.

18. Compute the standard error of the estimate for Exercise 16 in Section 10–3. The regression line equation was found in Exercise 16 in Section 10–4.
19. For the data in Exercises 13 in Sections 10–3 and 10–4 and 15 in Section 10–5, find the 90% prediction interval when $x = 20$ years.
20. For the data in Exercises 14 in Sections 10–3 and 10–4 and 16 in Section 10–5, find the 95% prediction interval when $x = 60$.
21. For the data in Exercises 15 in Sections 10–3 and 10–4 and 17 in Section 10–5, find the 90% prediction interval when $x = 4$ years.
22. For the data in Exercises 16 in Sections 10–3 and 10–4 and 18 in Section 10–5, find the 98% prediction interval when $x = 47$ years.

10–6

Multiple Regression (Optional)

Objective 8

Be familiar with the concept of multiple regression.

The previous sections explained the concepts of simple linear regression and correlation. In simple linear regression, the regression equation contains one independent variable x and one dependent variable y' and is written as

$$y' = a + bx$$

where a is the y' intercept and b is the slope of the regression line.

In **multiple regression**, there are several independent variables and one dependent variable, and the equation is

$$y' = a + b_1x_1 + b_2x_2 + \cdots + b_kx_k$$

where x_1, x_2, \dots, x_k are the independent variables.

For example, suppose a nursing instructor wishes to see whether there is a relationship between a student's grade point average, age, and score on the state board nursing examination. The two independent variables are GPA (denoted by x_1) and age (denoted by x_2). The instructor will collect the data for all three variables for a sample of nursing students. Rather than conduct two separate simple regression studies, one using the GPA and state board scores and another using ages and state board scores, the instructor can conduct one study using multiple regression analysis with two independent variables—GPA and ages—and one dependent variable—state board scores.

Unusual Stats

The most popular single-digit number played by people who purchase lottery tickets is 7.



Speaking of Statistics

In this study, researchers found a correlation between the cleanliness of the homes children are raised in and the years of schooling completed and earning potential for those children. What interfering variables were controlled? How might these have been controlled? Summarize the conclusions of the study.

SUCCESS

HOME SMART HOME

KIDS WHO GROW UP IN A CLEAN HOUSE FARE BETTER AS ADULTS

Good-bye, GPA. So long, SATs. New research suggests that we may be able to predict children's future success from the level of cleanliness in their homes.

A University of Michigan study presented at the annual meeting of the American Economic Association uncovered a surprising correlation: children raised in clean homes were later found to have completed more school and to have higher earning potential than those raised in dirty homes. The clean homes may indicate a family that values organization and similarly helpful skills at school and work, researchers say.

Cleanliness ratings for about 5,000 households were assessed between 1968 and 1972, and respondents were interviewed 25 years later to determine educational achievement and professional earnings of the young adults who had grown up there, controlling

for variables such as race, socioeconomic status and level of parental education. The data showed that those raised in homes rated "clean" to "very clean" had completed an average of 1.6 more years of school than those raised in "not very clean" or "dirty" homes. Plus, the first group's annual wages averaged about \$3,100 more than the second's.

But don't buy stock in Mr. Clean and Pine Sol just yet. "We're not advocating that everyone go out and clean their homes right this minute," explains Rachel Dunifon, a University of Michigan doctoral candidate and a researcher on the study. Rather, the main implication of the study, Dunifon says, is that there is significant evidence that non-cognitive factors, such as organization and efficiency, play a role in determining academic and financial success.

— Jackie Fisherman

Source: Reprinted with permission from *Psychology Today*, Copyright © (2001) Sussex Publishers, Inc.

A multiple regression correlation R can also be computed to determine if a significant relationship exists between the independent variables and the dependent variable. Multiple regression analysis is used when a statistician thinks there are several independent variables contributing to the variation of the dependent variable. This analysis then can be used to increase the accuracy of predictions for the dependent variable over one independent variable alone.

Two other examples for multiple regression analysis are when a store manager wants to see whether the amount spent on advertising and the amount of floor space used for a display affect the amount of sales of a product, and when a sociologist wants to see whether the amount of time children spend watching television and playing video games is related to their weight. Multiple regression analysis can also be conducted by using more than two independent variables, denoted by $x_1, x_2, x_3, \dots, x_m$. Since these computations are quite complicated and for the most part would be done on a computer, this chapter will show the computations for two independent variables only.



For example, the nursing instructor wishes to see whether a student's grade point average and age are related to the student's score on the state board nursing examination. She selects five students and obtains the following data.

Student	GPA x_1	Age x_2	State board score y
A	3.2	22	550
B	2.7	27	570
C	2.5	24	525
D	3.4	28	670
E	2.2	23	490

The multiple regression equation obtained from the data is

$$y' = -44.81 + 87.64x_1 + 14.533x_2$$

If a student has a GPA of 3.0 and is 25 years old, her predicted state board score can be computed by substituting these values in the equation for x_1 and x_2 , respectively, as shown.

$$\begin{aligned} y' &= -44.81 + 87.64(3.0) + 14.533(25) \\ &= 581.44 \text{ or } 581 \end{aligned}$$

Hence, if a student has a GPA of 3.0 and is 25 years old, the student's predicted state board score is 581.

The Multiple Regression Equation

A multiple regression equation with two independent variables (x_1 and x_2) and one dependent variable would have the form

$$y' = a + b_1x_1 + b_2x_2$$

A multiple regression with three independent variables (x_1 , x_2 , and x_3) and one dependent variable would have the form

$$y' = a + b_1x_1 + b_2x_2 + b_3x_3$$

General Form of the Multiple Regression Equation

The general form of the multiple regression equation with k independent variables is

$$y' = a + b_1x_1 + b_2x_2 + \cdots + b_kx_k$$

The x 's are the independent variables. The value for a is more or less an intercept, although a multiple regression equation with two independent variables constitutes a plane rather than a line. The b 's are called *partial regression coefficients*. Each b represents the amount of change in y' for one unit of change in the corresponding x value when the other x values are held constant. In the example just shown, the regression equation was $y' = -44.81 + 87.64x_1 + 14.533x_2$. In this case, for each unit of change in the student's GPA, there is a change of 87.64 units in the state board score with the student's age x_2 being held constant. And for each unit of change in x_2 (the student's age), there is a change of 14.533 units in the state board score with the GPA held constant.

Assumptions for Multiple Regression

The assumptions for multiple regression are similar to those for simple regression.

1. For any specific value of the independent variable, the values of the y variable are normally distributed. (This is called the *normality* assumption.)
2. The variances (or standard deviations) for the y variables are the same for each value of the independent variable. (This is called the *equal-variance* assumption.)
3. There is a linear relationship between the dependent variable and the independent variables. (This is called the *linearity* assumption.)
4. The independent variables are not correlated. (This is called the *nonmulticollinearity* assumption.)
5. The values for the y variables are independent. (This is called the *independence* assumption.)

In multiple regression, as in simple regression, the strength of the relationship between the independent variables and the dependent variable is measured by a correlation coefficient. This **multiple correlation coefficient** is symbolized by R . The value of R can range from 0 to +1; R can never be negative. The closer to +1, the stronger the relationship; the closer to 0, the weaker the relationship. The value of R takes into account all the independent variables and can be computed by using the values of the individual correlation coefficients. The formula for the multiple correlation coefficient when there are two independent variables is shown next.

Formula for the Multiple Correlation Coefficient

The formula for R is

$$R = \sqrt{\frac{r_{yx_1}^2 + r_{yx_2}^2 - 2r_{yx_1} \cdot r_{yx_2} \cdot r_{x_1x_2}}{1 - r_{x_1x_2}^2}}$$

where r_{yx_1} is the value of the correlation coefficient for variables y and x_1 ; r_{yx_2} is the value of the correlation coefficient for variables y and x_2 ; and $r_{x_1x_2}$ is the value of the correlation coefficient for variables x_1 and x_2 .

In this case, R is 0.989, as shown in Example 10–15. The multiple correlation coefficient is always higher than the individual correlation coefficients. For this specific example, the multiple correlation coefficient is higher than the two individual correlation coefficients computed by using grade point average and state board scores ($r_{yx_1} = 0.845$) or age and state board scores ($r_{yx_2} = 0.791$). *Note:* $r_{x_1x_2} = 0.371$.

Example 10–15

For the data regarding state board scores, find the value of R .

Solution

The values of the correlation coefficients are

$$r_{yx_1} = 0.845$$

$$r_{yx_2} = 0.791$$

$$r_{x_1x_2} = 0.371$$

Substituting in the formula, one gets

$$\begin{aligned} R &= \sqrt{\frac{r_{yx_1}^2 + r_{yx_2}^2 - 2r_{yx_1} \cdot r_{yx_2} \cdot r_{x_1x_2}}{1 - r_{x_1x_2}^2}} \\ &= \sqrt{\frac{(0.845)^2 + (0.791)^2 - 2(0.845)(0.791)(0.371)}{1 - 0.371^2}} \\ &= \sqrt{\frac{0.8437569}{0.862359}} = \sqrt{0.9784288} = 0.989 \end{aligned}$$

Hence, the correlation between a student's grade point average and age with the student's score on the nursing state board examination is 0.989. In this case, there is a strong relationship among the variables; the value of R is close to 1.00.

As with simple regression, R^2 is the *coefficient of multiple determination*, and it is the amount of variation explained by the regression model. The expression $1 - R^2$ represents the amount of unexplained variation, called the *error or residual variation*. Since $R = 0.989$, $R^2 = 0.978$ and $1 - R^2 = 1 - 0.978 = 0.022$.

Testing the Significance of R

An F test is used to test the significance of R . The hypotheses are

$$H_0: \rho = 0 \quad \text{and} \quad H_1: \rho \neq 0$$

where ρ represents the population correlation coefficient for multiple correlation.

F Test for Significance of R

The formula for the F test is

$$F = \frac{R^2/k}{(1 - R^2)/(n - k - 1)}$$

where n is the number of data groups (x_1, x_2, \dots, y) and k is the number of independent variables.

The degrees of freedom are d.f.N. = $n - k$ and d.f.D. = $n - k - 1$.

Example 10–16

Test the significance of the R obtained in Example 10–15 at $\alpha = 0.05$.

Solution

$$\begin{aligned} F &= \frac{R^2/k}{(1 - R^2)/(n - k - 1)} \\ &= \frac{0.978/2}{(1 - 0.978)/(5 - 2 - 1)} = \frac{0.489}{0.011} = 44.45 \end{aligned}$$

The critical value obtained from Table H with $\alpha = 0.05$, d.f.N. = 3, and d.f.D. = $5 - 2 - 1 = 2$ is 19.16. Hence, the decision is to reject the null hypothesis and conclude that there is a significant relationship among the student's GPA, age, and score on the nursing state board examination.

Adjusted R^2

Since the value of R^2 is dependent on n (the number of data pairs) and k (the number of variables), statisticians also calculate what is called an **adjusted R^2** , denoted by R_{adj}^2 . This is based on the number of degrees of freedom.

Formula for the Adjusted R^2

The formula for the adjusted R^2 is

$$R_{\text{adj}}^2 = 1 - \left[\frac{(1 - R^2)(n - 1)}{n - k - 1} \right]$$

The adjusted R^2 is smaller than R^2 and takes into account the fact that when n and k are approximately equal, the value of R may be artificially high, due to sampling error rather than a true relationship among the variables. This occurs because the chance variations of all the variables are used in conjunction with each other to derive the regression equation. Even if the individual correlation coefficients for each independent variable and the dependent variable were all zero, the multiple correlation coefficient due to sampling error could be higher than zero.

Hence, both R^2 and R_{adj}^2 are usually reported in a multiple regression analysis.

Example 10–17

Calculate the adjusted R^2 for the data in Example 10–16. The value for R is 0.989.

Solution

$$\begin{aligned} R_{\text{adj}}^2 &= 1 - \left[\frac{(1 - R^2)(n - 1)}{n - k - 1} \right] \\ &= 1 - \left[\frac{(1 - 0.989^2)(5 - 1)}{5 - 2 - 1} \right] \\ &= 1 - 0.043758 \\ &= 0.956 \end{aligned}$$

In this case, when the number of data pairs and the number of independent variables are accounted for, the adjusted multiple coefficient of determination is 0.956.

Applying the Concepts 10–6

More Math Means More Money

In a study to determine a person's yearly income 10 years after high school, it was found that the two biggest predictors are number of math courses taken and number of hours worked per week during a person's senior year of high school. The multiple regression equation generated from a sample of 20 individuals is

$$y' = 6000 + 4540x_1 + 1290x_2$$

Let x_1 represent the number of mathematics courses taken and x_2 represent hours worked. The correlation between income and mathematics courses is 0.63. The correlation between income and hours worked is 0.84, and the correlation between mathematics courses and hours worked is 0.31. Use this information to answer the following questions.

1. What is the dependent variable?
2. What are the independent variables?

3. What are the multiple regression assumptions?
4. Explain what 4540 and 1290 in the equation tell us.
5. What is the predicted income if a person took 8 math classes and worked 20 hours per week during her or his senior year in high school?
6. What does a multiple correlation coefficient of 0.77 mean?
7. Compute R^2 .
8. Compute the adjusted R^2 .
9. Would the equation be considered a good predictor of income?
10. What are your conclusions about the relationship between courses taken, hours worked, and yearly income?

See page 581 for the answers.

Exercises 10–6

1. Explain the similarities and differences between simple linear regression and multiple regression.
2. What is the general form of the multiple regression equation? What does a represent? What do the b 's represent?
3. Why would a researcher prefer to conduct a multiple regression study rather than separate regression studies using one independent variable and the dependent variable?
4. What are the assumptions for multiple regression?
5. How do the values of the individual correlation coefficients compare to the value of the multiple correlation coefficient?
6. A researcher has determined that a significant relationship exists among an employee's age x_1 , grade point average x_2 , and income y . The multiple regression equation is $y' = -34,127 + 132x_1 + 20,805x_2$. Predict the income of a person who is 32 years old and has a GPA of 3.4.
7. A manufacturer found that a significant relationship exists among the number of hours an assembly line employee works per shift x_1 , the total number of items produced x_2 , and the number of defective items produced y . The multiple regression equation is $y' = 9.6 + 2.2x_1 - 1.08x_2$. Predict the number of defective items produced by an employee who has worked 9 hours and produced 24 items.
8. A real estate agent found that there is a significant relationship among the number of acres on a farm x_1 , the number of rooms in the farmhouse x_2 , and the selling price in thousands of dollars y of farms in a specific area. The regression equation is $y' = 44.9 - 0.0266x_1 + 7.56x_2$. Predict the selling price of a farm that has 371 acres and a farmhouse with six rooms.
9. An educator has found a significant relationship among a college graduate's IQ x_1 , score on the verbal section of the SAT x_2 , and income for the first year following graduation from college y . Predict the income of a college graduate whose IQ is 120 and verbal SAT score is 650. The regression equation is $y' = 5000 + 97x_1 + 35x_2$.
10. A medical researcher found a significant relationship among a person's age x_1 , cholesterol level x_2 , sodium level of the blood x_3 , and systolic blood pressure y . The regression equation is $y' = 97.7 + 0.691x_1 + 219x_2 - 299x_3$. Predict the systolic blood pressure of a person who is 35 years old and has a cholesterol level of 194 milligrams per deciliter (mg/dl) and a sodium blood level of 142 milliequivalents per liter (mEq/l).
11. Explain the meaning of the multiple correlation coefficient R .
12. What is the range of values R can assume?
13. Define R^2 and R^2_{adj} .
14. What are the hypotheses used to test the significance of R ?
15. What test is used to test the significance of R ?
16. What is the meaning of the adjusted R^2 ? Why is it computed?

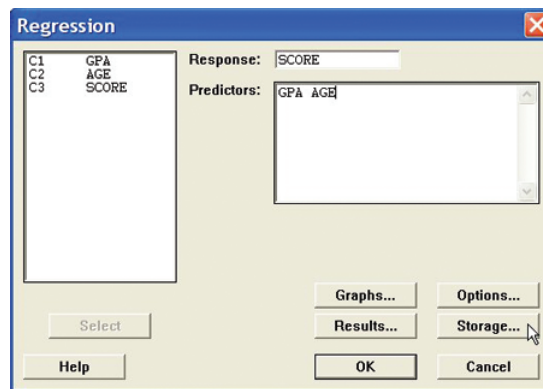
Technology Step by Step

MINITAB
Step by Step

Multiple Regression

In Example 10–15, is there a correlation between a student's score and her or his age and grade point average?

1. Enter the data for the example into three columns of MINITAB. Name the columns **GPA**, **AGE**, and **SCORE**.
2. Click **Stat>Regression>Regression**.
3. Double-click on C3 SCORE, the response variable.
4. Double-click C1 GPA, then C2 AGE.
5. Click on [Storage].
 - a) Check the box for Residuals.
 - b) Check the box for Fits.
6. Click [OK] twice.

**Regression Analysis: SCORE versus GPA, AGE**

The regression equation is

$$\text{SCORE} = -44.8 + 87.6 \text{ GPA} + 14.5 \text{ Age}$$

Predictor	Coef	SE Coef	T	P
Constant	-44.81	69.25	-0.65	0.584
GPA	87.64	15.24	5.75	0.029
AGE	14.533	2.914	4.99	0.038

S = 14.0091 R-Sq = 97.9% R-Sq(adj) = 95.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	18027.5	9013.7	45.93	0.021
Residual Error	2	392.5	196.3		
Total	4	18420.0			

The test statistic and *P*-value are 45.93 and 0.021, respectively. Since the *P*-value is less than α , reject the null hypothesis. There is enough evidence in the sample to conclude the scores are related to age and grade point average.

TI-83 Plus or
TI-84 Plus
Step by Step

The TI-83 Plus and the TI-84 Plus do not have a built-in function for multiple regression. However, the downloadable program named MULREG is available on your CD and Online Learning Center. Follow the instructions with your CD for downloading the program.

Finding a Multiple Regression Equation

1. Enter the sets of data values into L_1 , L_2 , L_3 , etc. Make note of which lists contain the independent variables and which list contains the dependent variable as well as how many data values are in each list.
2. Press **PRGM**, move the cursor to the program named MULREG, and press **ENTER** twice.
3. Type the number of independent variables and press **ENTER**.
4. Type the number of cases for each variable and press **ENTER**.
5. Type the name of the list that contains the data values for the first independent variable and press **ENTER**. Repeat this for all independent variables and the dependent variable.
6. The program will show the regression coefficients.

7. Press **ENTER** to see the values of R^2 and adjusted R^2 .
8. Press **ENTER** to see the values of the F test statistics and the P -value.

Find the multiple regression equation for these data used in this section:

Student	GPA x_1	Age x_2	State board score y
A	3.2	22	550
B	2.7	27	570
C	2.5	24	525
D	3.4	28	670
E	2.2	23	490

```
HOW MANY IND.
VARIABLES?
?2
HOW MANY CASES
FOR EACH VAR ?
?5
```

```
IN WHICH LIST
IS IND VAR NUM1
?L1
```

```
IN WHICH LIST
IS IND VAR NUM2
?L2
```

```
IN WHICH LIST IS
DEPENDENT VAR?
?L3
```

```
REG COEF IN ORD.:
A, B1, B2, ...
-44.81018805
87.64015185
14.53297431
ENTER FOR MORE
```

```
R^2= .9786911481
ADJ R^2= .9573822961
ENTER FOR MORE
```

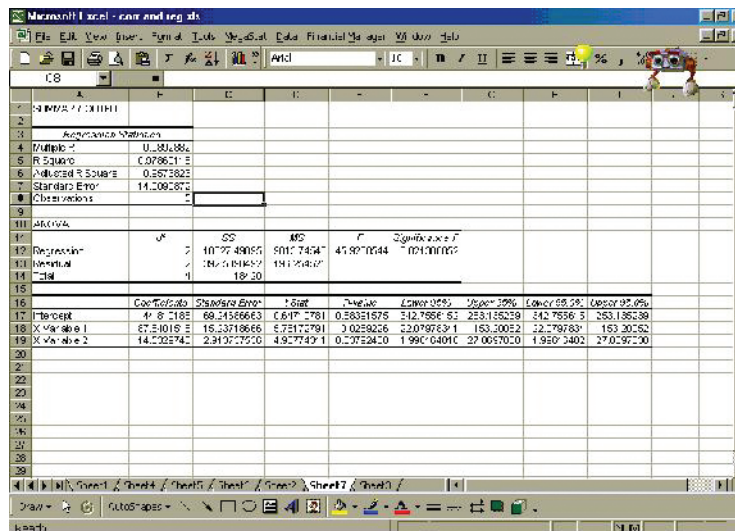
```
F STATISTIC=
45.92885392
P-VALUE =
.0213833419
Done
```

Excel Step by Step

Multiple Regression

These instructions use data from the nursing examination example discussed at the beginning of Section 10–6.

1. Enter the data from the example into three separate columns of a new worksheet—GPAs in cells A1:A5, ages in cells B1:B5, and scores in cells C1:C5.
2. Select **Tools>Data Analysis>Regression**.
3. Select cells C1:C5 for the Input Y Range.
4. Select cells A1:B5 for the Input X Range.
5. Click [OK].



Regression Analysis: SCORE versus GPA, AGE

The regression equation is

$$\text{SCORE} = -44.8 + 87.6 \text{ GPA} + 14.5 \text{ AGE}$$

Predictor	Coef	SE Coef	T	P
Constant	-44.81	69.25	-0.65	0.584
GPA	87.64	15.24	5.75	0.029
AGE	14.533	2.914	4.99	0.038

S = 14.01 R-Sq = 97.9% R-Sq(adj) = 95.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	18027.5	9013.7	45.93	0.021
Residual Error	2	392.5	196.3		
Total	4	18420.0			

Source	DF	Seq SS
GPA	1	13145.2
AGE	1	4882.3

The session window shows the correlation coefficient for each pair of variables. The multiple correlation coefficient is significant at 0.021. Ninety-six percent of the variation from the mean is explained by regression. The regression equation is $\text{SCORE} = -44.8 + 87.6\text{GPA} + 14.5\text{AGE}$.

10-7

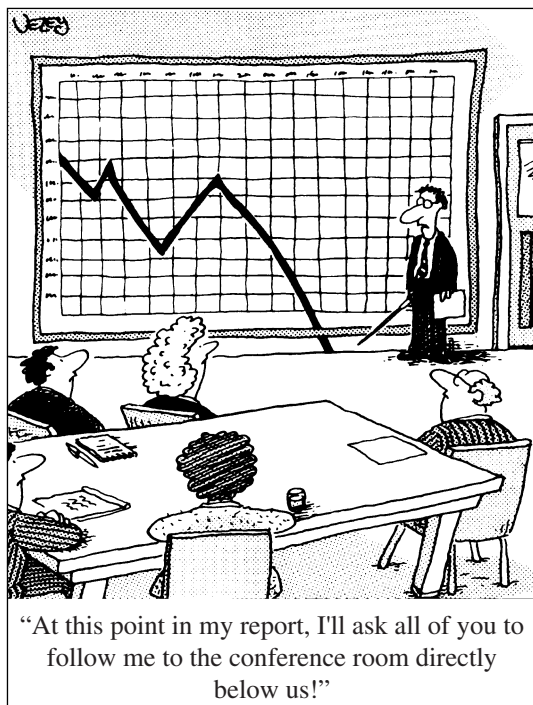
Summary

Many relationships among variables exist in the real world. One way to determine whether a relationship exists is to use the statistical techniques known as correlation and regression.

The strength and direction of the relationship are measured by the value of the correlation coefficient. It can assume values between and including +1 and -1. The closer the value of the correlation coefficient is to +1 or -1, the stronger the linear relationship is between the variables. A value of +1 or -1 indicates a perfect linear relationship. A positive relationship between two variables means that for small values of the independent variable, the values of the dependent variable will be small, and that for large values of the independent variable, the values of the dependent variable will be large. A negative relationship between two variables means that for small values of the independent variable, the values of the dependent variable will be large, and that for large values of the independent variable, the values of the dependent variable will be small.

Relationships can be linear or curvilinear. To determine the shape, one draws a scatter plot of the variables. If the relationship is linear, the data can be approximated by a straight line, called the *regression line*, or the *line of best fit*. The closer the value of r is to +1 or -1, the closer the points will fit the line.

In addition, relationships can be multiple. That is, there can be two or more independent variables and one dependent variable. A coefficient of correlation and a regression equation can be found for multiple relationships, just as they can be found for simple relationships.



Source: Cartoon by Bradford Veley, Marquette, Michigan. Reprinted with permission.

The coefficient of determination is a better indicator of the strength of a relationship than the correlation coefficient. It is better because it identifies the percentage of variation of the dependent variable that is directly attributable to the variation of the independent variable. The coefficient of determination is obtained by squaring the correlation coefficient and converting the result to a percentage.

Another statistic used in correlation and regression is the standard error of the estimate, which is an estimate of the standard deviation of the y values about the predicted y' values. The standard error of the estimate can be used to construct a prediction interval about a specific value point estimate y' of the mean of the y values for a given value of x .

Finally, remember that a significant relationship between two variables does not necessarily mean that one variable is a direct cause of the other variable. In some cases this is true, but other possibilities that should be considered include a complex relationship involving other (perhaps unknown) variables, a third variable interacting with both variables, or a relationship due solely to chance.

Important Terms

adjusted R^2 571	influential point or observation 550	multiple relationship 529	regression 529
coefficient of determination 561	least-squares line 560	negative relationship 529	regression line 544
correlation 528	lurking variable 540	Pearson product moment correlation coefficient 533	residual 560
correlation coefficient 533	marginal change 548	population correlation coefficient 537	scatter plot 530
dependent variable 529	multiple correlation coefficient 569	positive relationship 529	simple relationship 529
extrapolation 549	multiple regression 566	prediction interval 563	standard error of the estimate 561

Important Formulas

Formula for the correlation coefficient:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

Formula for the t test for the correlation coefficient:

$$t = r \sqrt{\frac{n-2}{1-r^2}} \quad \text{d.f.} = n - 2$$

The regression line equation: $y' = a + bx$, where

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

Formula for the standard error of the estimate:

$$s_{\text{est}} = \sqrt{\frac{\sum (y - y')^2}{n - 2}}$$

or

$$s_{\text{est}} = \sqrt{\frac{\sum y^2 - a \sum y - b \sum xy}{n - 2}}$$

Formula for the prediction interval for a value y' :

$$y' - t_{\alpha/2} s_{\text{est}} \sqrt{1 + \frac{1}{n} + \frac{n(x - \bar{X})^2}{n \sum x^2 - (\sum x)^2}} < y < y' + t_{\alpha/2} s_{\text{est}} \sqrt{1 + \frac{1}{n} + \frac{n(x - \bar{X})^2}{n \sum x^2 - (\sum x)^2}}$$

d.f. = $n - 2$

Formula for the multiple correlation coefficient:

$$R = \sqrt{\frac{r_{yx_1}^2 + r_{yx_2}^2 - 2r_{yx_1} \cdot r_{yx_2} \cdot r_{x_1x_2}}{1 - r_{x_1x_2}^2}}$$

Formula for the F test for the multiple correlation coefficient:

$$F = \frac{R^2/k}{(1 - R^2)/(n - k - 1)}$$

with d.f.N = $n - k$ and d.f.D = $n - k - 1$.


Formula for the adjusted R^2 :

$$R_{\text{adj}}^2 = 1 - \left[\frac{(1 - R^2)(n - 1)}{n - k - 1} \right]$$

Review Exercises


For Exercises 1 through 7, do a complete regression analysis by performing the following steps.

- Draw the scatter plot.
- Compute the value of the correlation coefficient.
- Test the significance of the correlation coefficient at $\alpha = 0.01$, using Table I.
- Determine the regression line equation.
- Plot the regression line on the scatter plot.
- Predict y' for a specific value of x .

-  1. These data represent the number of hits and the number of strikeouts for 15 players on a college baseball team. If there is a significant relationship between the variables, predict the number of strikeouts a baseball player is likely to have if he has 30 hits.


Hits x	54	16	41	43	24	21	6	2
Strikeouts y	12	6	30	33	21	29	10	4
Hits x	54	41	29	39	11	24	1	
Strikeouts y	26	20	23	27	10	12	3	

Source: University of Findlay baseball statistics.


-  2. A researcher wishes to determine if there is a relationship between the number of day-care centers and the number of group day-care homes for counties in Pennsylvania. If there is a significant relationship, predict the number of group care homes a county has if the county has 20 day-care centers.

Day-care centers x	5	28	37	16	16	48
Group day-care homes y	2	7	4	10	6	9


Source: State Department of Public Welfare.

-  3. A study is done to see whether there is a relationship between a mother's age and the number of children she has. The data are shown here. If there is a significant relationship, predict the number of children of a mother whose age is 34.

Mother's age x	18	22	29	20	27	32	33	36
No. of children y	2	1	3	1	2	4	3	5

-  4. A study is conducted to determine the relationship between a driver's age and the number of accidents he or she has over a 1-year period. The data are shown here. (This information will be used for Exercise 8.) If there is a significant relationship, predict the number of accidents of a driver who is 28.


Driver's age x	16	24	18	17	23	27	32
No. of accidents y	3	2	5	2	0	1	1

-  5. A researcher desires to know whether the typing speed of a secretary (in words per minute) is related to the time (in hours) that it takes the secretary to learn to


use a new word processing program. The data are shown.

Speed x	48	74	52	79	83	56	85	63	88	74	90	92
Time y	7	4	8	3.5	2	6	2.3	5	2.1	4.5	1.9	1.5

If there is a significant relationship, predict the time it will take the average secretary who has a typing speed of 72 words per minute to learn the word processing program. (This information will be used for Exercises 9 and 11.)

-  6. A study was conducted with vegetarians to see whether the number of grams of protein each ate per day was related to diastolic blood pressure. The data are given here. (This information will be used for Exercises 10 and 12.) If there is a significant relationship, predict the diastolic pressure of a vegetarian who consumes 8 grams of protein per day.

Grams x	4	6.5	5	5.5	8	10	9	8.2	10.5
Pressure y	73	79	83	82	84	92	88	86	95

-  7. A researcher wishes to determine the relationship between the number of cows (in thousands) in counties in southwestern Pennsylvania and the milk production (in millions of pounds). The data are shown. Describe the relationship.

Cows x	70	3	194	12	46	65
Pounds y	115	5	289	15	72	92

Source: Pittsburgh Tribune-Review.

- For Exercise 4, find the standard error of the estimate.
- For Exercise 5, find the standard error of the estimate.
- For Exercise 6, find the standard error of the estimate.
- For Exercise 5, find the 90% prediction interval for time when the speed is 72 words per minute.
- For Exercise 6, find the 95% prediction interval for pressure when the number of grams is 8.
- (Opt.) A study found a significant relationship among a person's years of experience on a particular job x_1 , the number of workdays missed per month x_2 , and the person's age y . The regression equation is $y' = 12.8 + 2.09x_1 + 0.423x_2$. Predict a person's age if he or she has been employed for 4 years and has missed 2 workdays a month.
- (Opt.) Find R when $r_{yx_1} = 0.681$ and $r_{yx_2} = 0.872$ and $r_{x_1x_2} = 0.746$.
- (Opt.) Find R_{adj}^2 when $R = 0.873$, $n = 10$, and $k = 3$.

Statistics Today

Do Dust Storms Affect Respiratory Health?—Revisited

The researchers correlated the dust pollutant levels in the atmosphere and the number of daily emergency room visits for several respiratory disorders, such as bronchitis, sinusitis, asthma, and pneumonia. Using the Pearson correlation coefficient, they found overall a significant but low correlation, $r = 0.13$, for bronchitis visits only. However, they found a much higher correlation value for sinusitis, P -value = 0.08, when pollutant levels exceeded maximums set by the Environmental Protection Agency (EPA). In addition, they found statistically significant correlation coefficients $r = 0.94$ for sinusitis visits and $r = 0.74$ for upper-respiratory-tract infection visits 2 days after the dust pollutants exceeded the maximum levels set by the EPA.

Data Analysis

The Data Bank is found in Appendix D, or on the World Wide Web by following links from www.mhhe.com/math/stat/bluman/

- From the Data Bank, choose two variables that might be related: for example, IQ and educational level; age and cholesterol level; exercise and weight; or weight and systolic pressure. Do a complete correlation and regression analysis by performing the following steps. Select a random sample of at least 10 subjects.
 - Draw a scatter plot.
 - Compute the correlation coefficient.
 - Test the hypothesis $H_0: \rho = 0$.
 - Find the regression line equation.
 - Summarize the results.
- Repeat Exercise 1, using samples of values of 10 or more obtained from Data Set V in Appendix D. Let x = the number of suspensions and y = the enrollment size.
- Repeat Exercise 1, using samples of 10 or more values obtained from Data Set XIII. Let x = the number of beds and y = the number of personnel employed.

Chapter Quiz

Determine whether each statement is true or false. If the statement is false, explain why.

- A negative relationship between two variables means that for the most part, as the x variable increases, the y variable increases.
- A correlation coefficient of -1 implies a perfect linear relationship between the variables.
- Even if the correlation coefficient is high or low, it may not be significant.
- When the correlation coefficient is significant, one can assume x causes y .
- It is not possible to have a significant correlation by chance alone.
- In multiple regression, there are several dependent variables and one independent variable.
- To test the significance of r , a(n) _____ test is used.

a. t	c. χ^2
b. F	d. None of the above
- The test of significance for r has _____ degrees of freedom.

a. 1	c. $n - 1$
b. n	d. $n - 2$
- The equation of the regression line used in statistics is

a. $x = a + by$	c. $y' = a + bx$
b. $y = bx + a$	d. $x = ay + b$
- The coefficient of determination is

a. r	c. a
b. r^2	d. b

Complete the following statements with the best answer.

Select the best answer.


- The strength of the relationship between two variables is determined by the value of

a. r	c. x
b. a	d. s_{est}
- A statistical graph of two variables is called a(n) _____.
- The x variable is called the _____ variable.
- The range of r is from _____ to _____.


- 15. The sign of r and _____ will always be the same.
- 16. The regression line is called the _____.
- 17. If all the points fall on a straight line, the value of r will be _____ or _____.

For Exercises 18 through 21, do a complete regression analysis.


- a. Draw the scatter plot.
- b. Compute the value of the correlation coefficient.
- c. Test the significance of the correlation coefficient at $\alpha = 0.05$.
- d. Determine the regression line equation.
- e. Plot the regression line on the scatter plot.
- f. Predict y' for a specific value of x .

 **18.** A medical researcher wants to determine the relationship between the price per dose of prescription drugs in the United States and the price of the same dose in Australia. The data are shown. Describe the relationship.

U.S. price x	3.31	3.16	2.27	3.13	2.54	1.98	2.22
Australian price y	1.29	1.75	0.82	0.83	1.32	0.84	0.82


 **19.** A study is conducted to determine the relationship between a driver's age and the number of accidents he or she has over a 1-year period. The data are shown here. If there is a significant relationship, predict the number of accidents of a driver who is 64.

Driver's age x	63	65	60	62	66	67	59
No. of accidents y	2	3	1	0	3	1	4

 **20.** A researcher desires to know if the age of a child is related to the number of cavities he or she has. The data are shown here. If there is a significant

relationship, predict the number of cavities for a child of 11.


Age of child x	6	8	9	10	12	14
No. of cavities y	2	1	3	4	6	5

 **21.** A study is conducted with a group of dieters to see if the number of grams of fat each consumes per day is related to cholesterol level. The data are shown here. If there is a significant relationship, predict the cholesterol level of a dieter who consumes 8.5 grams of fat per day.

Fat grams x	6.8	5.5	8.2	10	8.6	9.1	8.6	10.4
Cholesterol level y	183	201	193	283	222	250	190	218

- 22.** For Exercise 20, find the standard error of the estimate.
- 23.** For Exercise 21, find the standard error of the estimate.
- 24.** For Exercise 20, find the 90% prediction interval of the number of cavities for a 7-year-old.
- 25.** For Exercise 21, find the 95% prediction interval of the cholesterol level of a person who consumes 10 grams of fat.
- 26. (Opt.)** A study was conducted, and a significant relationship was found among the number of hours a teenager watches television per day x_1 , the number of hours the teenager talks on the telephone per day x_2 , and the teenager's weight y . The regression equation is $y' = 98.7 + 3.82x_1 + 6.51x_2$. Predict a teenager's weight if she averages 3 hours of TV and 1.5 hours on the phone per day.
- 27. (Opt.)** Find R when $r_{yx_1} = 0.561$ and $r_{yx_2} = 0.714$ and $r_{x_1x_2} = 0.625$.
- 28. (Opt.)** Find R_{adj}^2 when $R = 0.774$, $n = 8$, and $k = 2$.

Critical Thinking Challenges

 When the points in a scatter plot show a curvilinear trend rather than a linear trend, statisticians have methods of fitting curves rather than straight lines to the data, thus obtaining a better fit and a better prediction model. One type of curve that can be used is the logarithmic regression curve. The data shown are the number of items of a new product sold over a period of 15 months at a certain store. Notice that sales rise during the beginning months and then level off later on.

Month x	1	3	6	8	10	12	15
No. of items sold y	10	12	15	19	20	21	21

- 1. Draw the scatter plot for the data.
- 2. Find the equation of the regression line.

- 3. Describe how the line fits the data.
- 4. Using the log key on your calculator, transform the x values into $\log x$ values.
- 5. Using the $\log x$ values instead of the x values, find the equation of a and b for the regression line.
- 6. Next, plot the curve $y = a + b \log x$ on the graph.
- 7. Compare the line $y = a + bx$ with the curve $y = a + b \log x$ and decide which one fits the data better.
- 8. Compute r , using the x and y values; then compute r , using the $\log x$ and y values. Which is higher?
- 9. In your opinion, which (the line or the logarithmic curve) would be a better predictor for the data? Why?



Data Projects

Where appropriate, use MINITAB, the TI-83 Plus, the TI-84 Plus, or a computer program of your choice to complete the following exercises.

- Select two variables that might be related, such as the age of a person and the number of cigarettes the person smokes, or the number of credits a student has and the number of hours the student watches television. Sample at least 10 people.
 - Write a brief statement as to the purpose of the study.
 - Define the population.
 - State how the sample was selected.
 - Show the raw data.
 - Draw a scatter plot for the data values.
 - Write a statement analyzing the scatter plot.
 - Compute the value of the correlation coefficient.
 - Test the significance of r . (State the hypotheses, select α , find the critical values, make the decision, and analyze the results.)
- For the data in Exercise 1, use MINITAB to answer these.
 - Does a linear correlation exist between x and y ?
 - If so, find the regression equation.
 - Explain how good a model the regression equation is by finding the coefficient of determination and coefficient of correlation and interpreting the strength of these values.
 - Find the prediction interval for y . Use the α value that you selected in Exercise 1.

You may use the following websites to obtain raw data:

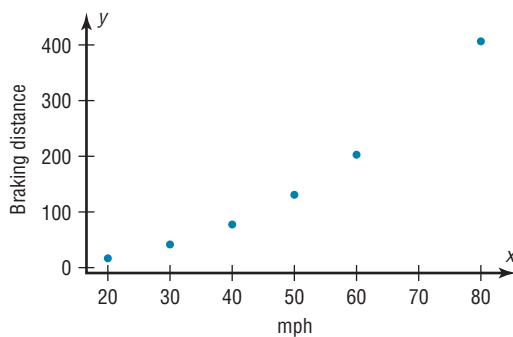
Visit the data sets at the book's website found at <http://www.mhhe.com/math/stat/bluman>
 Click on the 6th edition.
<http://lib.stat.cmu.edu/DASL>
<http://www.statcan.ca>

Answers to Applying the Concepts

Section 10-3 Stopping Distances

- The independent variable is miles per hour (mph).
- The dependent variable is braking distance (feet).
- Miles per hour is a continuous quantitative variable.
- Braking distance is a continuous quantitative variable.
- A scatter plot of the data is shown.

Scatterplot of braking distance vs. mph



- Changing the distances between the mph increments will change the appearance of the relationship.
- There is a positive relationship between the two variables—higher speeds are associated with longer braking distances.
- The strong relationship between the two variables suggests that braking distance can be accurately predicted from mph. We might still have some concern about the curve in the data.
- Answers will vary. Some other variables that might affect braking distance include road conditions, driver response time, and condition of the brakes.
- The correlation coefficient is $r = 0.966$.
- The value for $r = 0.966$ is significant at $\alpha = 0.05$. This confirms the strong positive relationship between the variables.

Section 10-4 Stopping Distances Revisited

- The linear regression equation is $\bar{y}' = -151.90 + 6.4514x$.
- The slope says that for each additional mile per hour a car is traveling, we expect the stopping distance to

increase by 6.45 feet, on average. The y intercept is the braking distance we would expect for a car traveling 0 mph—this is meaningless in this context, but is an important part of the model.

- $y' = -151.90 + 6.4514(45) = 138.4$ The braking distance for a car traveling 45 mph is approximately 138 feet.
- $y' = -151.90 + 6.4514(100) = 493.2$ The braking distance for a car traveling 100 mph is approximately 493 feet.
- It is not appropriate to make predictions of braking distance for speeds outside of the given data values (for example, the 100 mph above) because we know nothing about the relationship between the two variables outside of the range of the data.

Section 10–5 Interpreting Simple Linear Regression

- Both variables are moving in the same direction. In other words, the two variables are positively associated. We know this because the correlation coefficient is positive.
- The unexplained variation of 3026.49 measures the distances from the prediction line to the actual values.
- The slope of the regression line is 0.725983.
- The y intercept is 16.5523.
- The critical value of 0.378419 can be found in a table.
- The allowable risk of making a type I error is 0.10, the level of significance.
- The variation explained by the regression is 0.631319, or about 63.1%.
- The average scatter of points about the regression line is 12.9668, the standard error of the estimate.
- The null hypothesis is that there is no correlation, $H_0: \rho = 0$.
- We compare the test statistic of 0.794556 to the critical value to see if the null hypothesis should be rejected.
- Since $0.794556 > 0.378419$, we reject the null hypothesis and find that there is enough evidence to conclude that the correlation is not equal to zero.

Section 10–6 More Math Means More Money

- The dependent variable is yearly income 10 years after high school.
- The independent variables are number of math courses taken and number of hours worked per week during the senior year of high school.
- Multiple regression assumes that the independent variables are not highly correlated.
- We expect a person's yearly income 10 years after high school to be \$4540 more, on average, for each additional math course taken, all other variables held constant. We expect a person's yearly income 10 years after high school to be \$1290 more, on average, for each additional hour worked per week during the senior year of high school, all other variables held constant.
- $y' = 6000 + 4540(8) + 1290(20) = 68,120$. The predicted yearly income 10 years after high school is \$68,120.
- The multiple correlation coefficient of 0.77 means that there is a fairly strong positive relationship between the independent variables (number of math courses and hours worked during senior year of high school) and the dependent variable (yearly income 10 years after high school).
- $R^2 = (0.77)^2 = 0.5929$
- $$R_{\text{adj}}^2 = 1 - \left[\frac{(1 - R^2)(n - 1)}{n - k - 1} \right]$$

$$= 1 - \left[\frac{(1 - 0.5929)(20 - 1)}{20 - 2 - 1} \right]$$

$$= 1 - \left[\frac{(0.4071)(19)}{17} \right] = 0.5450$$
- The equation appears to be a fairly good predictor of income, since 54.5% of the variation in yearly income 10 years after high school is explained by the regression model.
- Answers will vary. One possible answer is that yearly income 10 years after high school increases with more math classes and more hours of work during the senior year of high school. The number of math classes has a higher coefficient, so more math does mean more money!

