Intermediate Analytics

Fatemeh Ahmadi

ALY 6015

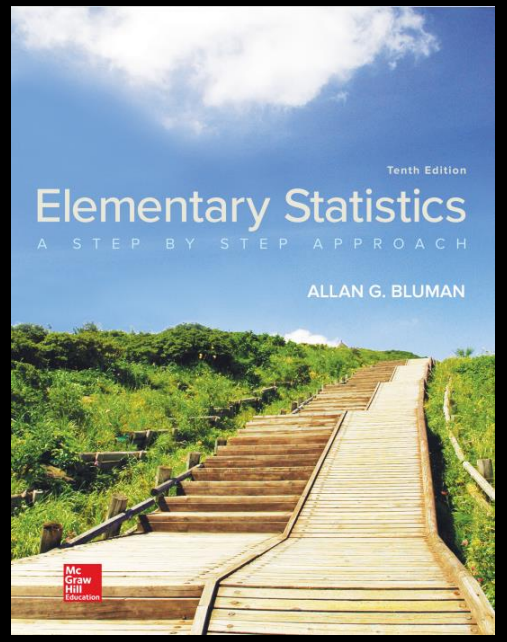
More on Correlation and Regression

Slides are mainly borrowed from the textbook:

- ***Elementary Statistics: A Step by-Step Approach****. 10th Edition,* ***Allen Bluman, McGraw Hill***

You will learn in this course:

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



1. Are two or more variables linearly 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?

Scatter Plots and Correlation

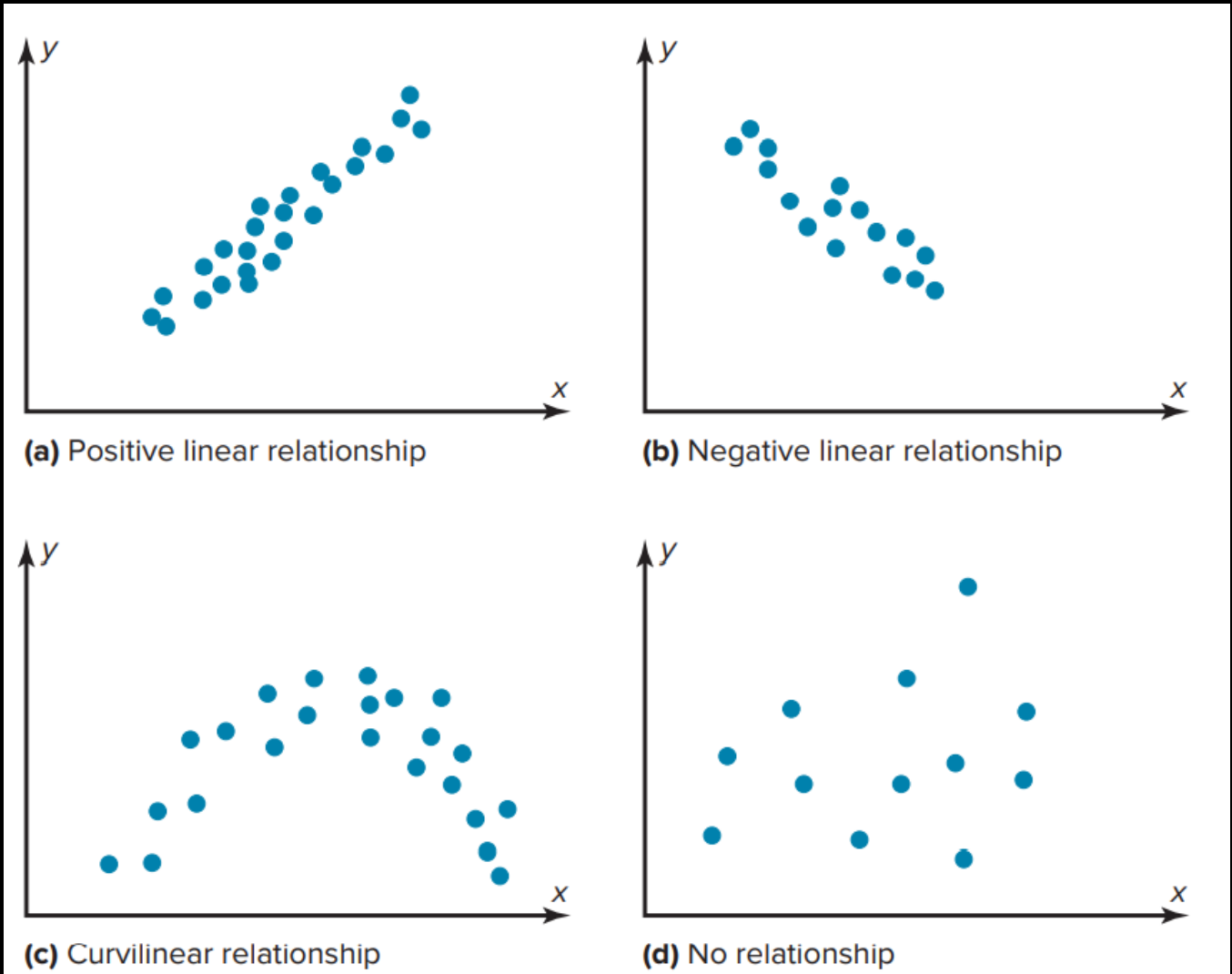
Two kinds of variables are called the **independent variable** and the **dependent variable**.

- The **independent variable** is the variable in regression that can be controlled or manipulated. In this case, the 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.

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

Types of Relationships



- 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.

Scatter Plots

- 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.
- Researchers look for various types of patterns in scatter plots. For example, in Figure 10 –1(a), the pattern in the points of the scatter plot shows a **positive linear** relationship. Here, as the values of the independent variable (*x* variable) increase, the values of the dependent variable (*y* variable) increase. Also, the points form somewhat of a straight line going in an upward direction from left to right.
- The pattern of the points of the scatter plot shown in Figure 10–1(b) shows a **negative linear relationship**. In this case, as the values of the independent variable increase, the values of the dependent variable decrease. Also, the points show a somewhat straight line going in a downward direction from left to right.

Example 1 - Car Rental Companies

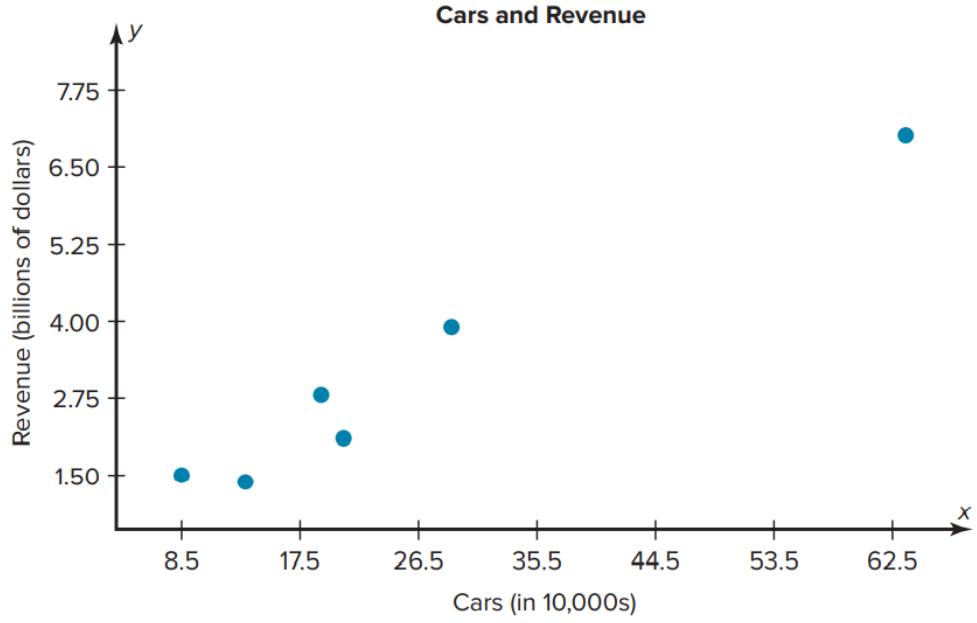
Construct a scatter plot for the data shown for car rental companies in the United States for a recent year.

Company	Cars (in ten thousands)	Revenue (in billions)
A	63.0	$7.0
B	29.0	3.9
C	20.8	2.1
D	19.1	2.8
E	13.4	1.4
F	8.5	1.5

Source: Auto Rental News.

In this example, it looks as if a **positive linear relationship** exists between the number of cars that an agency owns and the total revenue that is made by the company.

FIGURE 10–2 Scatter Plot for Example 10–1



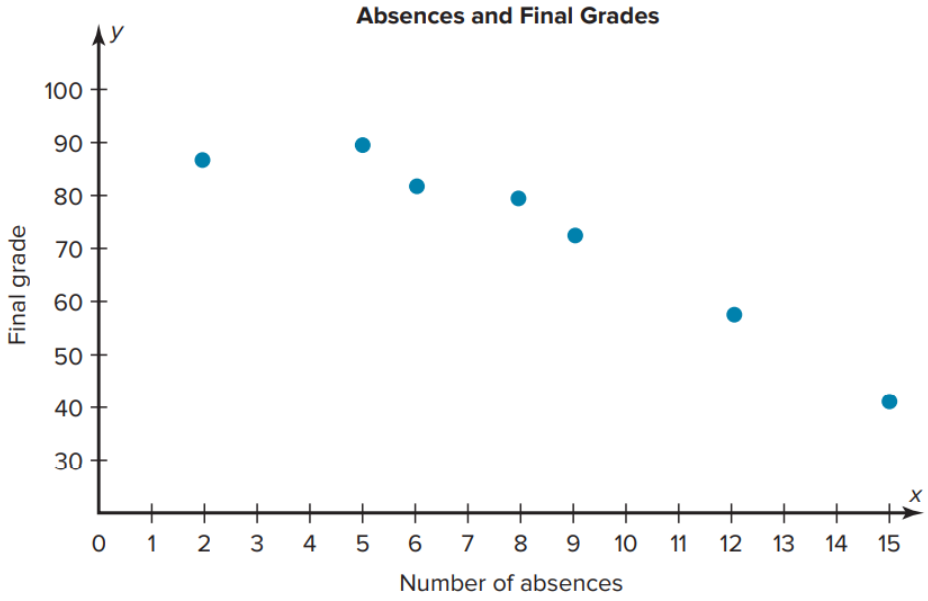
Example 2 - Absences and Final Grades

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

In this example, it looks as if a **negative linear relationship** exists between the number of student absences and the final grade of the students.

FIGURE 10–3 Scatter Plot for Example 10–2



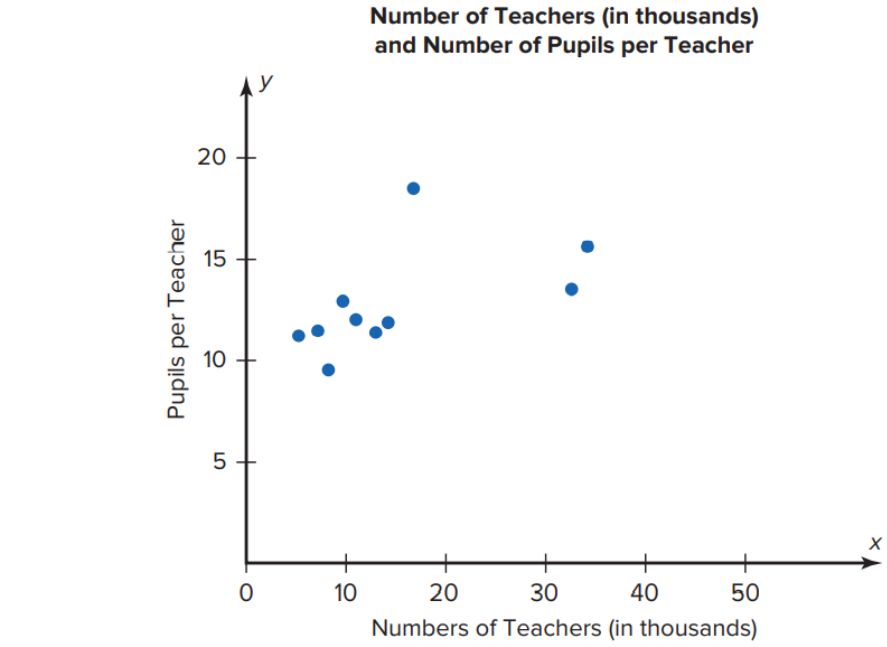
Example 3

A researcher wishes to see if there is a relationship between the number of pupils per teacher and the number of teachers (in thousands) employed by the school district. She randomly selects 10 school districts throughout the United States. The data are shown.

School district	Number of teachers (in thousands)	Pupils per teacher
1	7	12.4
2	34	14.3
3	9	14.3
4	8	9.2
5	16	18.3
6	15	12.1
7	6	12.3
8	14	12.4
9	32	15.2
10	10	13.4

Source: U.S. Department of Education.

FIGURE 10–4 Scatter Plot for Example 10–3



In this case, there is no indication of a strong positive or negative linear relationship between the number of pupils per teacher and the number of teachers (in thousands) in a school district.

Correlation

Correlation Coefficient

- Statisticians use a measure called the **correlation coefficient** to determine the strength of the linear relationship between two variables. There are several types of correlation coefficients.
- The population correlation coefficient denoted by the Greek letter ρ is the correlation computed by using all possible pairs of data values (x, y) taken from a population.
- The linear correlation coefficient computed from the sample data measures the **strength and direction** of a linear relationship between two **quantitative variables**. The symbol for the sample correlation coefficient is r.
- The linear correlation coefficient 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.

Correlation

- The range of the linear 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. When the value of *r* is 0 or close to zero, it implies only that there is no linear relationship between the variables. The data may be related in some other **nonlinear way**.

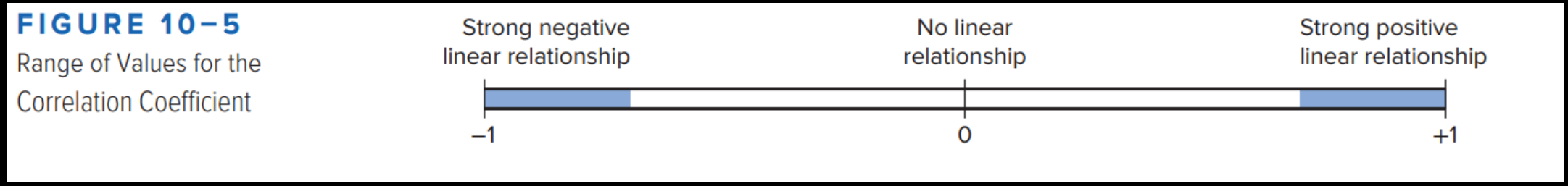
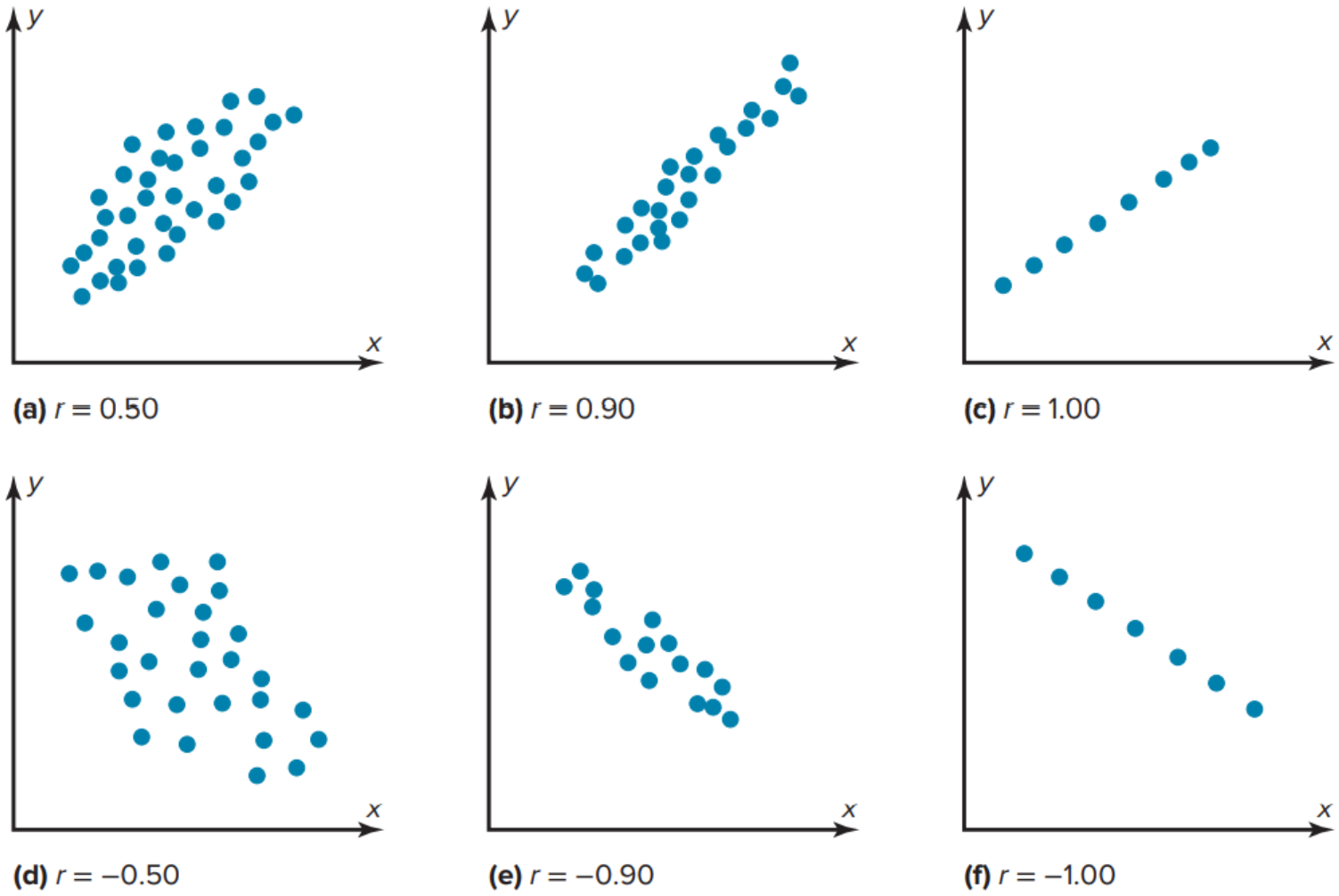


FIGURE 10–5
Range of Values for the Correlation Coefficient

Correlation

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



Correlation

Properties of the Linear Correlation Coefficient

1. The correlation coefficient is a unitless measure.
2. The value of r will always be between -1 and $+1$ inclusively. That is, $-1 \leq r \leq 1$.
3. If the values of x and y are interchanged, the value of r will be unchanged.
4. If the values of x and/or y are converted to a different scale, the value of r will be unchanged.
5. The value of r is sensitive to outliers and can change dramatically if they are present in the data.

Assumptions for the Correlation Coefficient

1. The sample is a random sample.
2. The data pairs fall approximately on a straight line and are measured at the interval or ratio level.
3. The variables have a bivariate normal distribution. (This means that given any specific value of x, the y values are normally distributed; and given any specific value of y, the x values are normally distributed.)

Correlation

Formula for the Linear Correlation Coefficient *r*

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

where n is the number of data pairs.

Procedure Table

Finding the Value of the Linear Correlation Coefficient

Step 1 Make a table as shown.

x	y	xy	x^2	y^2

Step 2 Place the values of x in the x column and the values of y in the y column. Multiply each x value by the corresponding y value, and place the products in the xy column.

Square each x value and place the squares in the x^2 column.

Square each y value and place the squares in the y^2 column.

Find the sum of each column.

Step 3 Substitute in the formula and find the value for r.

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

where n is the number of data pairs.

Example 1 - Car Rental Companies

Compute the linear correlation coefficient for the data in Example 10–1.

SOLUTION

Step 1 Make a table as shown here.

Company	Cars x (in ten thousands)	Revenue y (in billions)	xy	x^2	y^2
A	63.0	$7.0			
B	29.0	3.9			
C	20.8	2.1			
D	19.1	2.8			
E	13.4	1.4			
F	8.5	1.5			

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

Example1 (Car Rental Companies)

Company	Cars x (in 10,000s)	Revenue y (in billions of dollars)	xy	x^2	y^2
A	63.0	7.0	441.00	3969.00	49.00
B	29.0	3.9	113.10	841.00	15.21
C	20.8	2.1	43.68	432.64	4.41
D	19.1	2.8	53.48	364.81	7.84
E	13.4	1.4	18.76	179.56	1.96
F	8.5	1.5	12.75	72.25	2.25
	$\Sigma x = 153.8$	$\Sigma y = 18.7$	$\Sigma xy = 682.77$	$\Sigma x^2 = 5859.26$	$\Sigma y^2 = 80.67$

Step 3 Substitute in the formula and solve for r.

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

$$= \frac{(6)(682.77) - (153.8)(18.7)}{\sqrt{[(6)(5859.26) - (153.8)^2][(6)(80.67) - (18.7)^2]}} = 0.982$$

The linear correlation coefficient suggests a **strong positive linear relationship** between the number of cars a rental agency has and its annual revenue. That is, the more cars a rental agency has, the more annual revenue the company will have.

Example 2 - Absences and Final Grades

The value of r suggests a **strong negative linear 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 his or her grade.

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.

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

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

Example 3 - Numbers of Teachers and Pupils per Teacher

Compute the value of the linear correlation coefficient for the data given in Example 10–3 for the number of teachers (in thousands) and the number of pupils per teacher.

SOLUTION

Step 1 Make a table.

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

School district	Number of teachers, x	Pupils per teacher, y	xy	x^2	y^2
1	7	12.4	86.8	49	153.76
2	34	14.3	486.2	1156	204.49
3	9	14.3	128.7	81	204.49
4	8	9.2	73.6	64	84.64
5	16	18.3	292.8	256	334.89
6	15	12.1	181.5	225	146.41
7	6	12.3	73.8	36	151.29
8	14	12.4	173.6	196	153.76
9	32	15.2	486.4	1024	231.04
10	10	13.4	134	100	179.56
	$\Sigma x = 151$	$\Sigma y = 133.9$	$\Sigma xy = 2117.4$	$\Sigma x^2 = 3187$	$\Sigma y^2 = 1844.33$

Example3 (Numbers of Teachers and Pupils per Teacher)

Step 3 Substitute in the formula and solve for r.

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

$$= \frac{10(2117.4) - (151)(133.9^2)}{\sqrt{[10(3187) - (151)^2][10(1844.33) - (133.9^2)]}}$$

$$= \frac{955.1}{\sqrt{(9069)(514.09)}}$$

$$= \frac{955.1}{2159.23} = 0.442$$

The value of r indicates **a weak positive linear relationship** between the number of teachers (in thousands) employed and the number of pupils per teacher.

The Significance of the Correlation Coefficient

Assumptions for Testing the Significance of the Linear Correlation Coefficient

1. The data are quantitative and are obtained from a simple random sample.
2. The scatter plot shows that the data are approximately linearly related.
3. There are no outliers in the data.
4. The variables x and y must come from normally distributed populations.

In this book, the assumptions will be stated in the exercises; however, when encountering statistics in other situations, you must check to see that these assumptions have been met before proceeding.

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.

The Significance of the Correlation Coefficient

Formula for the *t* Test for the Correlation Coefficient

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

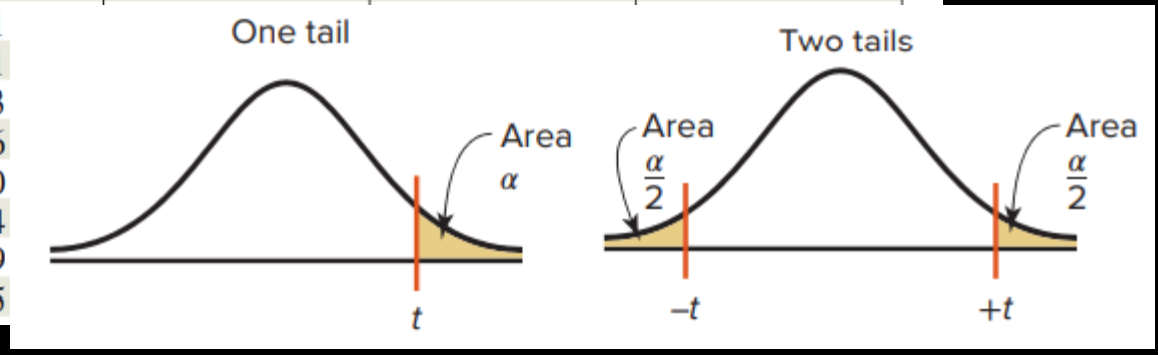
with degrees of freedom equal to $n - 2$, where n is the number of ordered pairs (x, y).

You do 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 A*. Also, when you are testing the significance of a correlation coefficient, both variables x and y must come from normally distributed populations.

The Significance of the Correlation Coefficient

TABLE F The *t* Distribution

	Confidence intervals	80%	90%	95%	98%	99%
	One tail, α	0.10	0.05	0.025	0.01	0.005
d.f.	Two tails, α	0.20	0.10	0.05	0.02	0.01
1		3.078	6.314	12.706	31.821	63.657
2		1.886	2.920	4.303	6.965	9.925
3		1.638	2.353	3.182	4.541	5.841
4		1.533	2.132	2.776	3.747	4.604
5		1.476	2.015	2.571	3.365	4.032
6		1.440	1.943	2.447	3.143	3.707
7		1.415	1.895	2.365	2.998	3.499
8		1.397	1.860	2.306	2.896	3.355
9		1.383	1.833	2.262	2.821	3.250
10		1.372	1.812	2.228	2.764	3.169
11		1.363	1.796	2.201	2.718	3.106
12		1.356	1.782	2.179	2.681	3.055
13		1.350	1.771			
14		1.345	1.761			
15		1.341	1.753			
16		1.337	1.746			
17		1.333	1.740			
18		1.330	1.734			
19		1.328	1.729			
20		1.325	1.725			



Example

Test the significance of the correlation coefficient found in Example 10–4.

Use $\alpha = 0.05$ and r = 0.982.

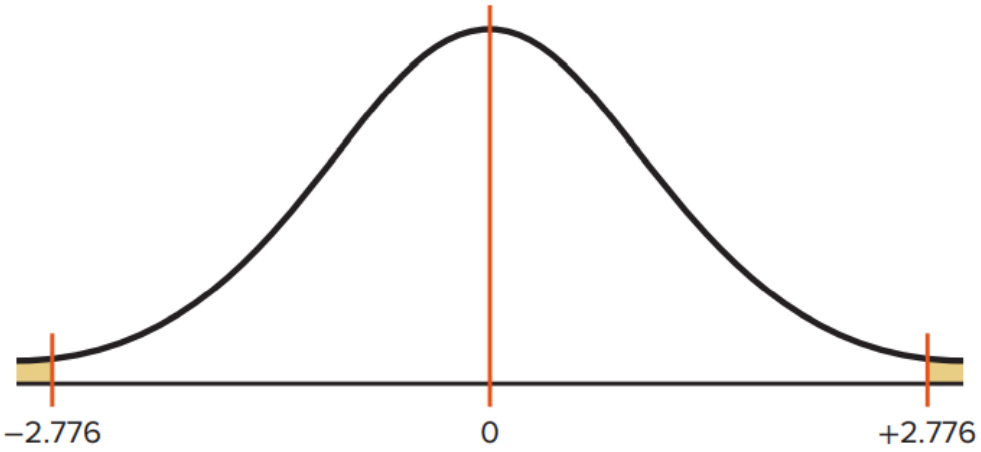
SOLUTION

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



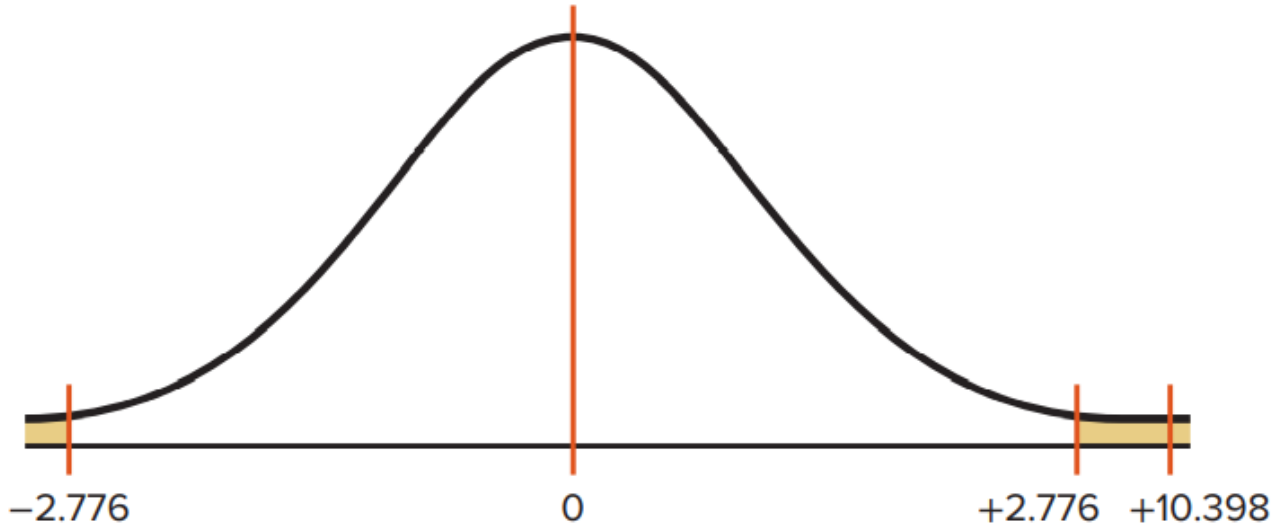
Step 3 Compute the test value.

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 0.982\sqrt{\frac{6-2}{1-(0.982)^2}} = 10.398$$

Example

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 number of cars a rental agency owns and its annual income.

The Significance of the Correlation Coefficient

The second method that can be used to test the significance of *r* is the *P-value* method.

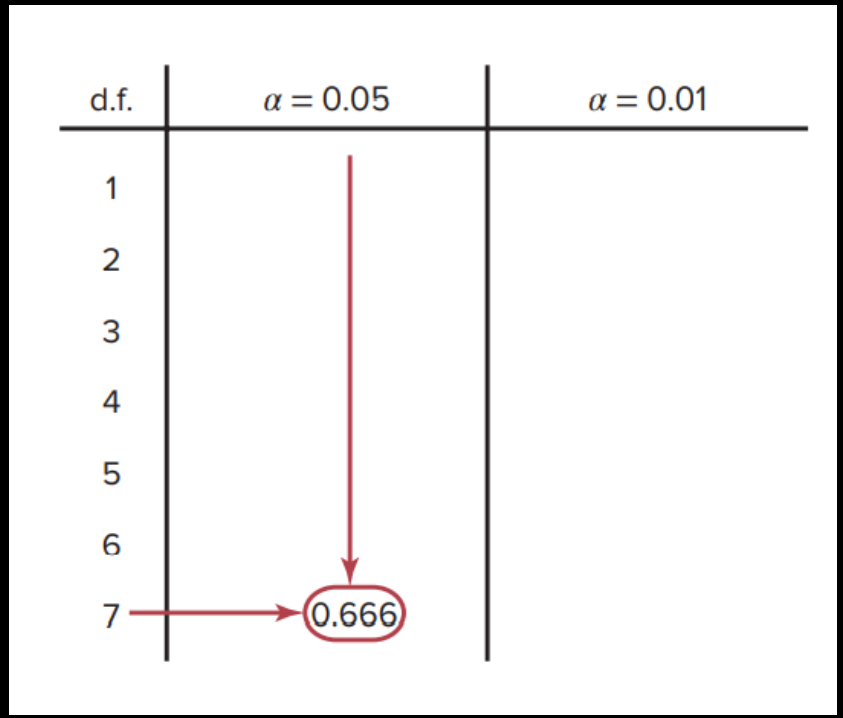
Consider an example where t = 4.059, d.f. = 4, and $\alpha = 0.05$. Using Table *F* with d.f. = 4 and the row Two tails, the value 4.059 falls between 3.747 and 4.604; hence, 0.01 < P-value < 0.02. (The P-value obtained from a calculator is 0.015.) That is, the P-value falls between 0.01 and 0.02.

The decision, then, is to reject the null hypothesis since P-value < 0.05.

Step 1	State the hypotheses.
Step 2	Find the test value. (In this case, use the *t* test.)
Step 3	Find the *P*-value. (In this case, use Table F.)
Step 4	Make the decision.
Step 5	Summarize the results.

The Significance of the Correlation Coefficient

The third method of testing the significance of r is to use Table *I* in Appendix *A*. 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. When Table *I* is used, you need not compute the *t-test* value. Table *I* is for two-tailed tests only.



The Significance of the Correlation Coefficient

TABLE I Critical Values for the PPMC

Reject H_0: $\rho = 0$ if the absolute value of r is greater than the value given in the table. The values are for a two-tailed test; d.f. $= n - 2$.

d.f.	$\alpha = 0.05$	$\alpha = 0.01$
1	0.999	0.999
2	0.950	0.999
3	0.878	0.959
4	0.811	0.917
5	0.754	0.875
6	0.707	0.834
7	0.666	0.798
8	0.632	0.765
9	0.602	0.735
10	0.576	0.708
11	0.553	0.684
12	0.532	0.661
13	0.514	0.641
14	0.497	0.623
15	0.482	0.606
16	0.468	0.590
17	0.456	0.575
18	0.444	0.561
19	0.433	0.549
20	0.423	0.537

Example

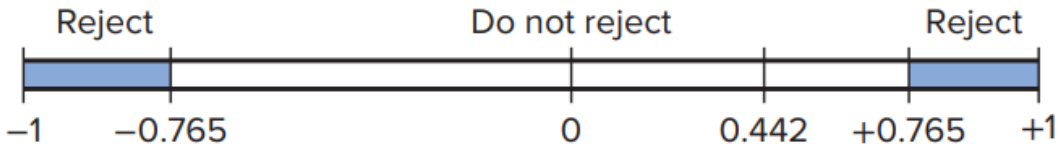
EXAMPLE 10–8

Using Table I, test the significance at $\alpha = 0.01$ of the correlation coefficient $r = 0.442$ obtained in Example 10–6.

SOLUTION

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

Since the sample size is 10, there are $n - 2 = 10 - 2 = 8$ degrees of freedom. The critical values obtained from Table I at $\alpha = 0.01$ and 8 degrees of freedom are ± 0.765. Since $0.442 < 0.765$, the decision is not to reject the null hypothesis. See Figure 10–10. Hence, there is not enough evidence to say that there is a significant linear relationship.



The Significance of the Correlation Coefficient

Generally, significance tests for correlation coefficients are two-tailed; however, they can be one-tailed. For example, if a researcher hypothesized a positive linear relationship between two variables, the hypotheses would be

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

If the researcher hypothesized a negative linear relationship between two variables, the hypotheses would be

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

In these cases, the *t* tests and the *P*-value tests would be one-tailed. Also, tables such as Table I are available for one-tailed tests. In this book, the examples and exercises will involve two-tailed tests.

Correlation and Causation

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.

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.
- 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.
- It is important to restate the fact that even if the correlation between two variables is high, it does not necessarily mean causation. There are other possibilities, such as lurking variables or just a coincidental relationship.

Regression

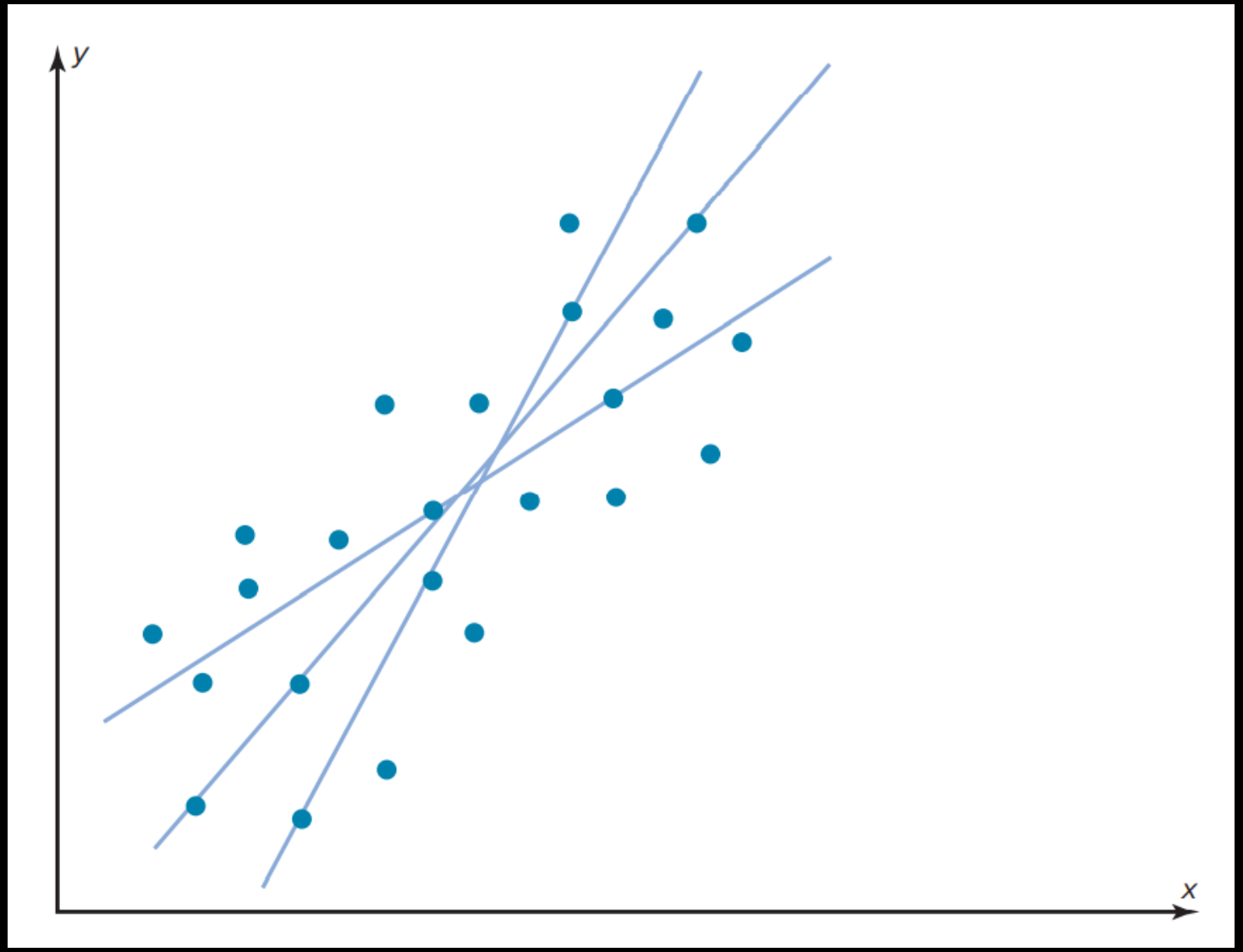
Regression

- 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 between the variables.
- The possibilities include a positive linear relationship, a negative linear relationship, a curvilinear relationship, or no discernible relationship. After the scatter plot is drawn and a linear relationship is determined, the next steps are to compute the value of the correlation coefficient and 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 dateline 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 the regression line is to enable the researcher to see the trend and make predictions on the basis of the data.

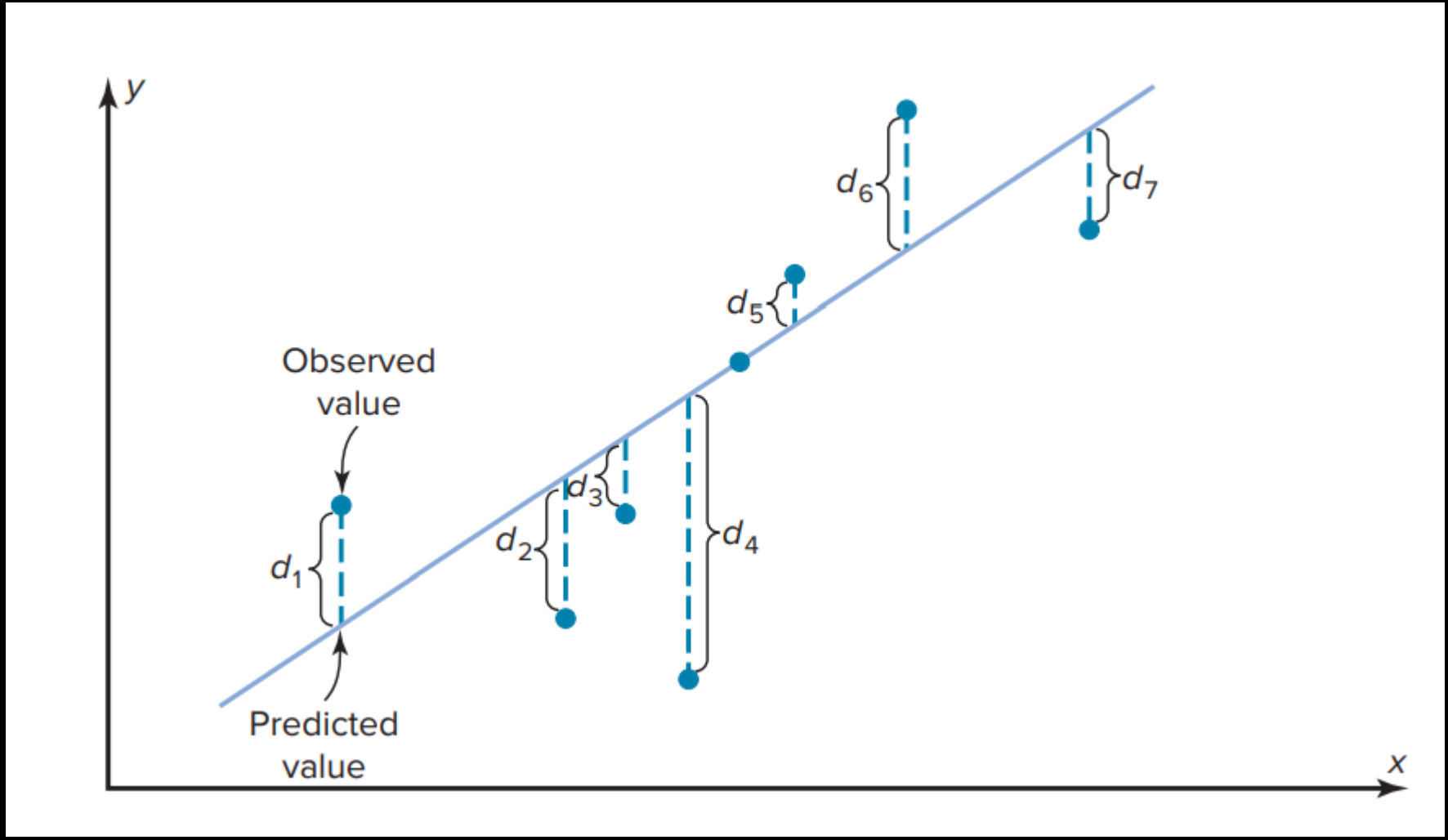
Line of Best Fit

- Next figure 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, you 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 difference between the actual value y and the predicted value y' (that is, the vertical distance) is called a *residual or a predicted error*. Residuals are used to determine the line that best describes the relationship between the two variables.
- The method used for making the residuals as small as possible is called **the method of least squares**. As a result of this method, **the regression line is also called the least squares regression line.**
- The reason you need 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.

Line of Best Fit

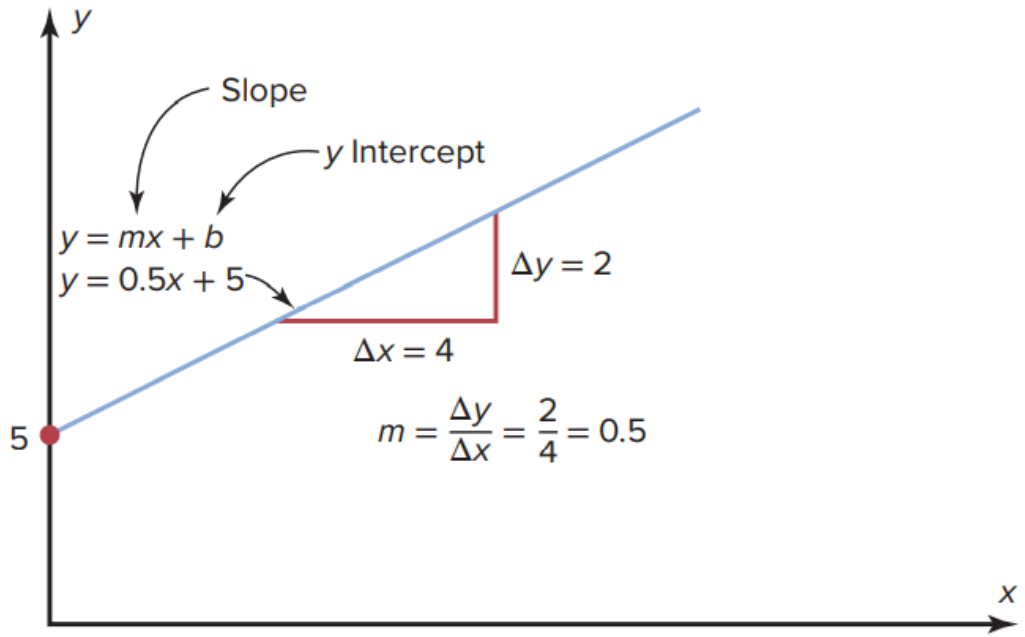


Line of Best Fit

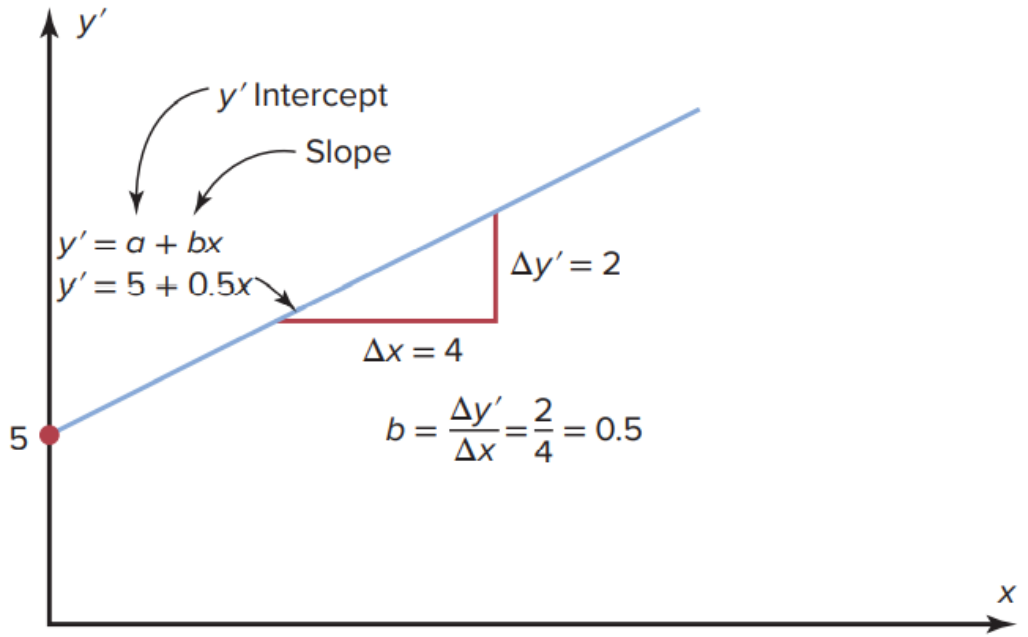


Line of Best Fit

FIGURE 10–13 A Line as Represented in Algebra and in Statistics



(a) Algebra of a line



(b) Statistical notation for a regression line

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.
- In 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.

Determination of the Regression Line Equation

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. The steps for finding the regression line equation are summarized in this Procedure Table below.

Procedure Table

Finding 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
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
$\Sigma x =$	$\Sigma y =$	$\Sigma xy =$	$\Sigma x^2 =$	$\Sigma y^2 =$

Step 3 When r is significant, substitute in the formulas to find the values of a and b for the regression line equation $y' = a + bx$.

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

Example 1- Car Rental Company

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 $n = 6$, $\Sigma x = 153.8$, $\Sigma y = 18.7$, $\Sigma xy = 682.77$, and $\Sigma x^2 = 5859.26$. Substituting in the formulas, you get

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{(18.7)(5859.26) - (153.8)(682.77)}{(6)(5859.26) - (153.8)^2} = 0.396$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{6(682.77) - (153.8)(18.7)}{(6)(5859.26) - (153.8)^2} = 0.106$$

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

$$y' = 0.396 + 0.106x$$

Example1- Car Rental Company

To graph the line, select any two points for x and find the corresponding values for y. Use any x values between 10 and 60. For example, let $x = 15$. Substitute in the equation and find the corresponding y' value.

$$\begin{aligned} y' &= 0.396 + 0.106x \\ &= 0.396 + 0.106(15) \\ &= 1.986 \end{aligned}$$

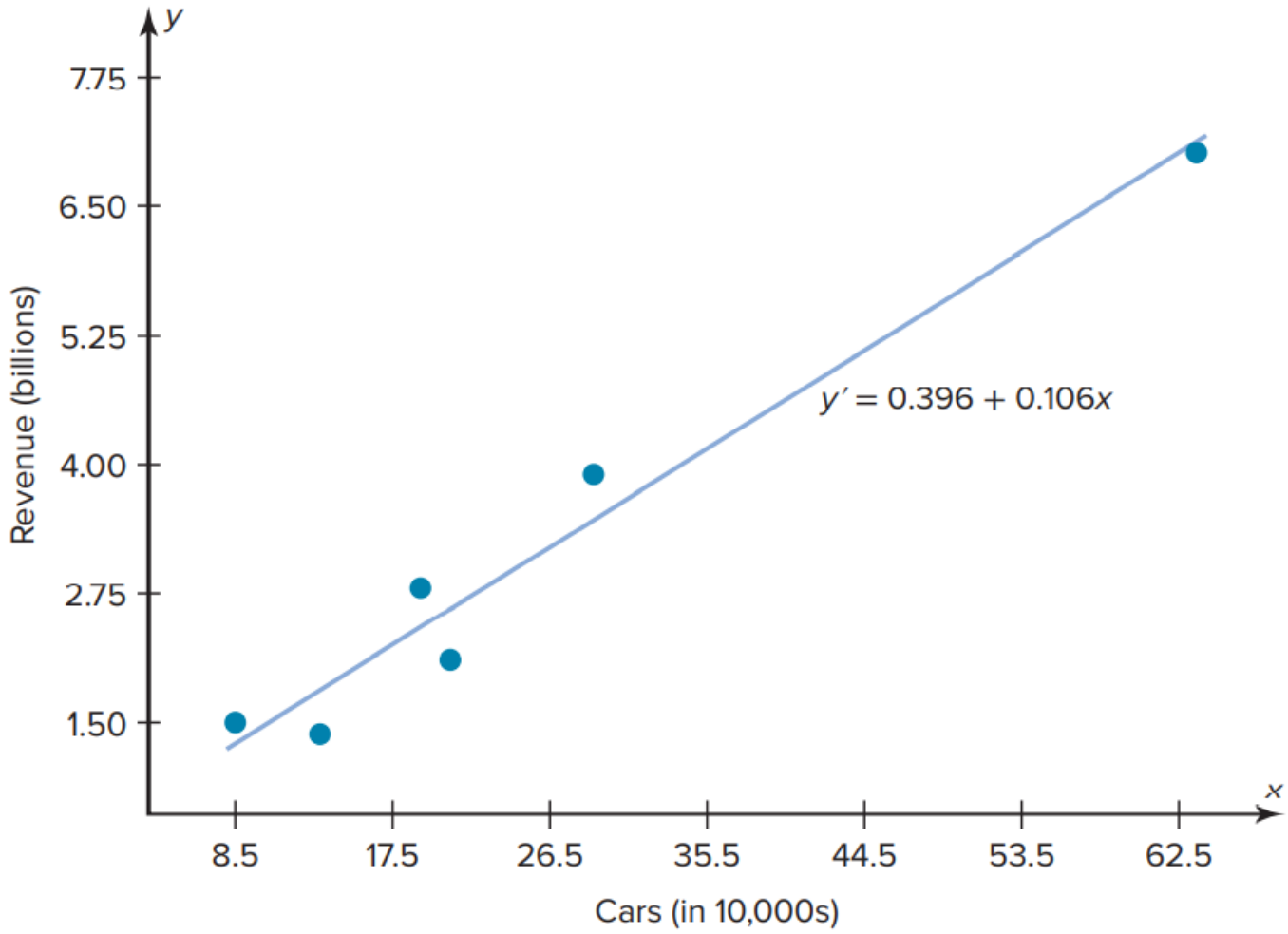
Let $x = 40$; then

$$\begin{aligned} y' &= 0.396 + 0.106x \\ &= 0.396 + 0.106(40) \\ &= 4.636 \end{aligned}$$

Then plot the two points $(15, 1.986)$ and $(40, 4.636)$ and draw a line connecting the two points. See Figure 10–14.

Example1- Car Rental Company

FIGURE 10–14 Regression Line for Example 10–9



Regression

- Note: When you draw the regression line, it is sometimes necessary to *truncate* the graph (see Chapter 2).
- This is done when the distance between the origin and the first labeled coordinate on the *x-axis* is not the same as the distance between the rest of the labeled *x* coordinates or the distance between the origin and the first labeled *y′* coordinate is not the same as the distance between the other labeled *y′* coordinates.
- When the *x-axis* or the *y-axis* has been truncated, do not use the *y′* intercept value to graph the line. When you graph the regression line, always select *x* values between the smallest *x* data value and the largest *x* data value.

Example 2- Absences and Final Grades

EXAMPLE 10–10 Absences and Final Grades

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, you get

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

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

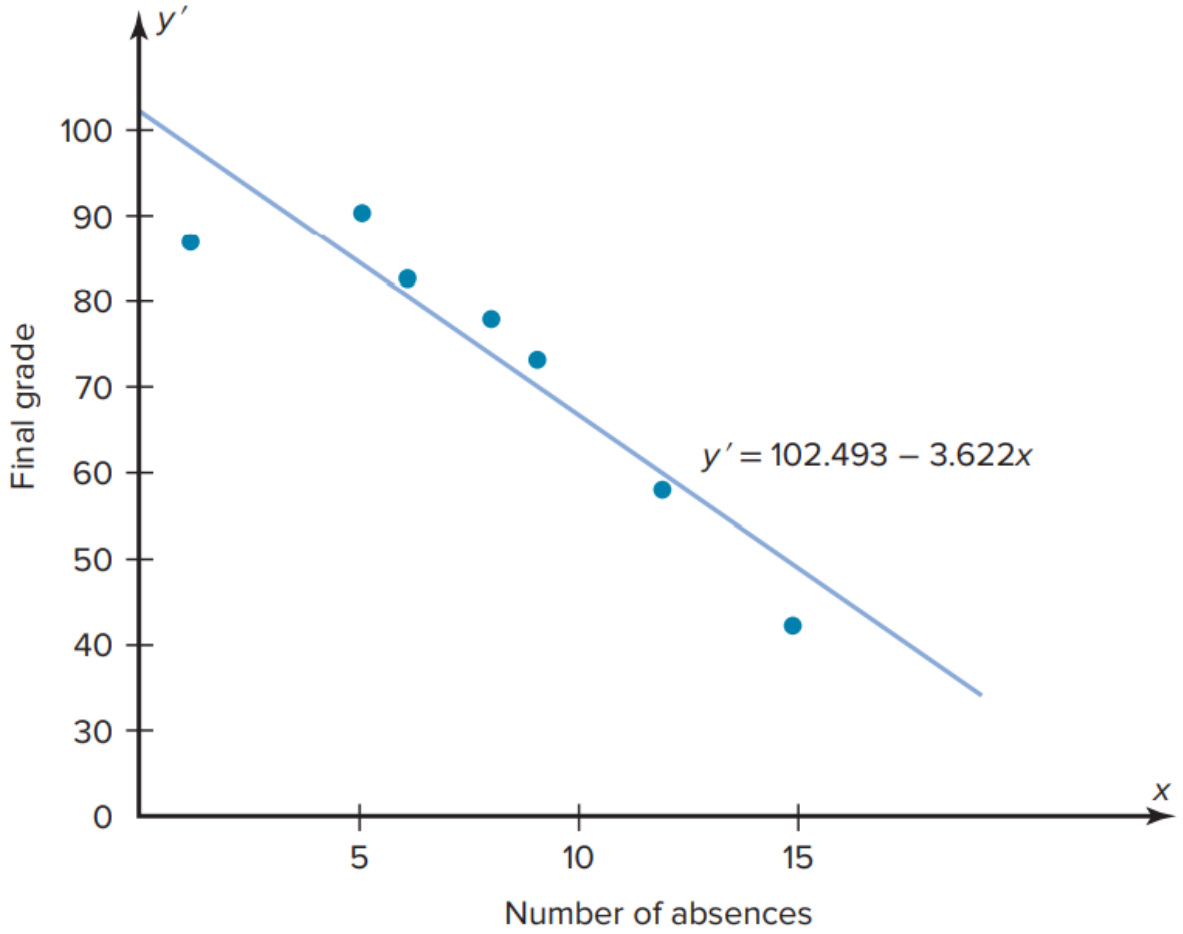
Hence, the equation of the regression line $y' = a + bx$ is

$$y' = 102.493 - 3.622x$$

Example 2- Absences and Final Grades

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

FIGURE 10–15
Regression Line for Example 10–10



Example 2- Absences and Final Grades

- 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.

- The regression line can be used to make predictions for the dependent variable.

Example3- Absences and Final Grades

EXAMPLE 10–11 Absences and Final Grades

Use the equation of the regression line in Example 10–10 to predict the final grade for a student who missed 4 classes.

SOLUTION

Substitute 4 for x in the regression line equation $y' = 102.493 - 3.622x$.

$$\begin{aligned} y' &= 102.493 - 3.622x \\ &= 102.493 - 3.622(4) \\ &= 88.005 \\ &= 88 \text{ (rounded)} \end{aligned}$$

Hence, when a student misses 4 classes, the student's grade on the final exam is predicted to be about 88.

Example 3- Absences and Final Grades

- 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 –3. 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, in Example 10 –9 the slope of the regression line is 0.106, which means for each additional increase of 10,000 cars, the value of *y* changes 0.106 unit ($106 million) on average.

Extrapolation

- 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.

Outliers

- 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.

Coefficient of Determination and Standard Error of the Estimate

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.*

Coefficient of Determination and Standard Error of the Estimate

In other words, the explained variation is the vertical distance $y' - \bar{y}$, which is the distance between the predicted value y' and the mean value $\bar{y}$. Most of the variation 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' is equal to y in each case.

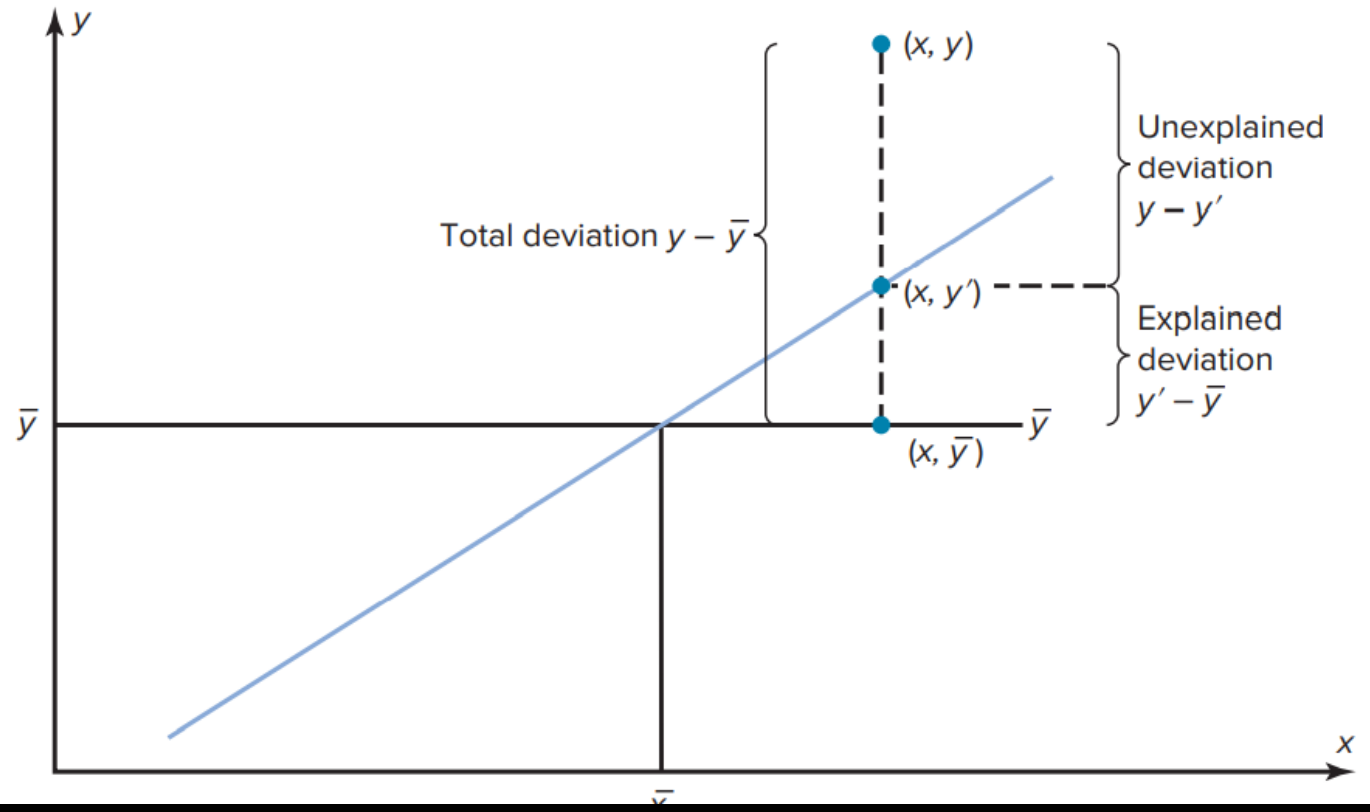
On the other hand, the variation due to chance, found by $\Sigma(y - y')^2$, is called the *unexplained variation.* In other words, the unexplained variation is the vertical distance $y - y'$, which is the distance between the observed value, y, and the predicted value y'. This variation cannot be attributed 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 $y = 10$, you get $y' = 7.6$ and $\bar{y} = 13.2$.

Coefficient of Determination and Standard Error of the Estimate

FIGURE 10–17
Deviations for the Regression Equation



Coefficient of Determination and Standard Error of the Estimate

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

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

Coefficient of Determination and Standard Error of the Estimate

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

$$(7.6 - 13.2)^2 = 31.36$$
$$(10.4 - 13.2)^2 = 7.84$$
$$(13.2 - 13.2)^2 = 0.00$$
$$(16 - 13.2)^2 = 7.84$$
$$(18.8 - 13.2)^2 = 31.36$$
$$\Sigma(y' - \bar{y})^2 = 78.4$$

Coefficient of Determination and Standard Error of the Estimate

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

$$92.8 = 78.4 + 14.4$$

Residual Plots

As previously stated, the values $y - y'$ are called *residuals* (sometimes called the *prediction errors*). These values can be plotted with the x values, and the plot, called a **residual plot,** can be used to determine how well the regression line can be used to make predictions.

The residuals for the previous example are calculated as shown.

x	y	y'	$y - y' =$ residual
1	10	7.6	$10 - 7.6 = 2.4$
2	8	10.4	$8 - 10.4 = -2.4$
3	12	13.2	$12 - 13.2 = -1.2$
4	16	16	$16 - 16 = 0$
5	20	18.8	$20 - 18.8 = 1.2$

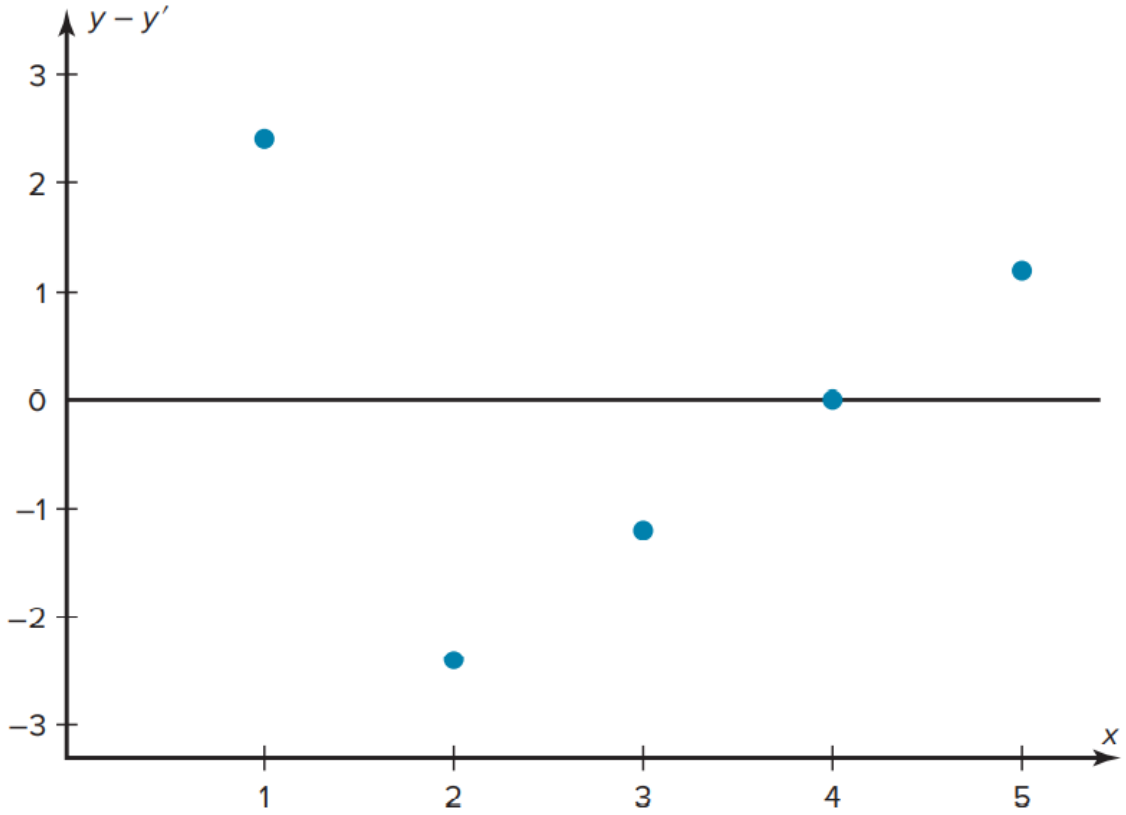
The x values are plotted using the horizontal axis, and the residuals are plotted using the vertical axis. Since the mean of the residuals is always zero, a horizontal line with a y coordinate of zero is placed on the y axis as shown in Figure 10–18.

Plot the x and residual values as shown in Figure 10–18.

x	1	2	3	4	5
$y - y'$	2.4	−2.4	−1.2	0	1.2

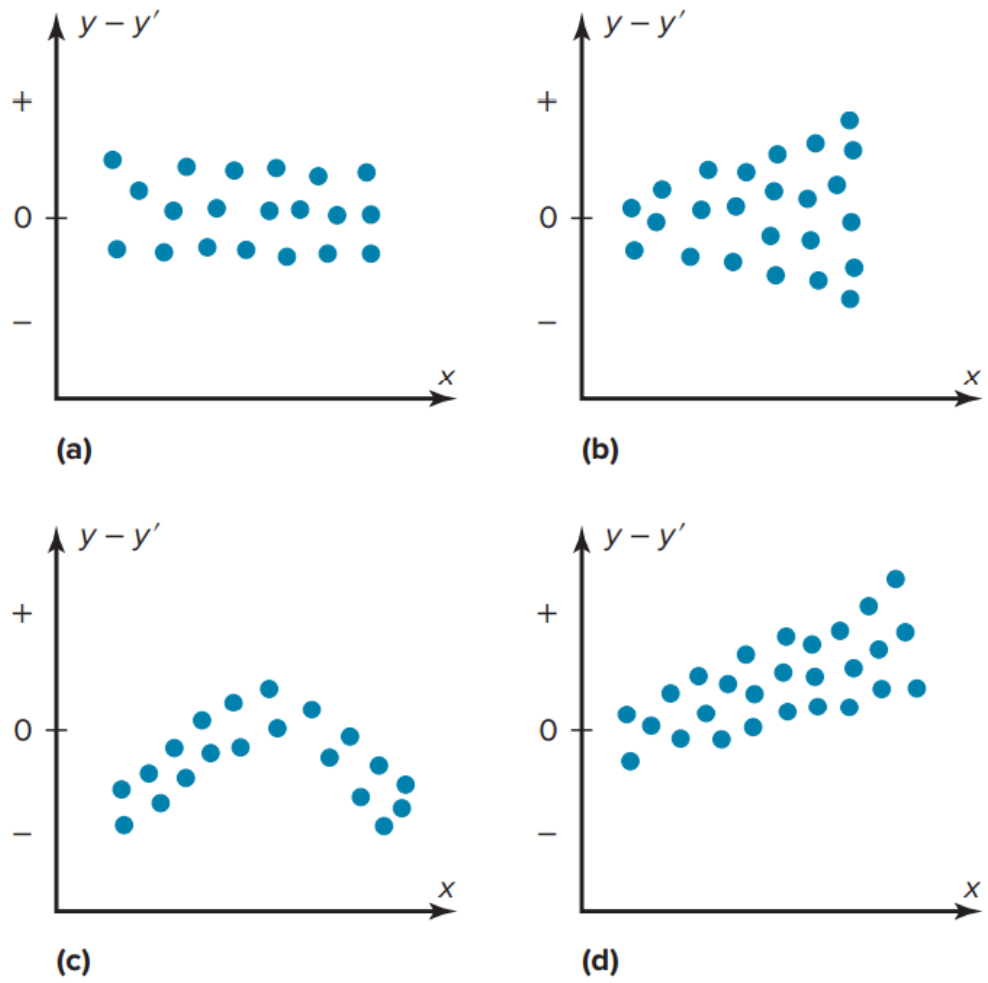
Residual Plots

To interpret a residual plot, you need to determine if the residuals form a pattern. Figure 10–19 shows four examples of residual plots. If the residual values are more or less evenly distributed about the line, as shown in Figure 10–19(a), then the relationship between x and y is linear and the regression line can be used to make predictions.



Residual Plots

FIGURE 10–19
Examples of
Residual Plots



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 the estimate is

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

EXAMPLE 10–12 Copy Machine Maintenance Costs

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

Standard Error of the Estimate

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'.

$$
\begin{aligned}
x = 1 \quad & y' = 55.57 + (8.13)(1) = 63.70 \\
x = 2 \quad & y' = 55.57 + (8.13)(2) = 71.83 \\
x = 3 \quad & y' = 55.57 + (8.13)(3) = 79.96 \\
x = 4 \quad & y' = 55.57 + (8.13)(4) = 88.09 \\
x = 6 \quad & y' = 55.57 + (8.13)(6) = 104.35
\end{aligned}
$$

Standard Error of the Estimate

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

$$62 - 63.70 = -1.70 \qquad 90 - 88.09 = 1.91$$
$$78 - 71.83 = 6.17 \qquad 93 - 88.09 = 4.91$$
$$70 - 79.96 = -9.96 \qquad 103 - 104.35 = -1.35$$

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
			$\Sigma(y-y')^2 = 169.7392$	

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

$$s_{est} = \sqrt{\frac{\Sigma(y - y')^2}{n - 2}} = \sqrt{\frac{169.7392}{6 - 2}} = 6.514$$

In this case, the standard deviation of observed values about the predicted values is 6.514.

Standard Error of the Estimate

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

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

This Procedure Table shows the alternate method for finding the standard error of the estimate.

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 as shown in the Procedure Table.

Step 2 Place the x values in the first column (the x column), and place the y values in the second column (the y column). Find the product of x and y values, and place the results in the third column. Square the y values, and place the results in the y^2 column.

Step 3 Find the sums of the y, xy, and y^2 columns. The completed table is shown here.

Standard Error of the Estimate

Step 3 Find the sums of the y, xy, and y^2 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
	$\Sigma y = 496$	$\Sigma xy = 1778$	$\Sigma y^2 = 42{,}186$

Step 4 From the regression equation $y' = 55.57 + 8.13x$, $a = 55.57$, and $b = 8.13$. Substitute in the formula and solve for s_{est}.

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

$$= \sqrt{\frac{42{,}186 - (55.57)(496) - (8.13)(1778)}{6-2}} = 6.483$$

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

References

- **Elementary Statistics: A Step-by-Step Approach, Allen Bluman, 10th Edition, McGraw Hill, 2017, ISBN 13: 978-1-259-755330, Chapter 10.**