CHAPTER 10 CORRELATION AND REGRESSION

Correlation and Regression

- Objectives:

1 Draw a scatter plot for a set of ordered pairs.

2 Compute the correlation coefficient.

3 Compute the equation of the regression line.

Introduction

- **Correlation** is a statistical method used to determine whether a linear 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?

There are two types of relationships:

simple and multiple.

In a simple relationship, there are two variables: an **independent variable** (predictor variable) and a **dependent variable** (response variable).

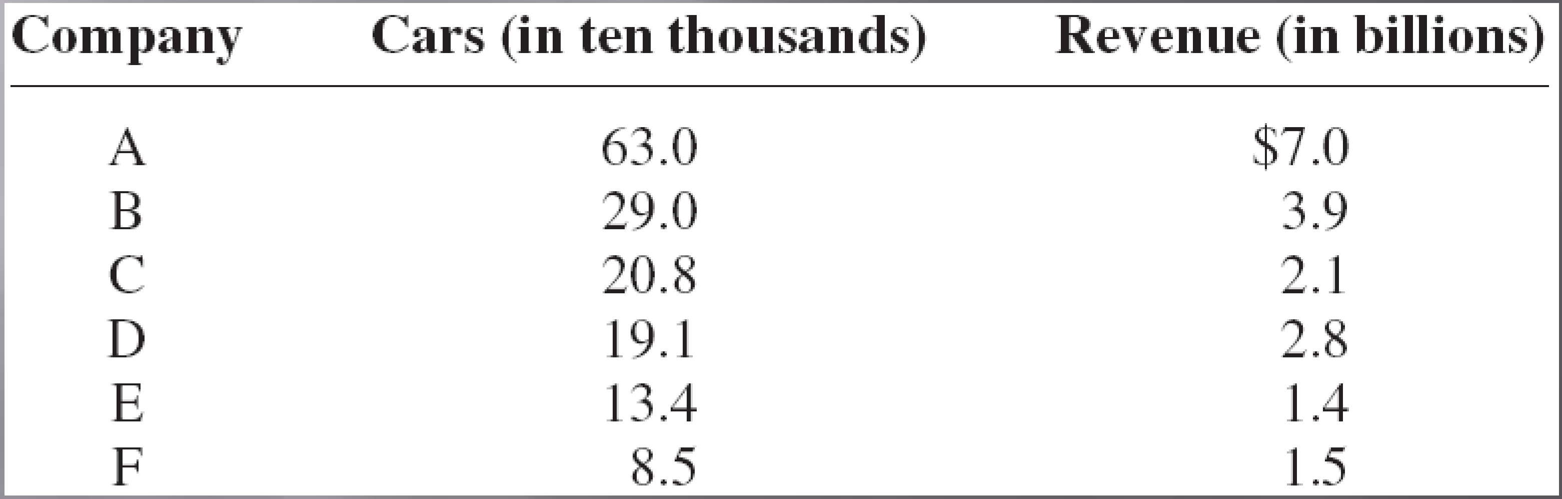
In a multiple relationship, there are two or more independent variables that are used to predict one dependent variable.

10.1 Scatter Plots and Correlation

- 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.

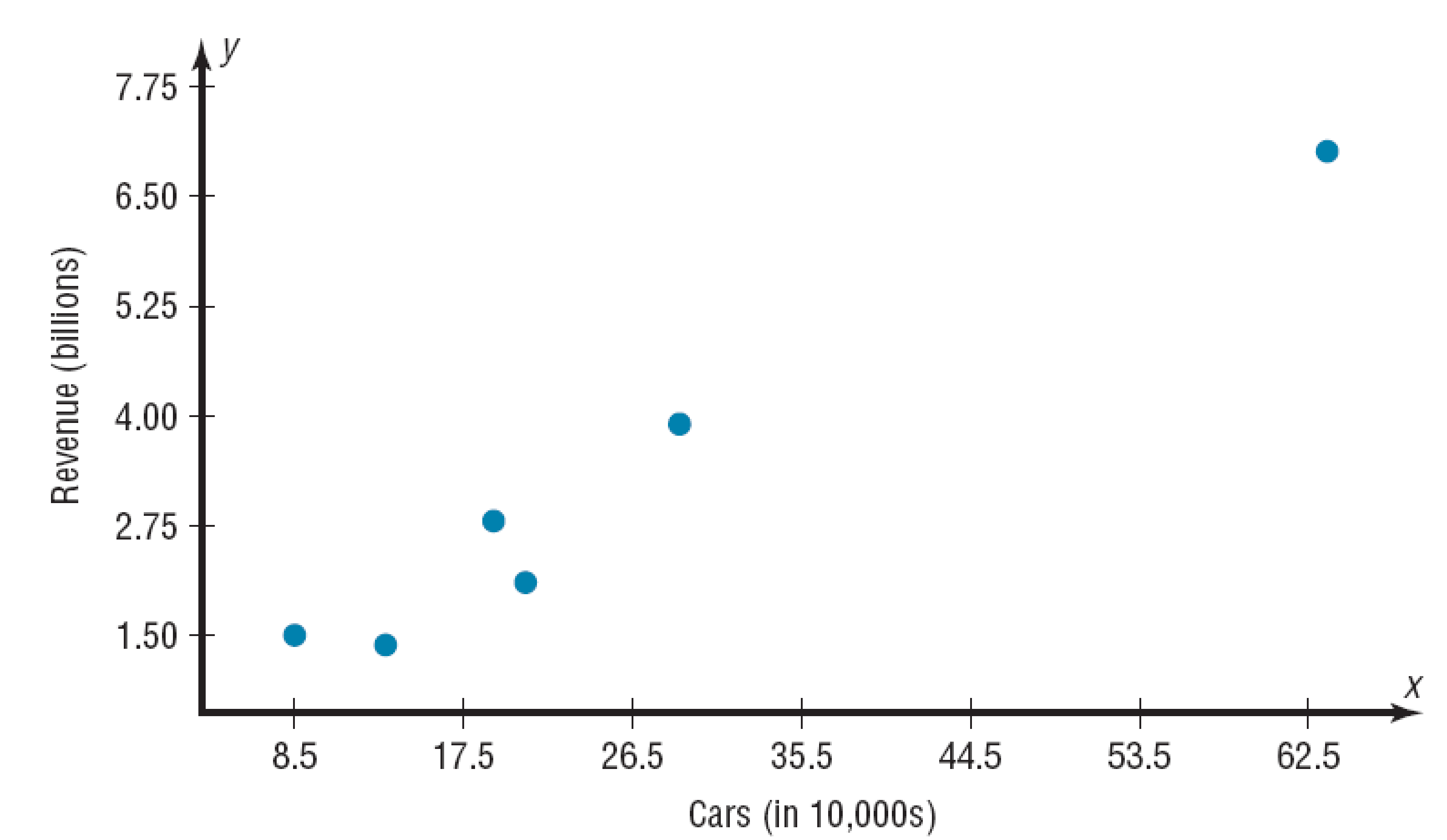
Example 10-1:

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

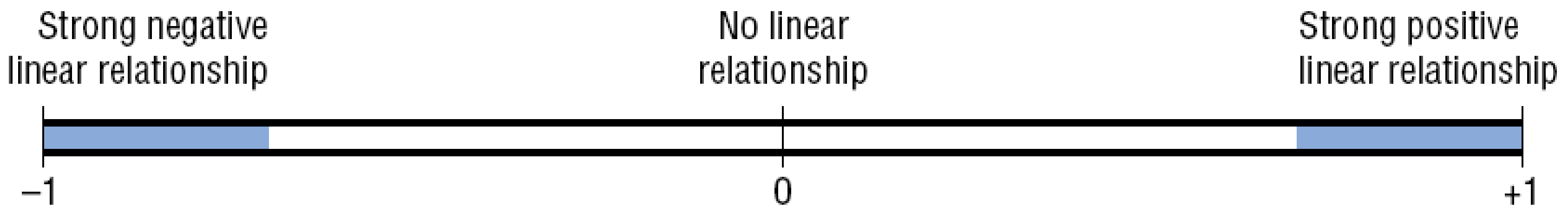
Step 1: Draw and label the *x* and *y* axes.
Step 2: Plot each point on the graph.


y
7.75
6.50
5.25
4.00
2.75
1.50
Revenue (billions)
x
8.5
17.5
26.5
35.5
44.5
53.5
62.5
Cars (in 10,000s)

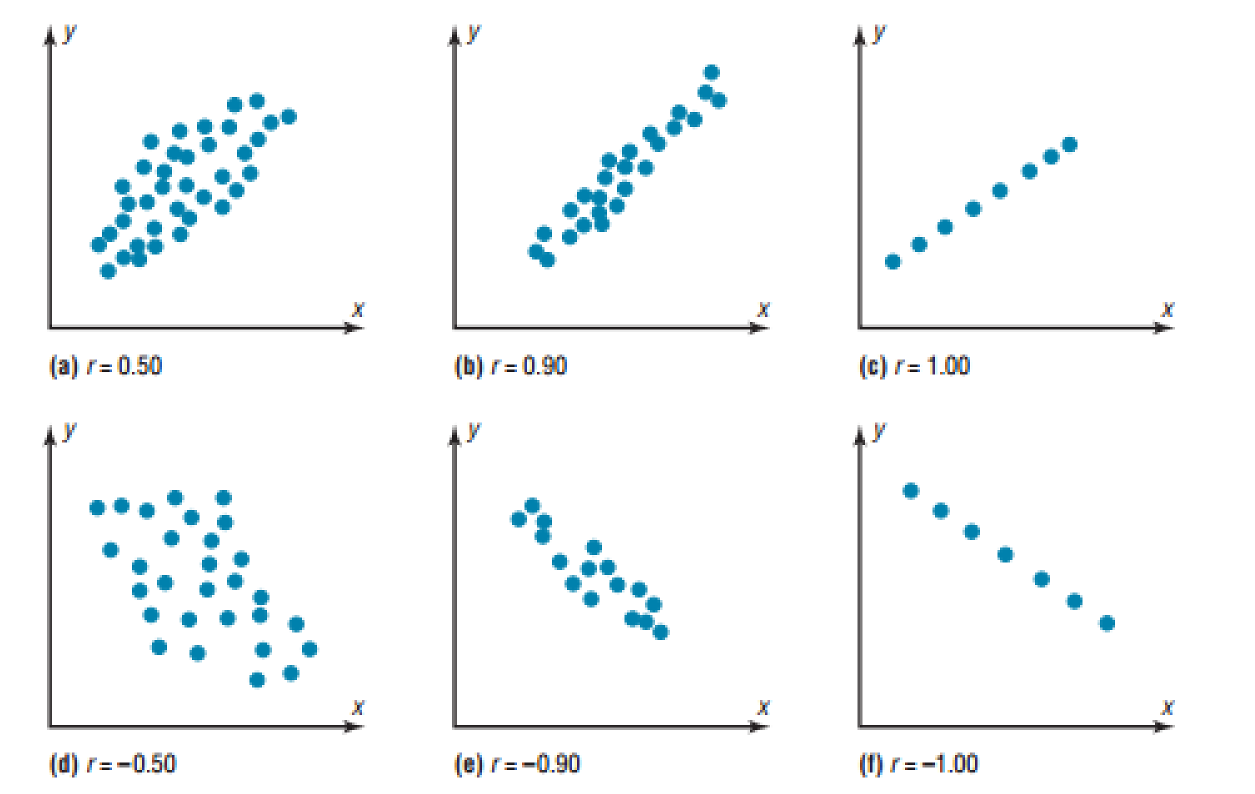
Correlation:

- The **correlation coefficient** computed from the sample data measures the strength and direction of a linear 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)**.
- The symbol for the sample correlation coefficient is r. The symbol for the population correlation coefficient is ρ.

- 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.


Strong negative linear relationship
No linear relationship
Strong positive linear relationship
–1
0
+1

RELATIONSHIP BETWEEN THE CORRELATION COEFFICIENT AND THE SCATTER PLOT



(a) $r = 0.50$ (b) $r = 0.90$ (c) $r = 1.00$

(d) $r = -0.50$ (e) $r = -0.90$ (f) $r = -1.00$

Correlation Coefficient

The formula for the correlation coefficient is:

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

where *n* is the number of data pairs.

Example 10-4:

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

$\Sigma x = 153.8$, $\Sigma y = 18.7$, $\Sigma xy = 682.77$, $\Sigma x^2 = 5859.26$, $\Sigma y^2 = 80.67$, $n = 6$

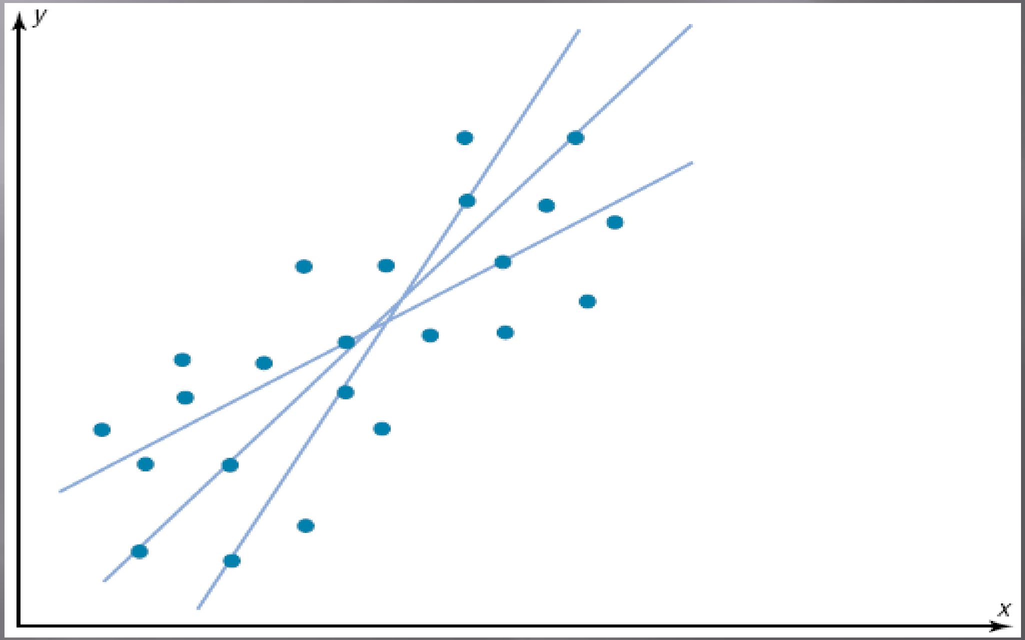
$$r = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left[n\left(\sum x^2\right) - \left(\sum x\right)^2\right]\left[n\left(\sum y^2\right) - \left(\sum y\right)^2\right]}}$$

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

$$r = 0.982 \text{ (strong positive relationship)}$$

10.2 Regression

- 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.



Regression Line: $y' = a + bx$

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

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

where

$a = y'$ intercept

b = the slope of the line.

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.

$\Sigma x = 153.8$, $\Sigma y = 18.7$, $\Sigma xy = 682.77$, $\Sigma x^2 = 5859.26$, $\Sigma y^2 = 80.67$
, $n = 6$

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

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

$$y' = a + bx \quad \rightarrow \quad y' = 0.396 + 0.106x$$

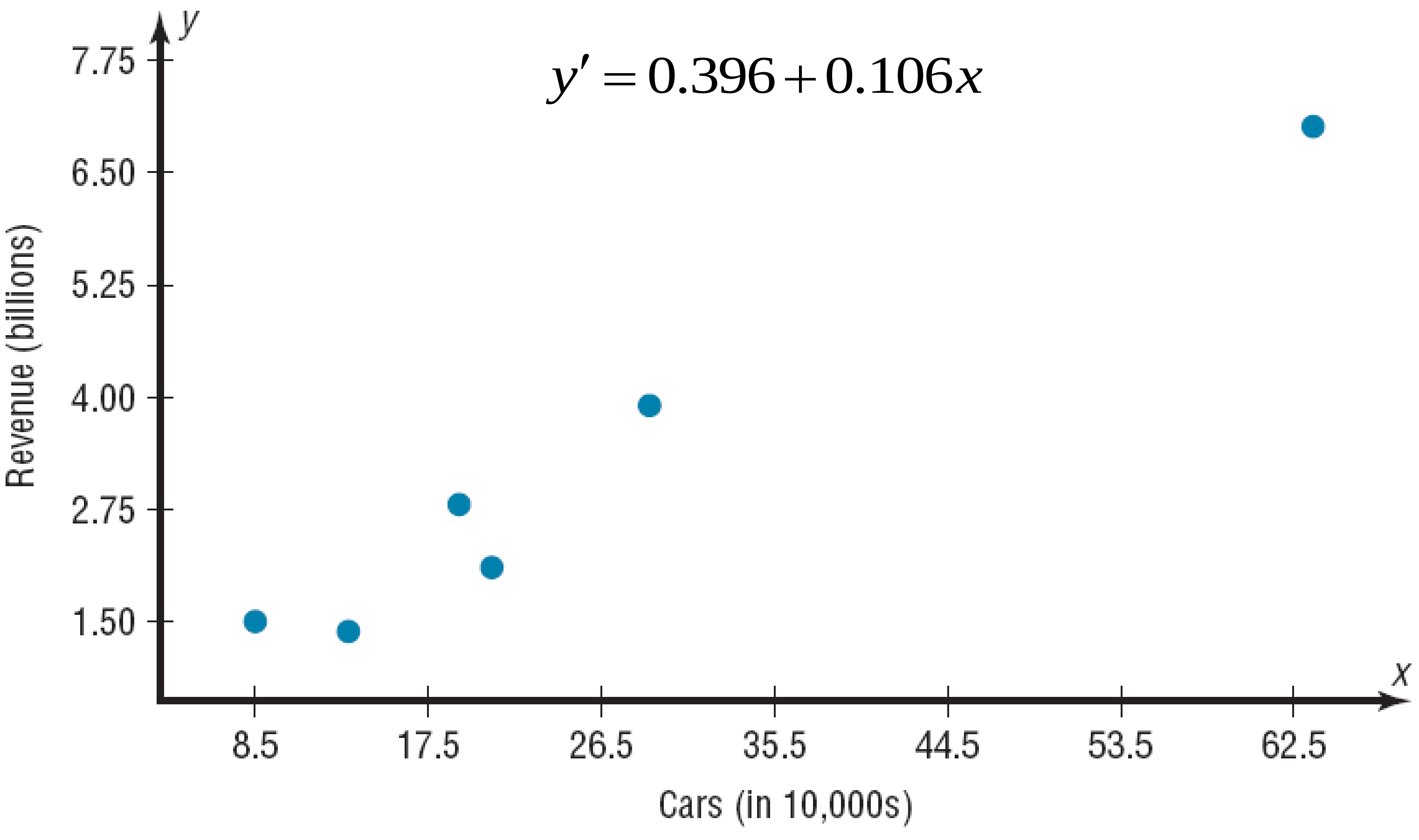
Find two points to sketch the graph of the regression line.

Use any *x* values between 10 and 60. For example, let *x* equal 15 and 40. Substitute in the equation and find the corresponding *y* value.

$$\begin{aligned} y' &= 0.396 + 0.106x \\ &= 0.396 + 0.106(15) \\ &= 1.986 \end{aligned} \qquad \begin{aligned} y' &= 0.396 + 0.106x \\ &= 0.396 + 0.106(40) \\ &= 4.636 \end{aligned}$$

Plot (15,1.986) and (40,4.636), and sketch the resulting line.

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



$$y' = 0.396 + 0.106x$$

Example 10-11:

Use the equation of the regression line to predict the income of a car rental agency that has 200,000 automobiles.

x = 20 corresponds to 200,000 automobiles.

$$\begin{aligned} y' &= 0.396 + 0.106x \\ &= 0.396 + 0.106(20) \\ &= 2.516 \end{aligned}$$

Hence, when a rental agency has 200,000 automobiles, its revenue will be approximately $2.516 billion.

Marginal change

- **The magnitude of the change in one Variable when the other variable change exactly 1 unit is called Marginal change.**
- **The value of slope *b* of the regression line equation represents the marginal change**.

13-6 The Spearman Rank Correlation Coefficient and the Runs Test

- One assumption for testing the hypothesis that $\rho = 0$ for the Pearson coefficient is that the populations from which the samples are obtained are normally distributed.
- If this requirement cannot be met, the nonparametric equivalent, called the **Spearman rank correlation coefficient** (denoted by r_s), can be used when the data are ranked.

Formula for the Spearman Rank Correlation Coefficient

$$r_s = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)}$$

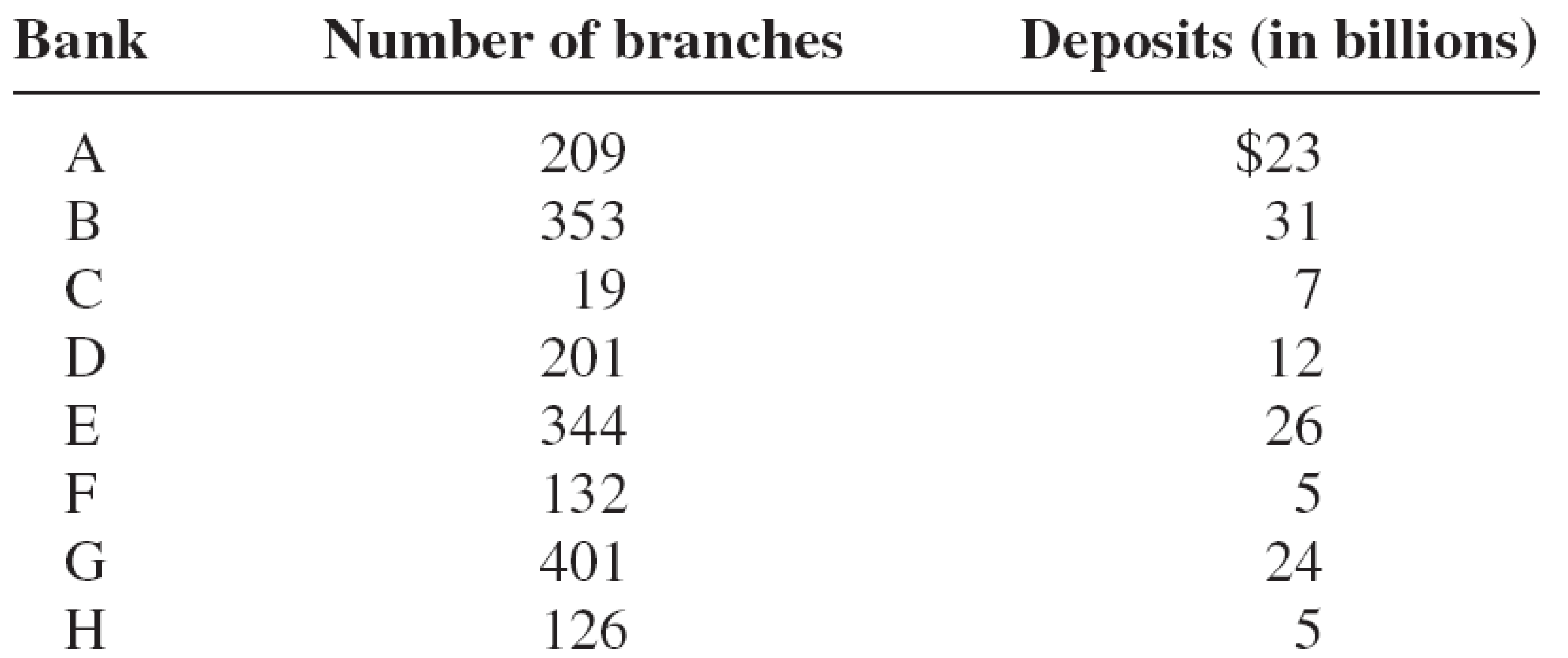
where
$d =$ difference in ranks
$n =$ number of data pairs

Example 13-7: Bank Branches/Deposits

A researcher wishes to see if there is a relationship between the number of branches a bank has and the total number of deposits (in billions of dollars) the bank receives. A sample of eight regional banks is selected (see next slide) , and the number of branches and the amount of deposits are shown in the table.

At $\alpha = 0.05$ is there a significant linear correlation between the number of branches and the amount of the deposits?

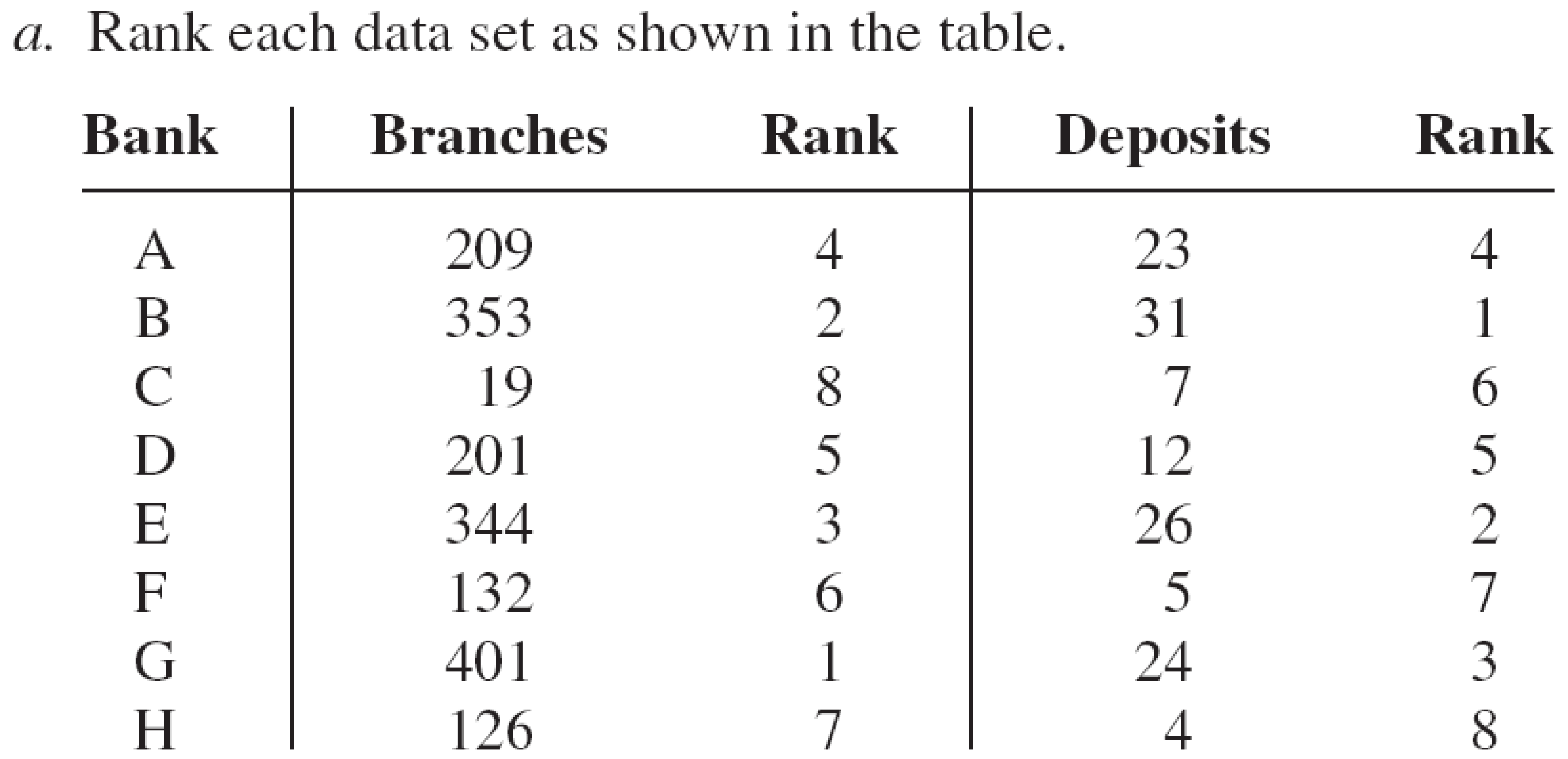
Example 13-7: Bank Branches/Deposits



Bank	Number of branches	Deposits (in billions)
A	209	$23
B	353	31
C	19	7
D	201	12
E	344	26
F	132	5
G	401	24
H	126	5

Example 13-7: Bank Branches/Deposits

Step 3: Find the test value.

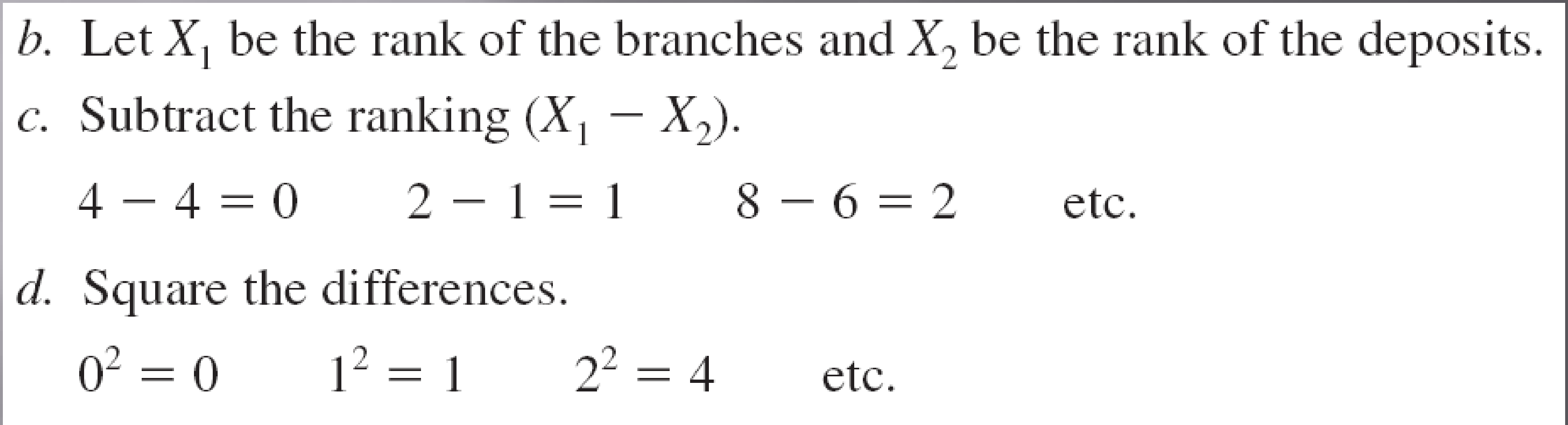


a. Rank each data set as shown in the table.

Bank	Branches	Rank	Deposits	Rank
A	209	4	23	4
B	353	2	31	1
C	19	8	7	6
D	201	5	12	5
E	344	3	26	2
F	132	6	5	7
G	401	1	24	3
H	126	7	4	8

Example 13-7: Bank Branches/Deposits

Step 3: Find the test value.



b. Let X_1 be the rank of the branches and X_2 be the rank of the deposits.

c. Subtract the ranking $(X_1 - X_2)$.

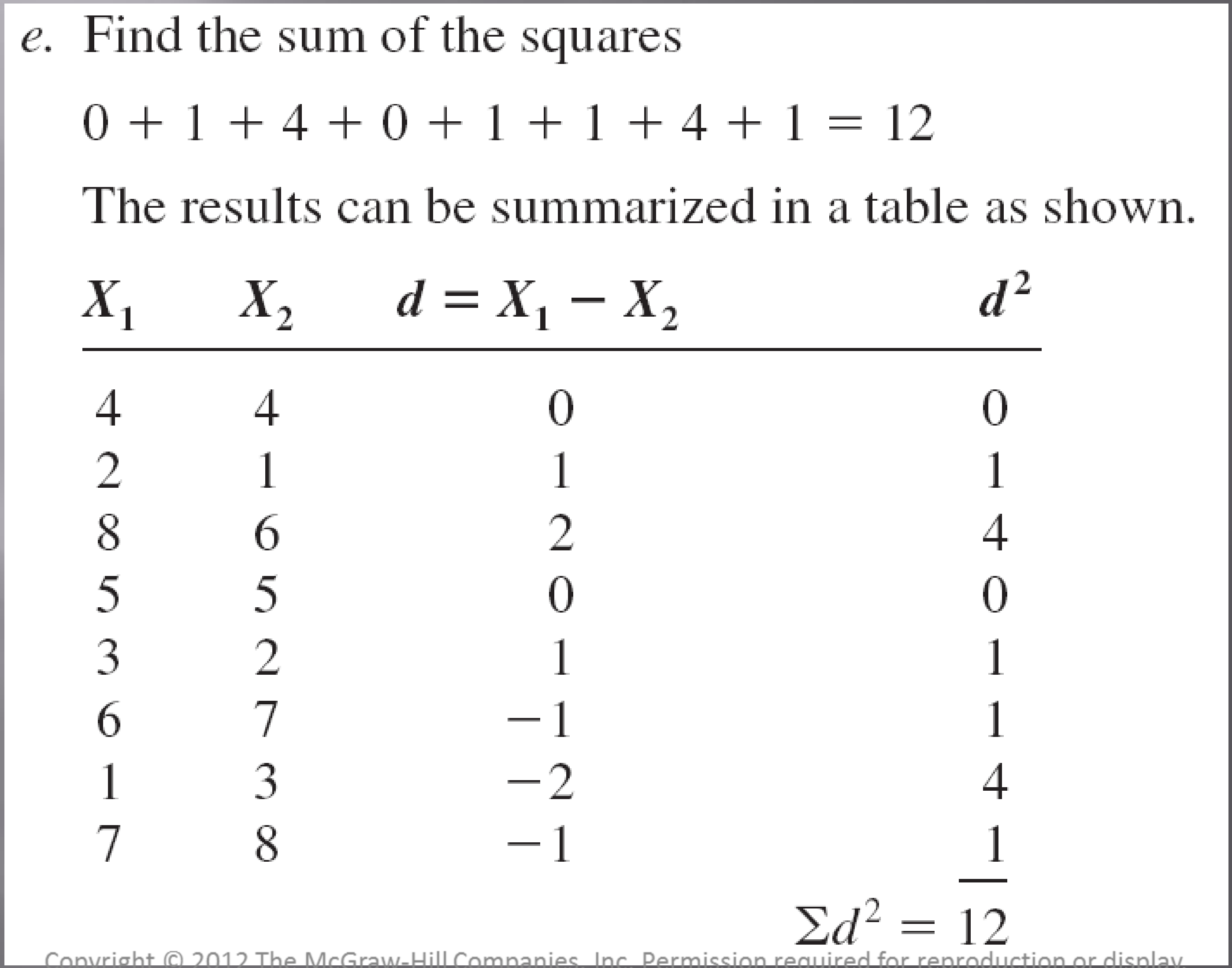
$4 - 4 = 0 \qquad 2 - 1 = 1 \qquad 8 - 6 = 2 \qquad$ etc.

d. Square the differences.

$0^2 = 0 \qquad 1^2 = 1 \qquad 2^2 = 4 \qquad$ etc.

Example 13-7: Bank Branches/Deposits

Step 3: Find the test value.



e. Find the sum of the squares

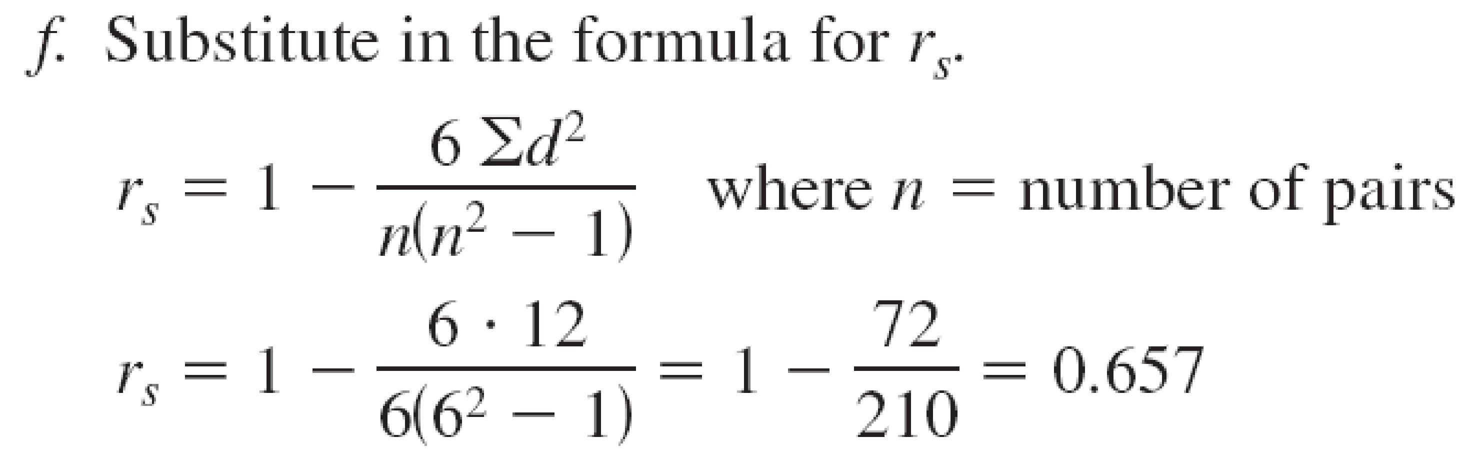
$0 + 1 + 4 + 0 + 1 + 1 + 4 + 1 = 12$

The results can be summarized in a table as shown.

X_1	X_2	$d = X_1 - X_2$	d^2
4	4	0	0
2	1	1	1
8	6	2	4
5	5	0	0
3	2	1	1
6	7	−1	1
1	3	−2	4
7	8	−1	1
			$\Sigma d^2 = 12$

Example 13-7: Bank Branches/Deposits

Step 3: Find the test value.



f. Substitute in the formula for r_s.

$$r_s = 1 - \frac{6\,\Sigma d^2}{n(n^2 - 1)} \quad \text{where } n = \text{number of pairs}$$

$$r_s = 1 - \frac{6 \cdot 12}{6(6^2 - 1)} = 1 - \frac{72}{210} = 0.657$$