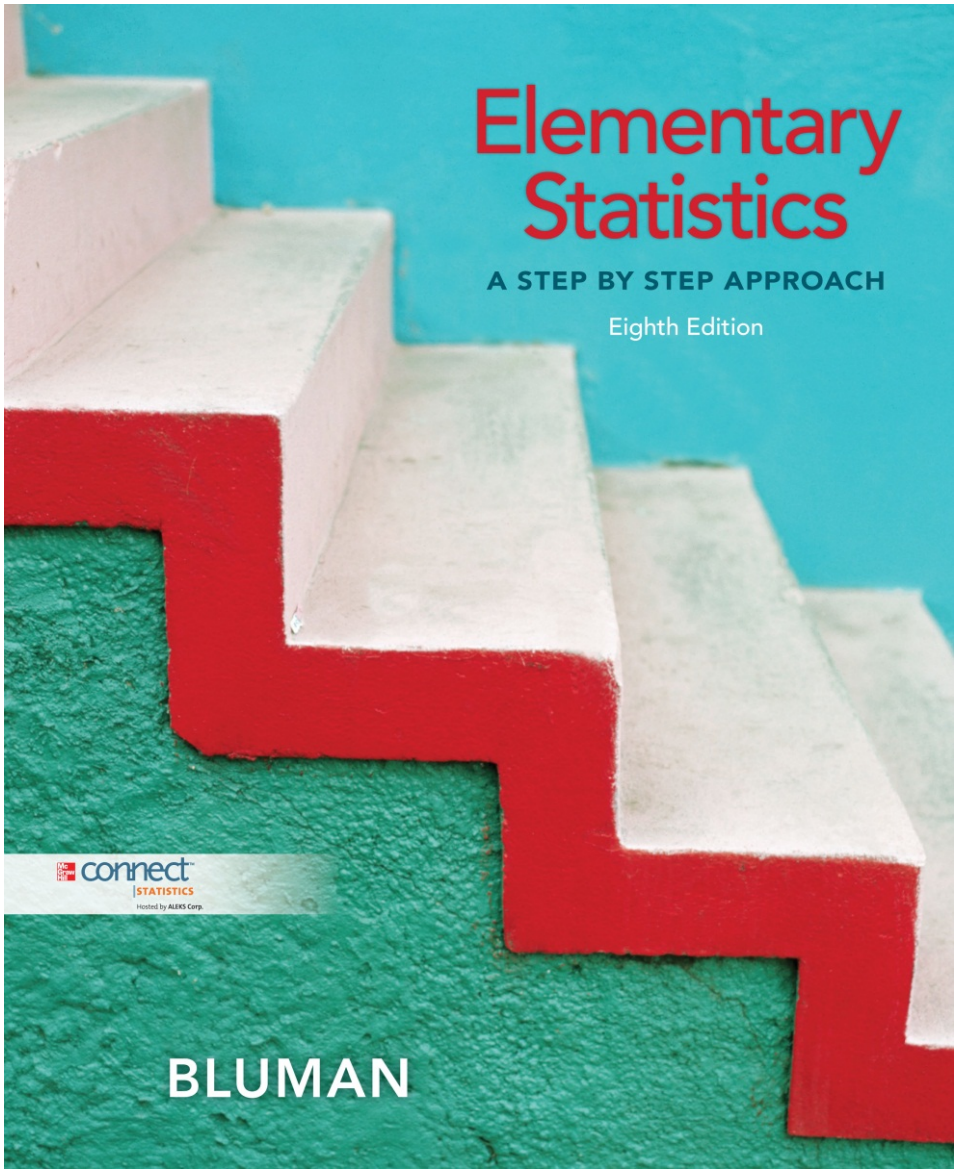


Chapter 10
Correlation and Regression

CHAPTER 10

Correlation and Regression

Outline

10-1 Scatter Plots and Correlation

10-2 Regression

10-3 Coefficient of Determination and Standard Error of the Estimate

10-4 Multiple Regression (Optional)

CHAPTER 10

Correlation and Regression

Objectives

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.

Introduction

- In addition to hypothesis testing and confidence intervals, inferential statistics involves determining whether a relationship between two or more numerical or quantitative variables exists.

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.

Introduction

- 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?

Introduction

1. *Are two or more variables related?*
2. *If so, what is the strength of the relationship?*

To answer these two questions, statisticians use the **correlation coefficient**, 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.

Introduction

3. 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** (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.

Introduction

4. What kind of predictions can be made from the relationship?

Predictions are made daily in all areas. Examples include weather forecasting, stock market analyses, sales predictions, crop predictions, gasoline price 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.

10.1 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.

Chapter 10
Correlation and Regression

Section 10-1

Example 10-1

Page #536

Example 10-1: Car Rental Companies

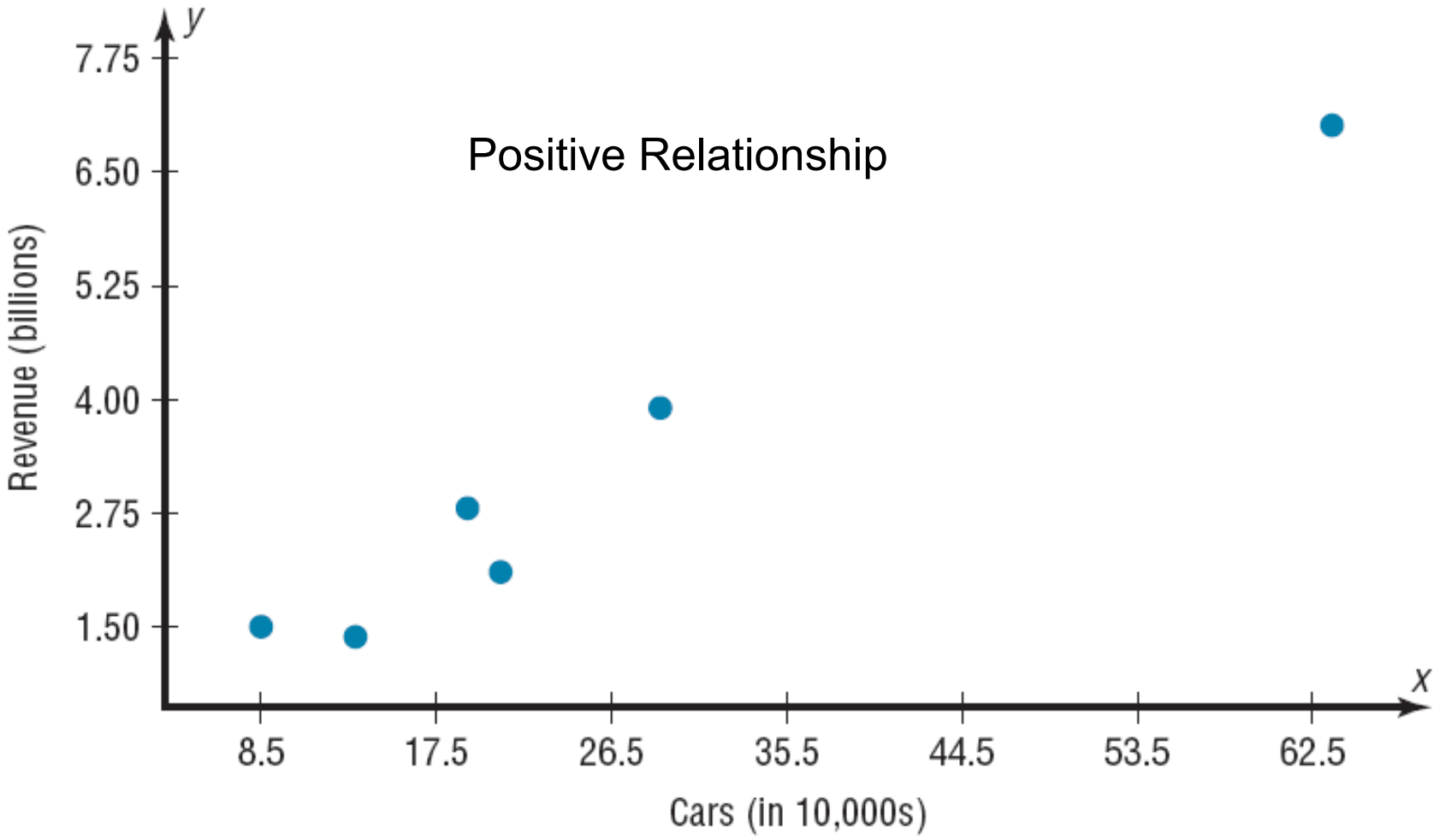
Construct a scatter plot for the data shown for car rental companies in the United States for a recent year.

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

Step 1: Draw and label the *x* and *y* axes.

Step 2: Plot each point on the graph.

Example 10-1: Car Rental Companies



Chapter 10
Correlation and Regression

Section 10-1

Example 10-2

Page #537

Example 10-2: Absences/Final Grades

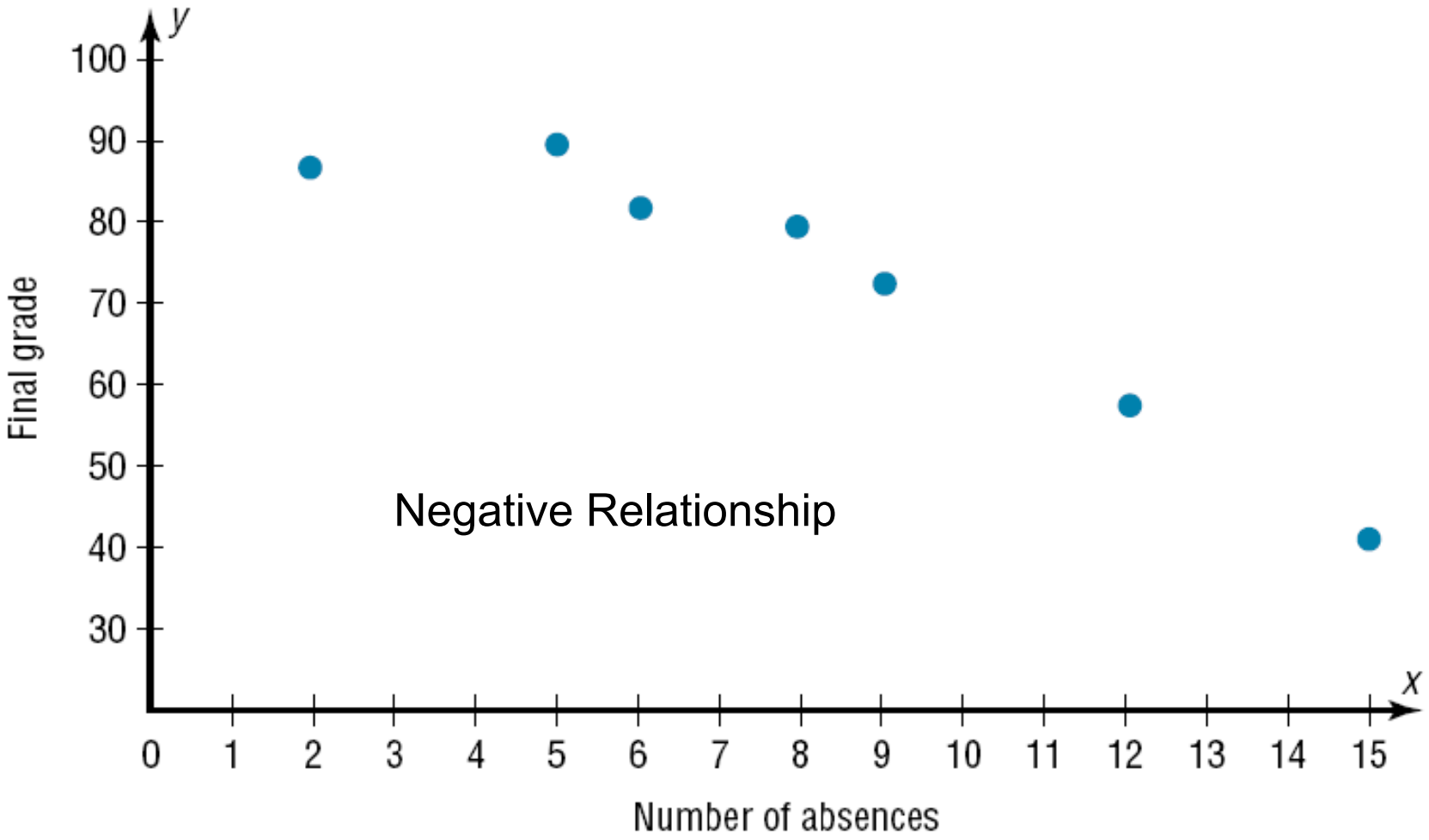
Construct a scatter plot for the data obtained in a study on the number of absences and the final grades of seven randomly selected students from a statistics class.

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

Step 1: Draw and label the x and y axes.

Step 2: Plot each point on the graph.

Example 10-2: Absences/Final Grades



Chapter 10
Correlation and Regression

Section 10-1

Example 10-3

Page #538

Example 10-3: Age and Wealth

A researcher wishes to see if there is a relationship between the ages and net worth of the wealthiest people in America. The data for a specific year are shown.

Person	Age x	Net wealth y ($ billions)
A	73	16
B	65	26
C	53	50
D	54	21.5
E	79	40
F	69	16
G	61	19.6
H	65	19

Step 1: Draw and label the x and y axes.

Step 2: Plot each point on the graph.

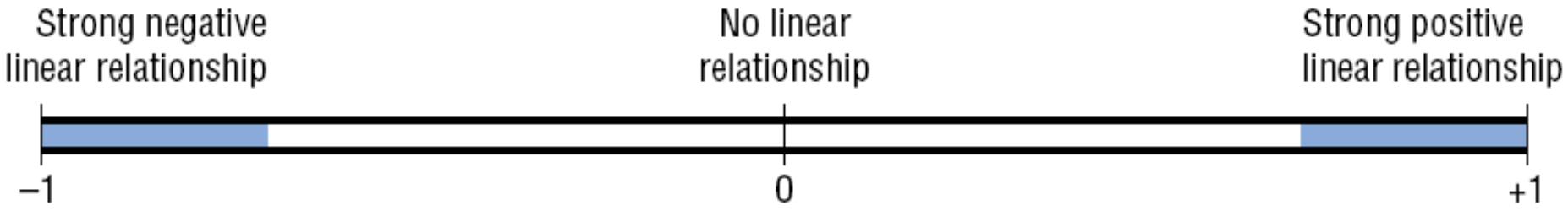
Example 10-3: Age and Wealth

Correlation

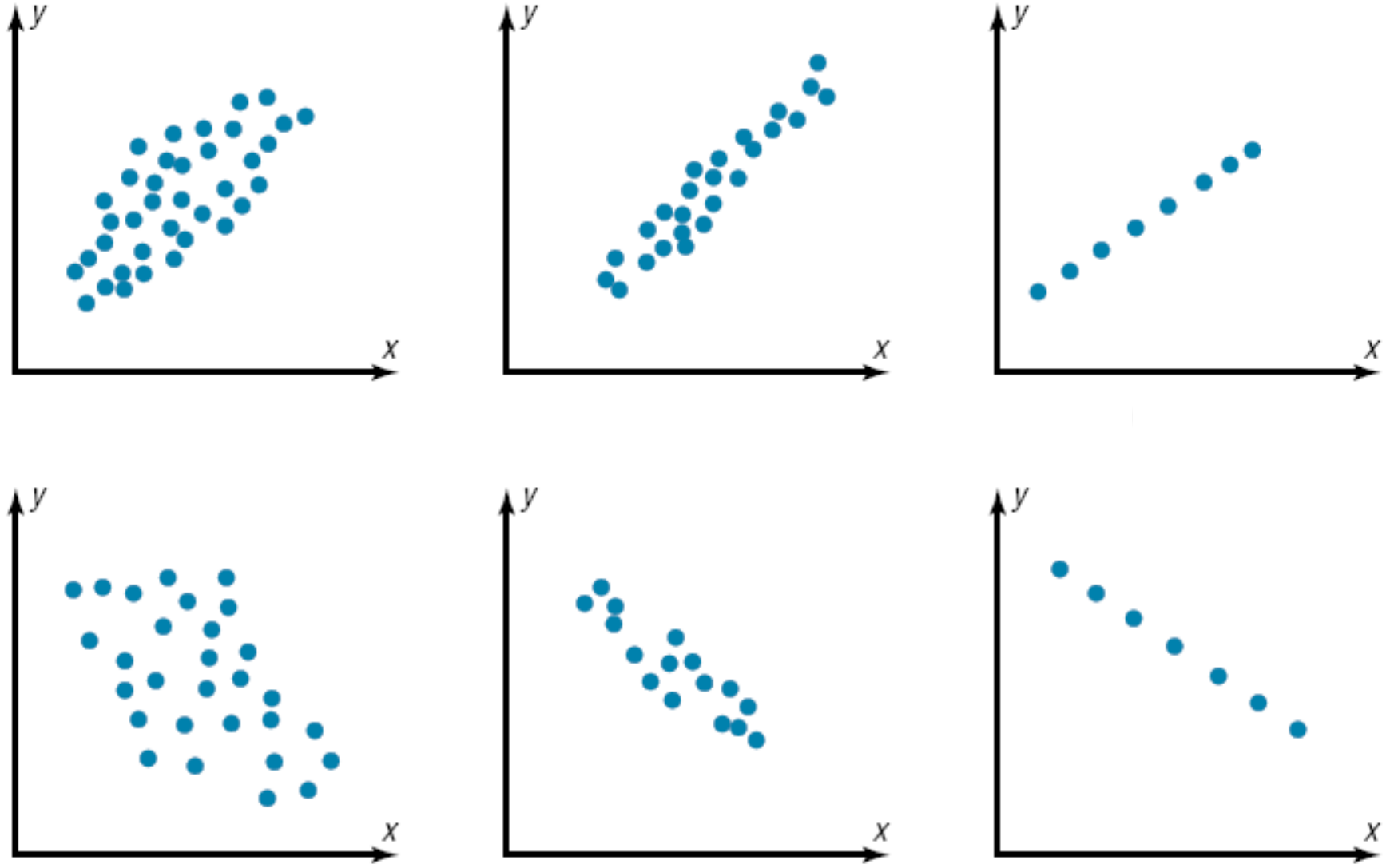
- The **correlation coefficient** computed from the sample data measures the strength and direction of a linear relationship between two variables.
- There are several types of correlation coefficients. The one explained in this section is called the **Pearson product moment correlation coefficient (PPMC)**.
- The symbol for the sample correlation coefficient is *r*. The symbol for the population correlation coefficient is ρ.

Correlation

- 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.



Correlation



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.

Rounding Rule: Round to three decimal places.

Chapter 10
Correlation and Regression

Section 10-1

Example 10-4

Page #540

Example 10-4: Car Rental Companies

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

Company	Cars x (in 10,000s)	Income y (in billions)	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

Example 10-4: Car Rental Companies

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

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

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

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

$r = 0.982$ (strong positive relationship)

Chapter 10
Correlation and Regression

Section 10-1

Example 10-5

Page #541

Example 10-5: Absences/Final Grades

Compute the correlation coefficient for the data in Example 10–2.

Student	Number of absences, x	Final Grade y (pct.)	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

Example 10-5: Absences/Final Grades

Compute the correlation coefficient for the data in Example 10–2.

$\Sigma x = 57,\ \Sigma y = 511,\ \Sigma xy = 3745,\ \Sigma x^2 = 579,$
$\Sigma y^2 = 38{,}993,\ n = 7$

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

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

$r = -0.944$ (strong negative relationship)

Chapter 10
Correlation and Regression

Section 10-1

Example 10-6

Page #542

Example 10-6: Age and Wealth

Compute the value of the correlation coefficient for the data given in Example 10–3 for the age and wealth of the richest persons in the United States.

Person	Age x	Net wealth y	xy	x^2	y^2
A	73	16	1,168	5,329	256
B	65	26	1,690	4,225	676
C	53	50	2,650	2,809	2,500
D	54	21.5	1,161	2,916	462.25
E	79	40	3,160	6,241	1,600
F	69	16	1,104	4,761	256
G	61	19.6	1,195.6	3,721	384.16
H	65	19	1,235	4,225	361
	$\Sigma x = 519$	$\Sigma y = 208.1$	$\Sigma xy = 13{,}363.6$	$\Sigma x^2 = 34{,}227$	$\Sigma y^2 = 6{,}495.41$

Example 10-6: Age and Wealth

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

$$= \frac{8(13{,}363.6) - (519)(208.1)}{\sqrt{[8(34{,}227) - (519)^2][8(6495.41) - (208.1)^2]}}$$

$$= \frac{-1095.1}{\sqrt{(4455)(8657.67)}}$$

$$= \frac{-1095.1}{6210.469}$$

$$= -0.176$$

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

Hypothesis Testing

- In hypothesis testing, one of the following 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.

t Test for the Correlation Coefficient

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

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

Chapter 10
Correlation and Regression

Section 10-1

Example 10-7

Page #544

Example 10-7: Car Rental Companies

Test the significance of the correlation coefficient found in Example 10–4. Use $\alpha = 0.05$ and $r = 0.982$.

Step 1: State the hypotheses.

$H_0: \rho = 0$ and $H_1: \rho \neq 0$

Step 2: Find the critical value.

Since $\alpha = 0.05$ and there are 6 – 2 = 4 degrees of freedom, the critical values obtained from Table F are ±2.776.

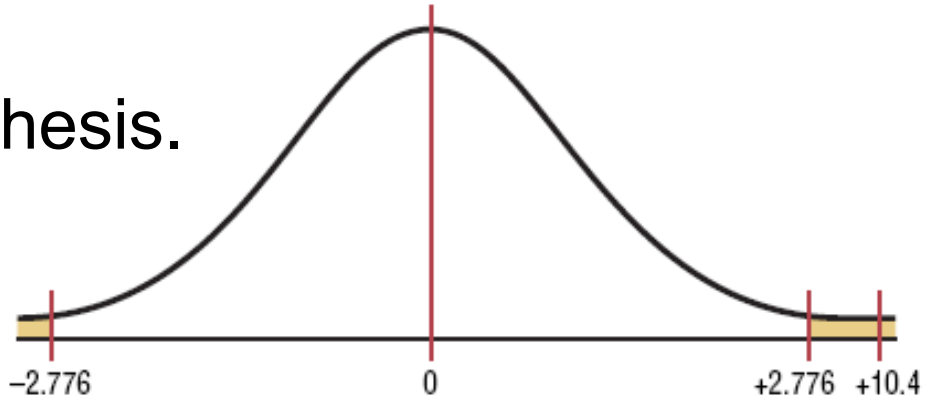
Example 10-7: Car Rental Companies

Step 3: Compute the test value.

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 0.982\sqrt{\frac{6-2}{1-(0.982)^2}} = 10.4$$

Step 4: Make the decision.

Reject the null hypothesis.



Step 5: Summarize the results.

There is a significant relationship between the number of cars a rental agency owns and its annual income.

Chapter 10
Correlation and Regression

Section 10-1

Example 10-8

Page #545

Example 10-8: Age and Wealth

Using Table I, test the significance at a 0.01 of the correlation coefficient $r = -0.176$, obtained in Example 10–6.

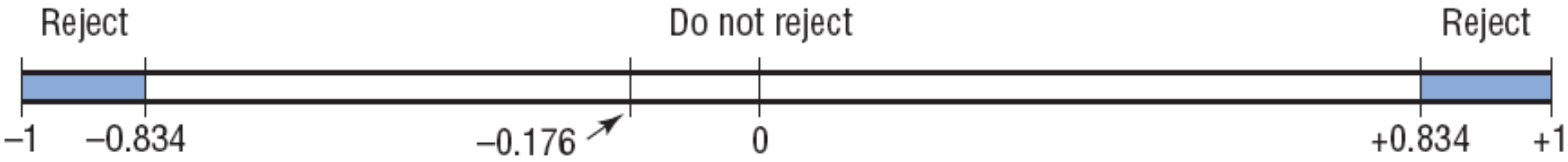
$$H_0: \rho = 0 \text{ and } H_1: \rho \neq 0$$

There are 8 – 2 = 6 degrees of freedom. The value in Table I when $\alpha = 0.01$ is 0.834.

For a significant relationship, r must be greater than 0.834 or less than –0.834. Since $r = -0.176$, do not reject the null hypothesis.

Example 10-8: Age and Wealth

Hence, there is not enough evidence to say that there is a significant linear relationship between the variables.



Possible Relationships Between Variables

When the null hypothesis has been rejected for a specific a 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*.
2. There is a *reverse cause-and-effect* relationship between the variables. That is, *y* causes *x*.
3. The relationship between the variables may be *caused by a third variable*.
4. There may be a *complexity of interrelationships* among many variables.
5. The relationship may be *coincidental*.

Possible Relationships Between Variables

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
   - ☐ heat causes ice to melt

Possible Relationships Between Variables

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.

Possible Relationships Between Variables

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.

Possible Relationships Between Variables

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.

Possible Relationships Between Variables

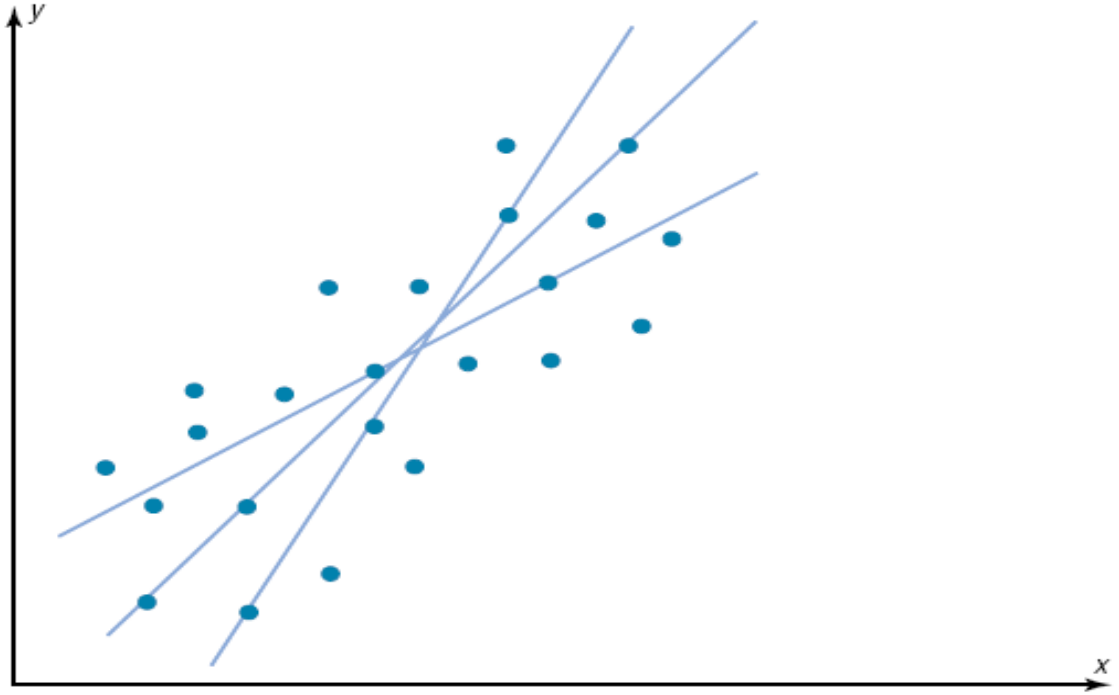
5. The relationship may be *coincidental.*

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.

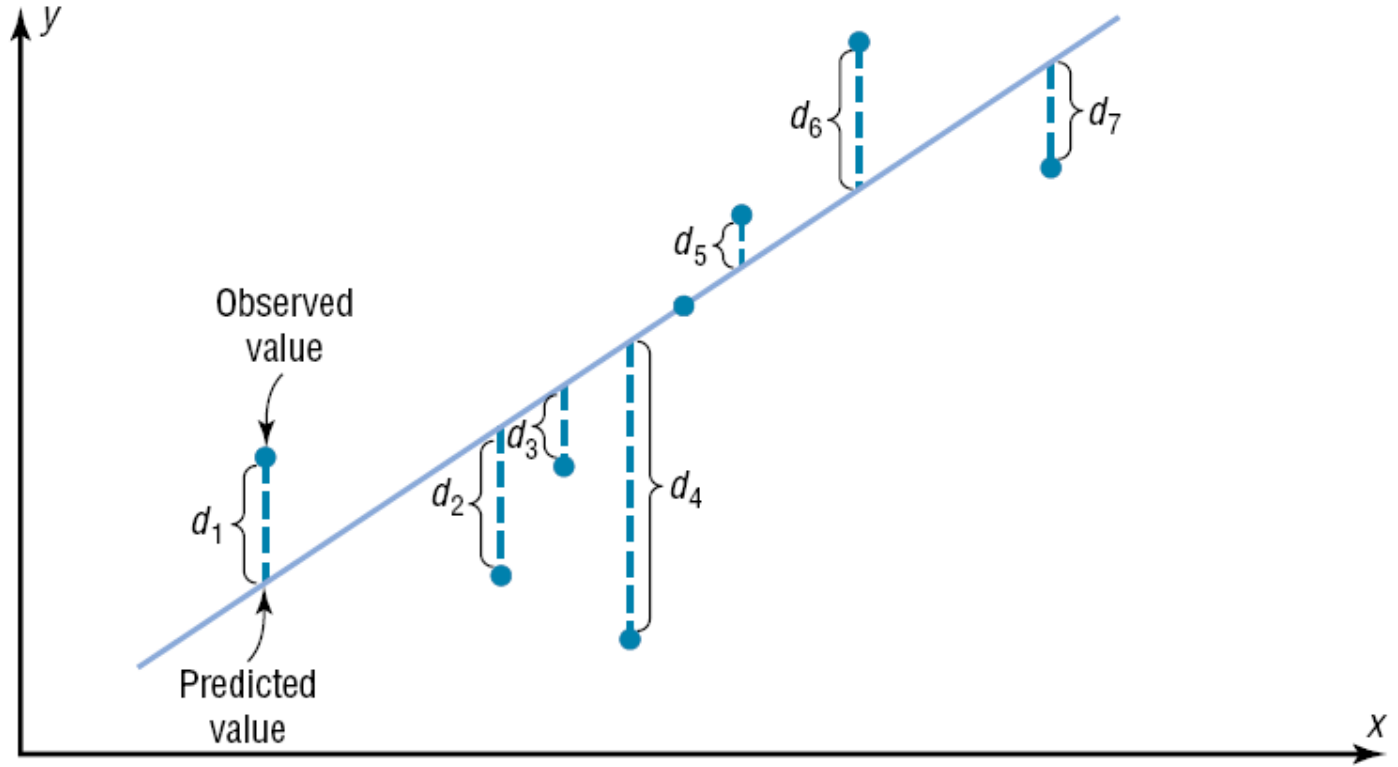
10.2 Regression

- 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

- **Best fit** means that the sum of the squares of the vertical distance from each point to the line is at a minimum.



Regression Line $y' = a + bx$

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

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

where

$a = y'$ intercept

$b =$ the slope of the line.

Chapter 10
Correlation and Regression

Section 10-2

Example 10-9

Page #553

Example 10-9: Car Rental Companies

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

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

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

$$= \frac{(18.7)(5859.26) - (153.8)(682.77)}{6(5859.26) - (153.8)^2} = 0.396$$

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

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

Example 10-9: Car Rental Companies

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

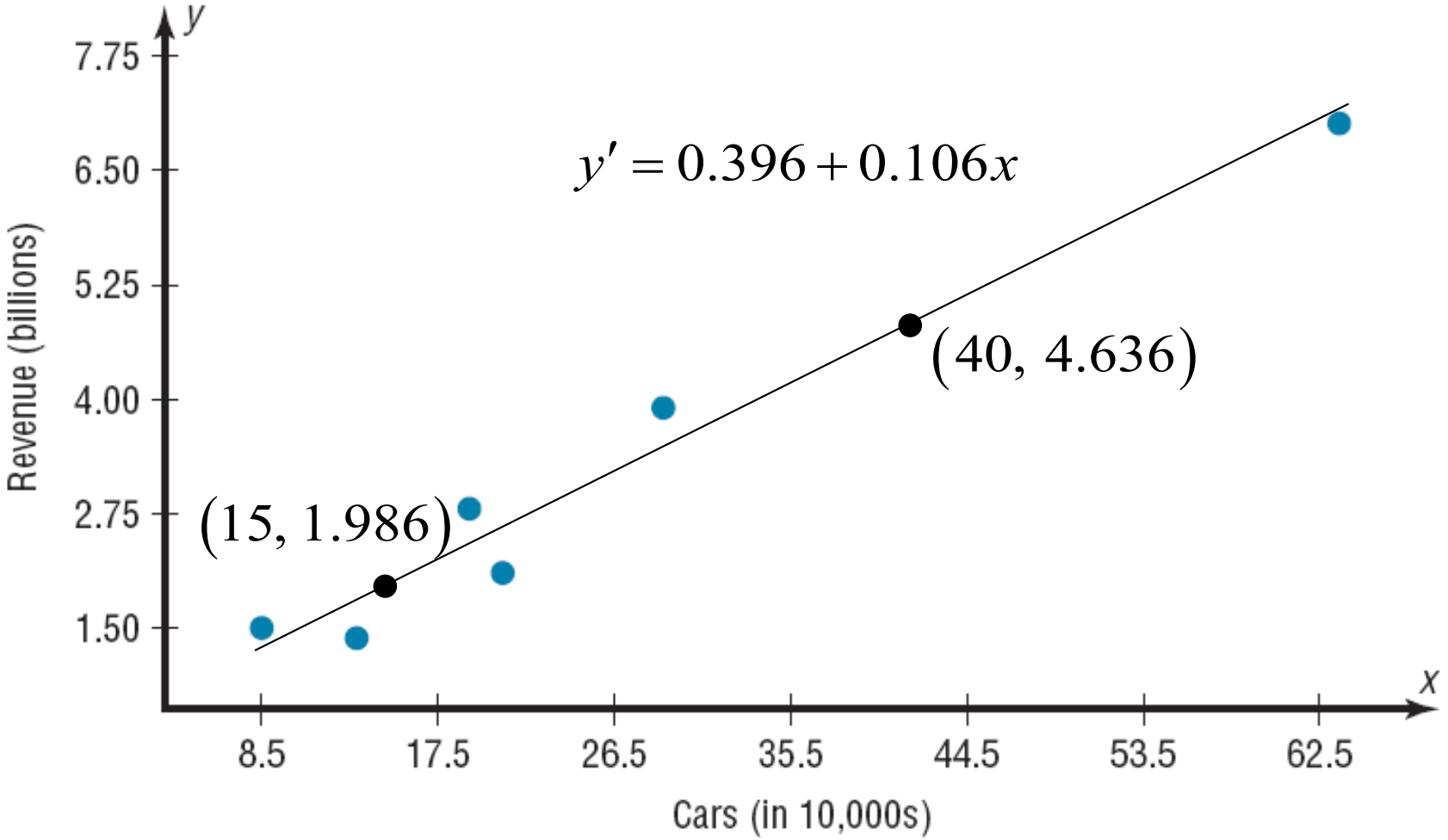
Use any *x* values between 10 and 60. For example, let *x* equal 15 and 40. Substitute in the equation and find the corresponding *y* value.

$$y' = 0.396 + 0.106x \qquad y' = 0.396 + 0.106x$$
$$= 0.396 + 0.106(15) \qquad = 0.396 + 0.106(40)$$
$$= 1.986 \qquad = 4.636$$

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

Example 10-9: Car Rental Companies

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



Chapter 10
Correlation and Regression

Section 10-2

Example 10-11

Page #555

Example 10-11: Car Rental Companies

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.

$$y' = 0.396 + 0.106x$$
$$= 0.396 + 0.106(20)$$
$$= 2.516$$

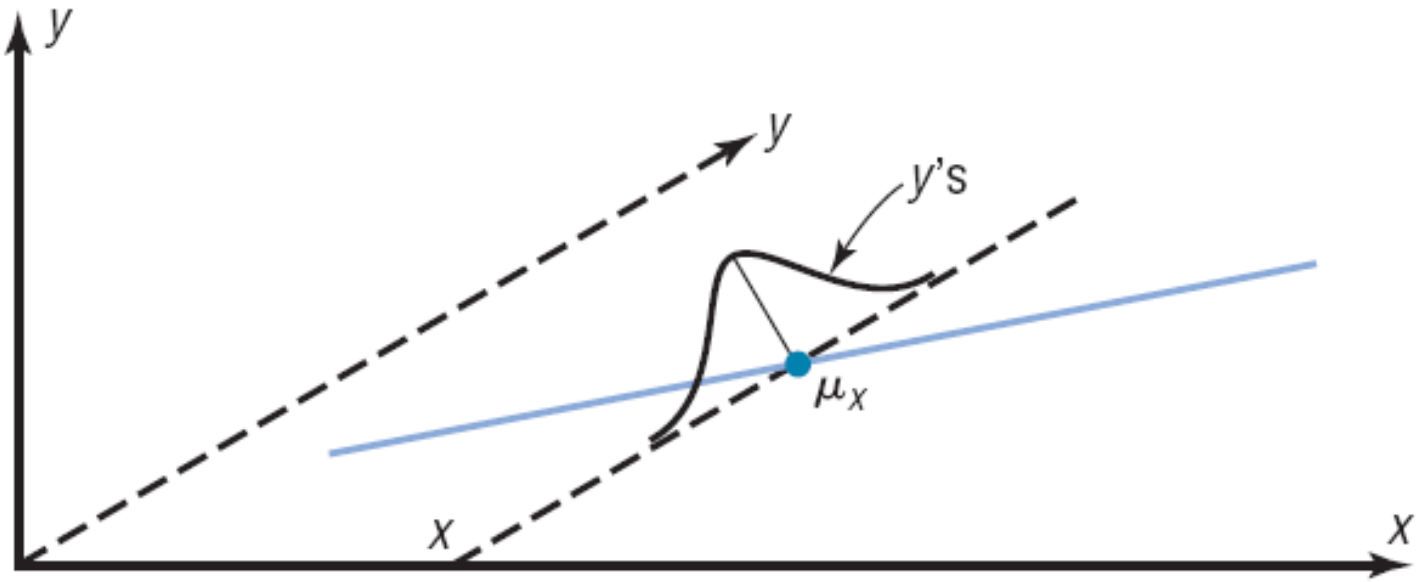
Hence, when a rental agency has 200,000 automobiles, its revenue will be approximately $2.516 billion.

Regression

- 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 valid predictions, the value of the correlation coefficient must be significant.
- When *r* is not significantly different from 0, the best predictor of *y* is the mean of the data values of *y*.

Assumptions for Valid Predictions

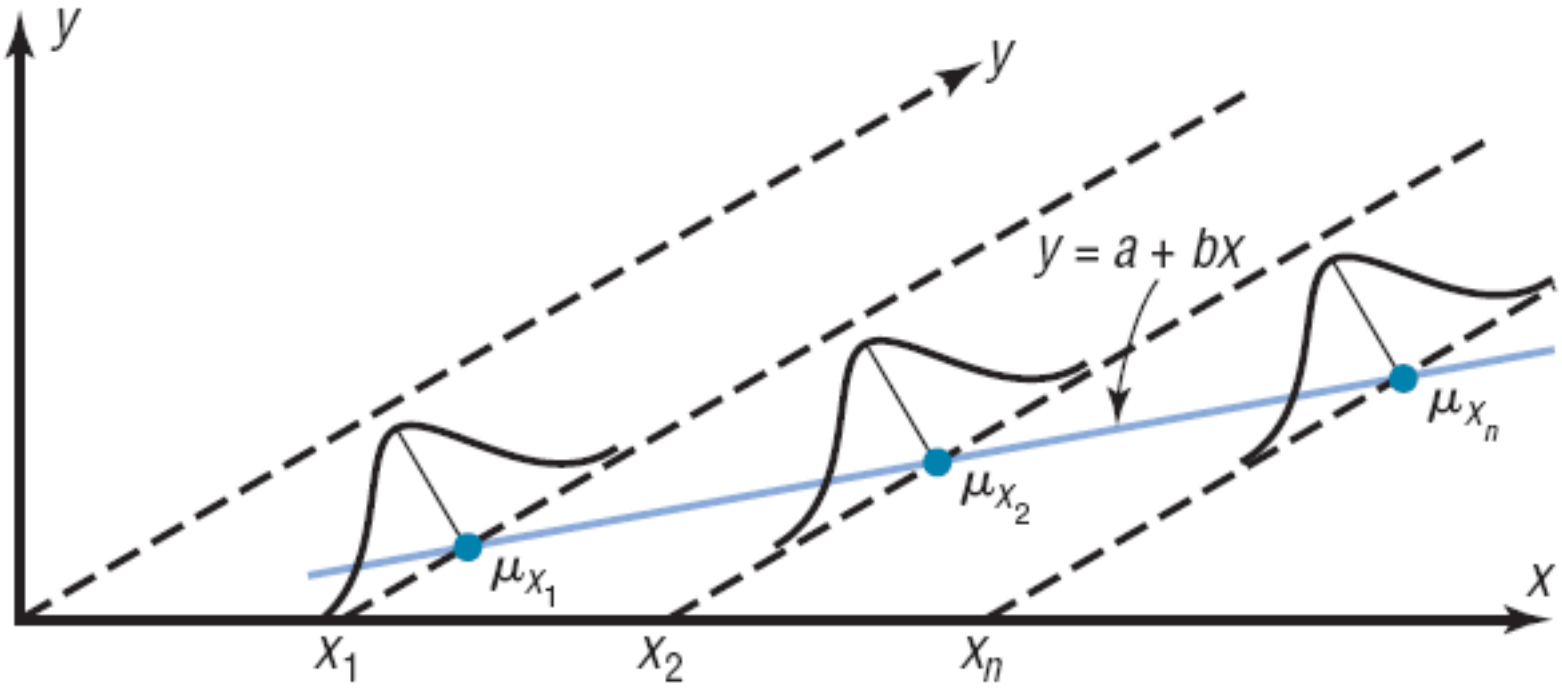
1. For any specific value of the independent variable *x*, the value of the dependent variable *y* must be normally distributed about the regression line.



(a) Dependent variable *y* normally distributed

Assumptions for Valid Predictions

2. The standard deviation of each of the dependent variables must be the same for each value of the independent variable.



(b) $\sigma_1 = \sigma_2 = \cdots = \sigma_n$

Extrapolations (Future Predictions)

- **Extrapolation**, or making predictions beyond the bounds of the data, must be interpreted cautiously.
- 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.

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

Procedure Table

Finding the Correlation Coefficient and the Regression Line Equation

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

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

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

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

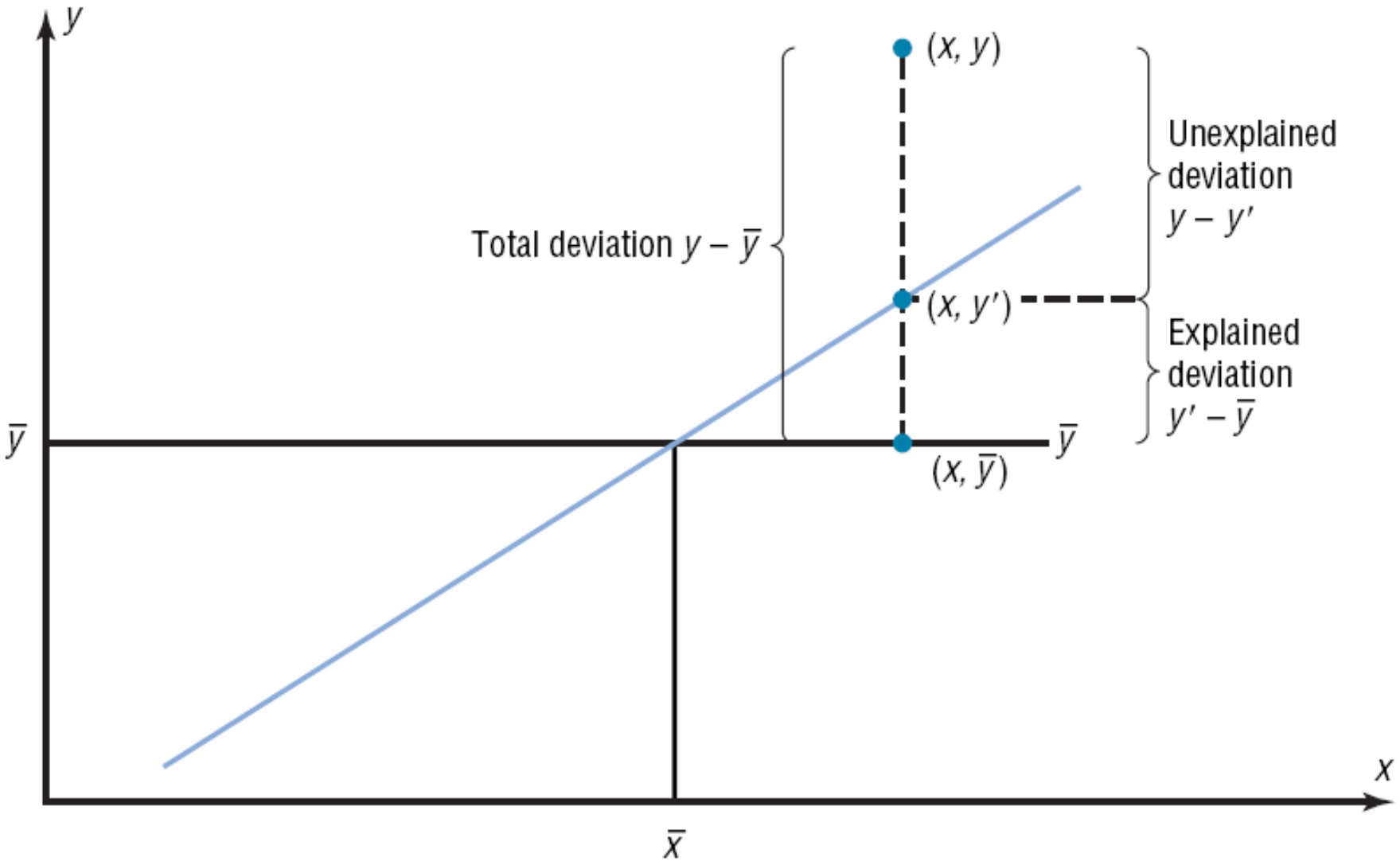
10.3 Coefficient of Determination and Standard Error of the Estimate

- The **total variation** $\sum(y-\bar{y})^2$ 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.

Variation

- The variation obtained from the relationship (i.e., from the predicted y' values) is $\sum(y'-\bar{y})^2$ and is called the **explained variation**.
- Variation due to chance, found by $\sum(y'-y)^2$, is called the **unexplained variation**. This variation cannot be attributed to the relationships.

Variation



Coefficient of Determiation

- The **coefficient of determination** is the ratio of the explained variation to the total variation.
- The symbol for the coefficient of determination is r^2.
- $$r^2 = \frac{\text{explained variation}}{\text{total variation}}$$
- Another way to arrive at the value for r^2 is to square the correlation coefficient.

Coefficient of Nondetermiation

- The **coefficient of nondetermination** is a measure of the unexplained variation.
- The formula for the coefficient of nondetermination is $1.00 - r^2$.

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:

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

Chapter 10
Correlation and Regression

Section 10-3

Example 10-12

Page #570

Example 10-12: Copy Machine 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

Example 10-12: Copy Machine Costs

Machine	Age x (years)	Monthly cost, y	y'	$y - y'$	$(y - y')^2$
A	1	62	63.70	–1.70	2.89
B	2	78	71.83	6.17	38.0689
C	3	70	79.96	–9.96	99.2016
D	4	90	88.09	1.91	3.6481
E	4	93	88.09	4.91	24.1081
F	6	103	104.35	–1.35	1.8225
					169.7392

$$y' = 55.57 + 8.13x$$

$$y' = 55.57 + 8.13(1) = 63.70$$

$$y' = 55.57 + 8.13(2) = 71.83$$

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

$$y' = 55.57 + 8.13(4) = 88.09$$

$$y' = 55.57 + 8.13(6) = 104.35$$

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

$$s_{est} = \sqrt{\frac{169.7392}{4}} = 6.51$$

Chapter 10
Correlation and Regression

Section 10-3

Example 10-13

Page #571

Example 10-13: Copy Machine Costs

Find the standard error of the estimate for the data for Example 10–12 by using the formula below. The equation of the regression line is y = 55.57 + 8.13*x*.

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

Example 10-13: Copy Machine Costs

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
	$\Sigma y = 496$	$\Sigma xy = 1778$	$\Sigma y^2 = 42{,}186$

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

$$s_{est} = \sqrt{\frac{42{,}186 - 55.57(496) - 8.13(1778)}{4}} = 6.48$$

Formula for the Prediction Interval about a Value y'

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

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

with d.f. $= n - 2$

Chapter 10
Correlation and Regression

Section 10-3

Example 10-14

Page #573

Example 10-14: Copy Machine Costs

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.

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.

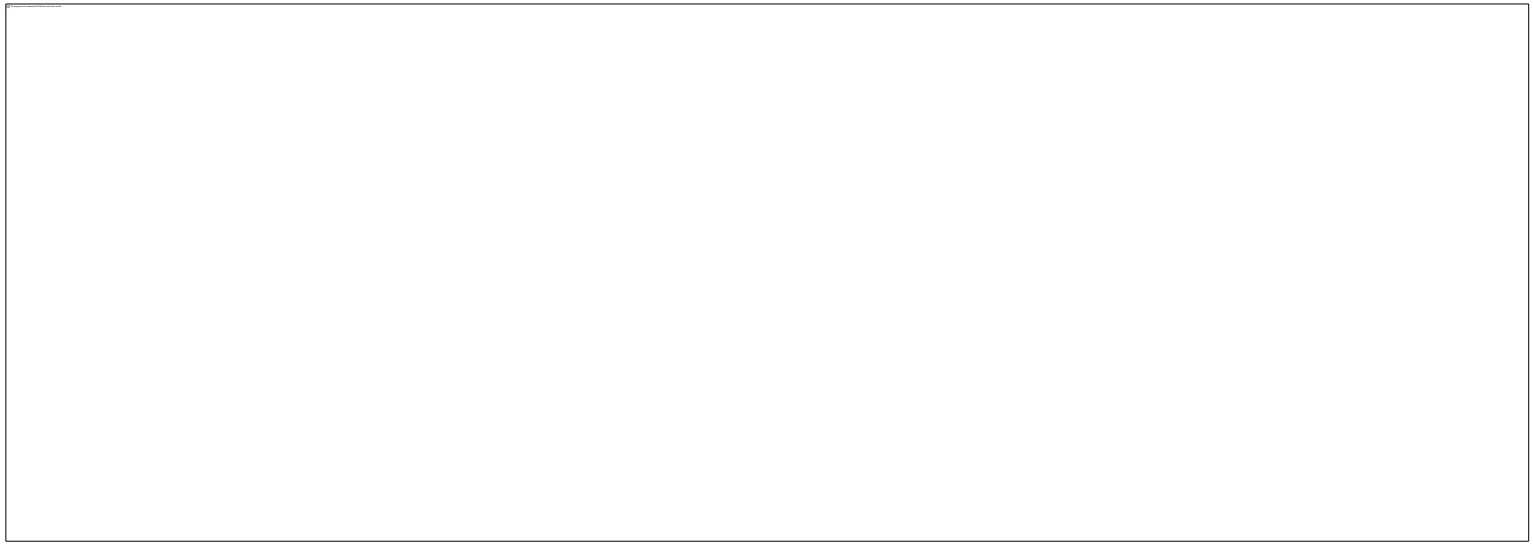
$$y' = 55.57 + 8.13(3) = 79.96$$

Step 3: Find s_{est}.

$$s_{est} = 6.48 \quad \text{(as shown in Example 10-13)}$$

Example 10-14: Copy Machine Costs

Step 4: Substitute in the formula and solve.



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

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

Example 10-14: Copy Machine Costs

Step 4: Substitute in the formula and solve.

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

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

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

$$60.53<y<99.39$$

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

10.4 Multiple Regression (Optional)

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

$$y' = a + b_1 x_1 + b_2 x_2 + \mathrm{L} \; + b_k x_k$$

where

$x_1, x_2, \mathrm{K} \; , x_k$ = independent variables.

Assumptions for Multiple Regression

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

Multiple Correlation Coefficient

- 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*.

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} = correlation coefficient for y and x_1

r_{yx_2} = correlation coefficient for y and x_2

$r_{x_1x_2}$ = correlation coefficient for x_1 and x_2

Chapter 10
Correlation and Regression

Section 10-4

Example 10-15

Page #578

Example 10-15: State Board Scores

A 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. Find the value of *R*.

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

Example 10-15: State Board Scores

A 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. Find the value of *R.*

The values of the correlation coefficients are

$$r_{yx_1} = 0.845$$

$$r_{yx_2} = 0.791$$

$$r_{x_1x_2} = 0.371$$

Example 10-15: State Board Scores

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

$$R = \sqrt{\frac{(0.845)^2 + (0.791)^2 - 2(0.845)(0.791)(0.371)}{1 - (0.371)^2}}$$

$$R = 0.989$$

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.

F Test for Significance of R

The formula for the F test is

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

where

n = the number of data groups

k = the number of independent variables.

d.f.N. = $n - k$

d.f.D. = $n - k - 1$

Chapter 10
Correlation and Regression

Section 10-4

Example 10-16

Page #579

Example 10-16: State Board Scores

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

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

$$F = \frac{0.978/2}{(1-0.978)/(5-2-1)} = 44.45$$

The critical value obtained from Table H with a 0.05, d.f.N. = 3, and d.f.D. = 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

The formula for the adjusted R^2 is

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

Chapter 10
Correlation and Regression

Section 10-4

Example 10-17

Page #580

Example 10-17: State Board Scores

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

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

$$R^2_{\text{adj}} = 1 - \frac{\left(1 - 0.989^2\right)\left(5 - 1\right)}{5 - 2 - 1} = 0.956$$

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.