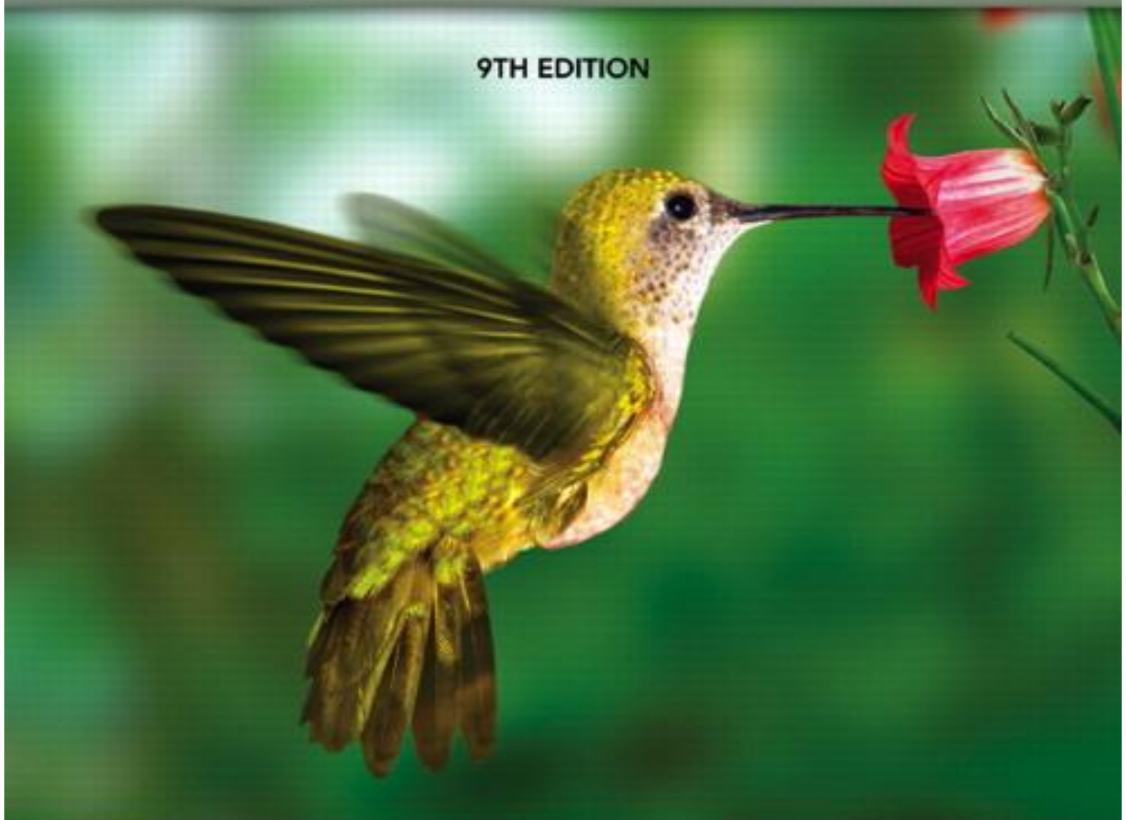
Introductory STATISTICS

9TH EDITION



Neil
WEISS

Chapter 14

Descriptive Methods in Regression and Correlation

Section 14.1

Linear Equations with One Independent Variable

Definition 14.1

y-Intercept and Slope

For a linear equation $y = b_0 + b_1 x$, the number b_0 is called the **y-intercept** and the number b_1 is called the **slope.**

Section 14.2

The Regression Equation

Table 14.2

Age and price data for a sample of 11 Orions

Car	Age (yr) x	Price ($100) y
1	5	85
2	4	103
3	6	70
4	5	82
5	5	89
6	5	98
7	6	66
8	6	95
9	2	169
10	7	70
11	7	48

Figure 14.7

Scatterplot for the age and price data of Orions from Table 14.2

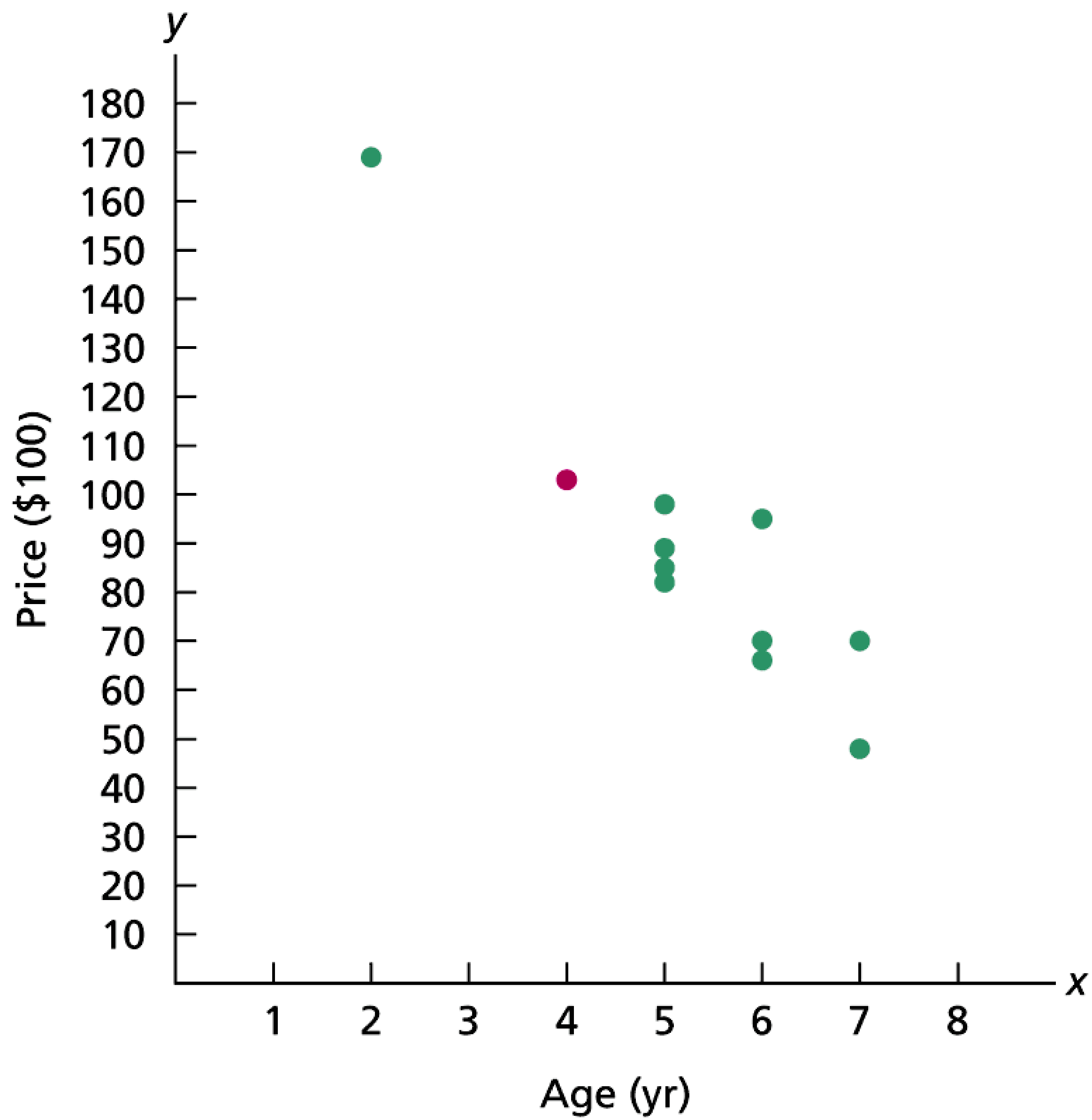


Table 14.3 & Figure 14.8

Four data points

x	y
1	1
1	2
2	2
4	6

Scatterplot for the data points in Table 14.3

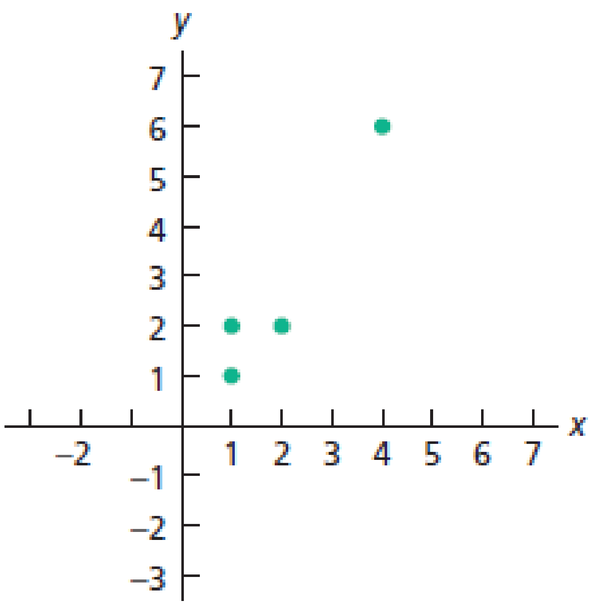


Figure 14.9

Two possible lines to fit the data points in Table 14.3

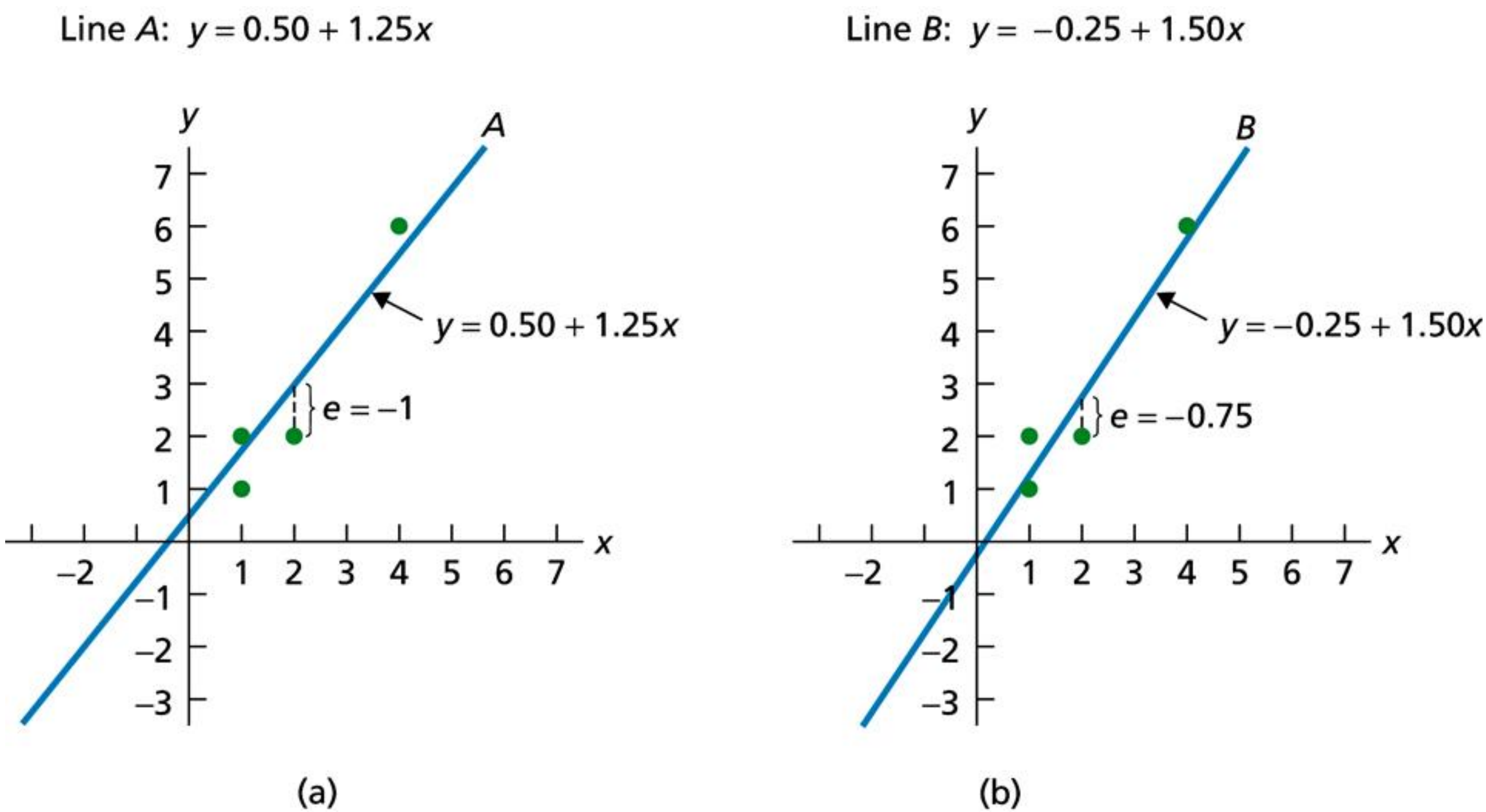


Table 14.4

Determining how well the data points in Table 14.3 are fit by (a) Line A and (b) Line B

Line A: $y = 0.50 + 1.25x$

x	y	$\hat{y}$	e	e^2
1	1	1.75	−0.75	0.5625
1	2	1.75	0.25	0.0625
2	2	3.00	−1.00	1.0000
4	6	5.50	0.50	0.2500
				1.8750

(a)

Line B: $y = -0.25 + 1.50x$

x	y	$\hat{y}$	e	e^2
1	1	1.25	−0.25	0.0625
1	2	1.25	0.75	0.5625
2	2	2.75	−0.75	0.5625
4	6	5.75	0.25	0.0625
				1.2500

(b)

Key Fact 14.2 & Definition 14.2

Least-Squares Criterion

The **least-squares criterion** is that the line that best fits a set of data points is the one having the smallest possible sum of squared errors.

Regression Line and Regression Equation

Regression line: The line that best fits a set of data points according to the least-squares criterion.

Regression equation: The equation of the regression line.

Definition 14.3

Notation Used in Regression and Correlation

For a set of n data points, the defining and computing formulas for $\boldsymbol{S_{xx}}$, $\boldsymbol{S_{xy}}$, and $\boldsymbol{S_{yy}}$ are as follows.

Quantity	Defining formula	Computing formula
S_{xx}	$\Sigma(x_i - \bar{x})^2$	$\Sigma x_i^2 - (\Sigma x_i)^2/n$
S_{xy}	$\Sigma(x_i - \bar{x})(y_i - \bar{y})$	$\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)/n$
S_{yy}	$\Sigma(y_i - \bar{y})^2$	$\Sigma y_i^2 - (\Sigma y_i)^2/n$

Formula 14.1

Regression Equation

The regression equation for a set of n data points is $\hat{y} = b_0 + b_1 x$, where

$$b_1 = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad b_0 = \frac{1}{n}(\Sigma y_i - b_1 \Sigma x_i) = \bar{y} - b_1 \bar{x}.$$

Table 14.5

Table for computing the regression equation for the Orion data

Age (yr) x	Price ($100) y	xy	x^2
5	85	425	25
4	103	412	16
6	70	420	36
5	82	410	25
5	89	445	25
5	98	490	25
6	66	396	36
6	95	570	36
2	169	338	4
7	70	490	49
7	48	336	49
58	975	4732	326

Figure 14.10

Regression line and data points for Orion data

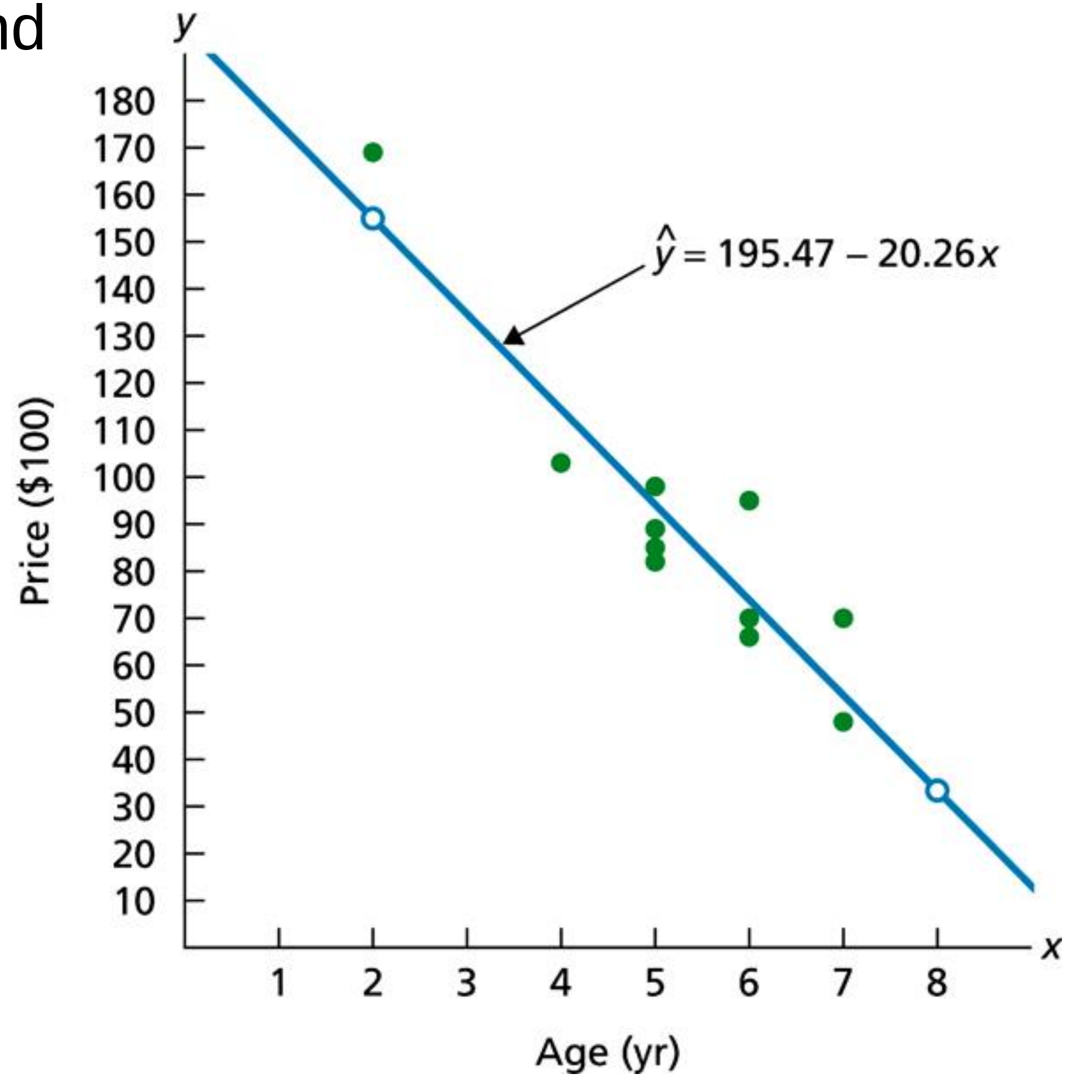


Figure 14.11

Extraplation in
the Orion example

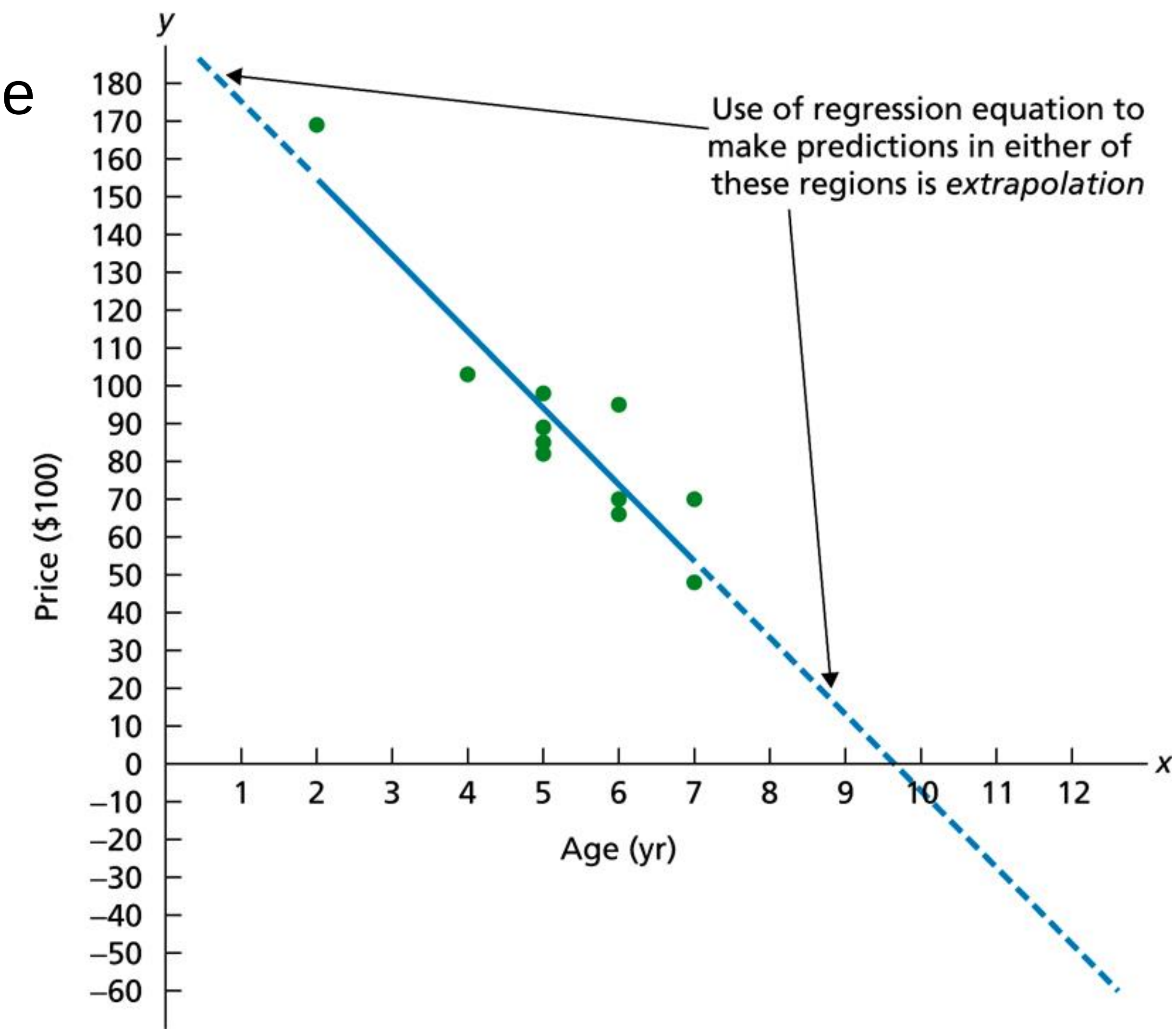
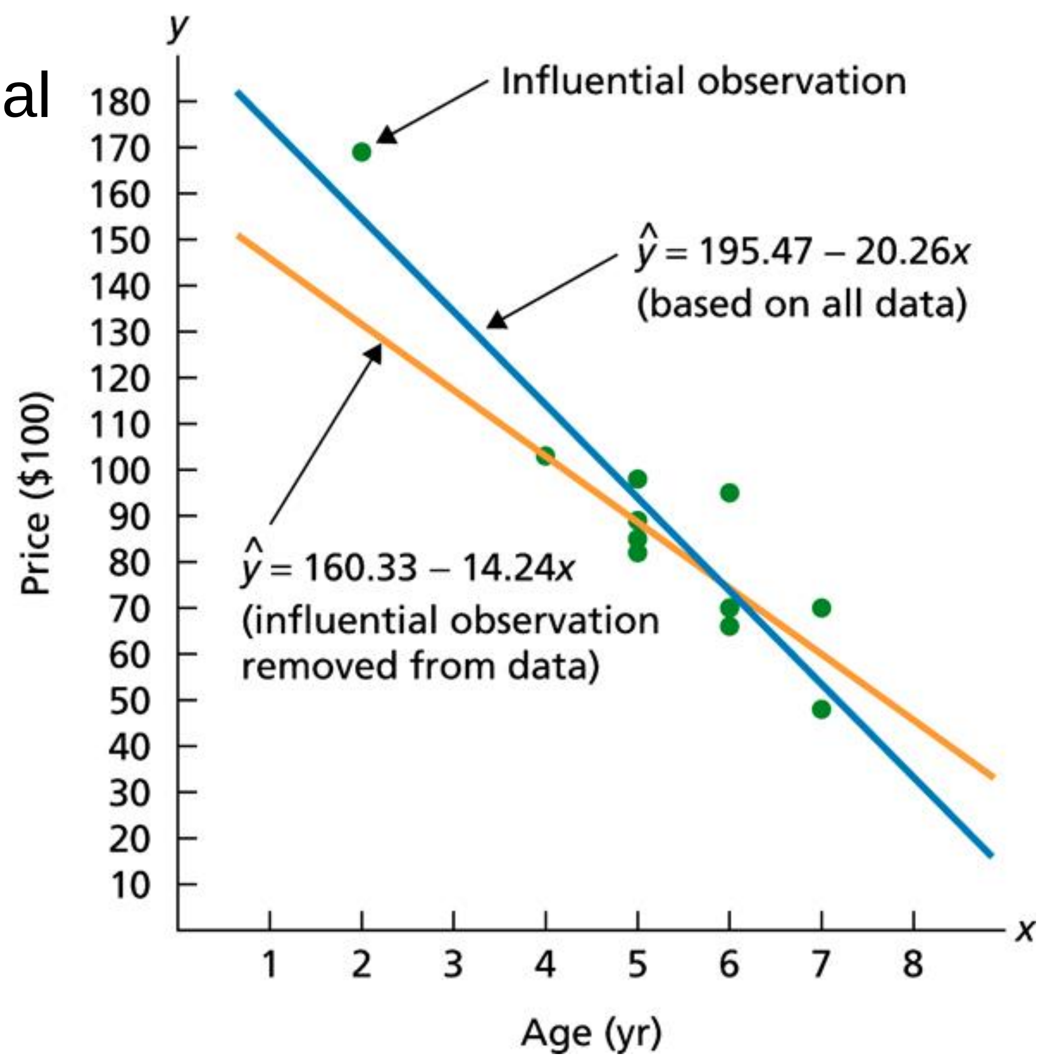


Figure 14.12

Regression lines with and without the influential observation removed



Section 14.3

The Coefficient of Determination

Definition 14.5

Sums of Squares in Regression

Total sum of squares, *SST*: The total variation in the observed values of the response variable: $SST = \Sigma(y_i - \bar{y})^2$.

Regression sum of squares, *SSR*: The variation in the observed values of the response variable explained by the regression: $SSR = \Sigma(\hat{y}_i - \bar{y})^2$.

Error sum of squares, *SSE*: The variation in the observed values of the response variable not explained by the regression: $SSE = \Sigma(y_i - \hat{y}_i)^2$.

Table 14.6

Table for computing *SST* for the Orion price data

Age (yr) x	Price ($100) y	$y - \bar{y}$	$(y - \bar{y})^2$
5	85	−3.64	13.2
4	103	14.36	206.3
6	70	−18.64	347.3
5	82	−6.64	44.0
5	89	0.36	0.1
5	98	9.36	87.7
6	66	−22.64	512.4
6	95	6.36	40.5
2	169	80.36	6458.3
7	70	−18.64	347.3
7	48	−40.64	1651.3
	975		9708.5

Table 14.7

Table for computing *SSR* for the Orion price data

Age (yr) x	Price ($100) y	$\hat{y}$	$\hat{y} - \bar{y}$	$(\hat{y} - \bar{y})^2$
5	85	94.16	5.53	30.5
4	103	114.42	25.79	665.0
6	70	73.90	−14.74	217.1
5	82	94.16	5.53	30.5
5	89	94.16	5.53	30.5
5	98	94.16	5.53	30.5
6	66	73.90	−14.74	217.1
6	95	73.90	−14.74	217.1
2	169	154.95	66.31	4397.0
7	70	53.64	−35.00	1224.8
7	48	53.64	−35.00	1224.8
				8285.0

Table 14.8

Table for computing *SSE* for the Orion data

Age (yr) x	Price ($100) y	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$
5	85	94.16	−9.16	83.9
4	103	114.42	−11.42	130.5
6	70	73.90	−3.90	15.2
5	82	94.16	−12.16	147.9
5	89	94.16	−5.16	26.6
5	98	94.16	3.84	14.7
6	66	73.90	−7.90	62.4
6	95	73.90	21.10	445.2
2	169	154.95	14.05	197.5
7	70	53.64	16.36	267.7
7	48	53.64	−5.64	31.8
				1423.5

Section 14.4

Linear Correlation

Definition 14.7 & Formula 14.3

Linear Correlation Coefficient

For a set of n data points, the **linear correlation coefficient, r,** is defined by

$$r = \frac{\frac{1}{n-1}\Sigma(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y},$$

where s_x and s_y denote the sample standard deviations of the x-values and y-values, respectively.

Using algebra, we can show that the linear correlation coefficient can be expressed as $r = S_{xy}/\sqrt{S_{xx}S_{yy}}$, where S_{xx}, S_{xy}, and S_{yy} are given in Definition 14.3 on page 637. Referring again to that definition, we get Formula 14.3.

Computing Formula for a Linear Correlation Coefficient

The computing formula for a linear correlation coefficient is

$$r = \frac{\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)/n}{\sqrt{[\Sigma x_i^2 - (\Sigma x_i)^2/n][\Sigma y_i^2 - (\Sigma y_i)^2/n]}}.$$

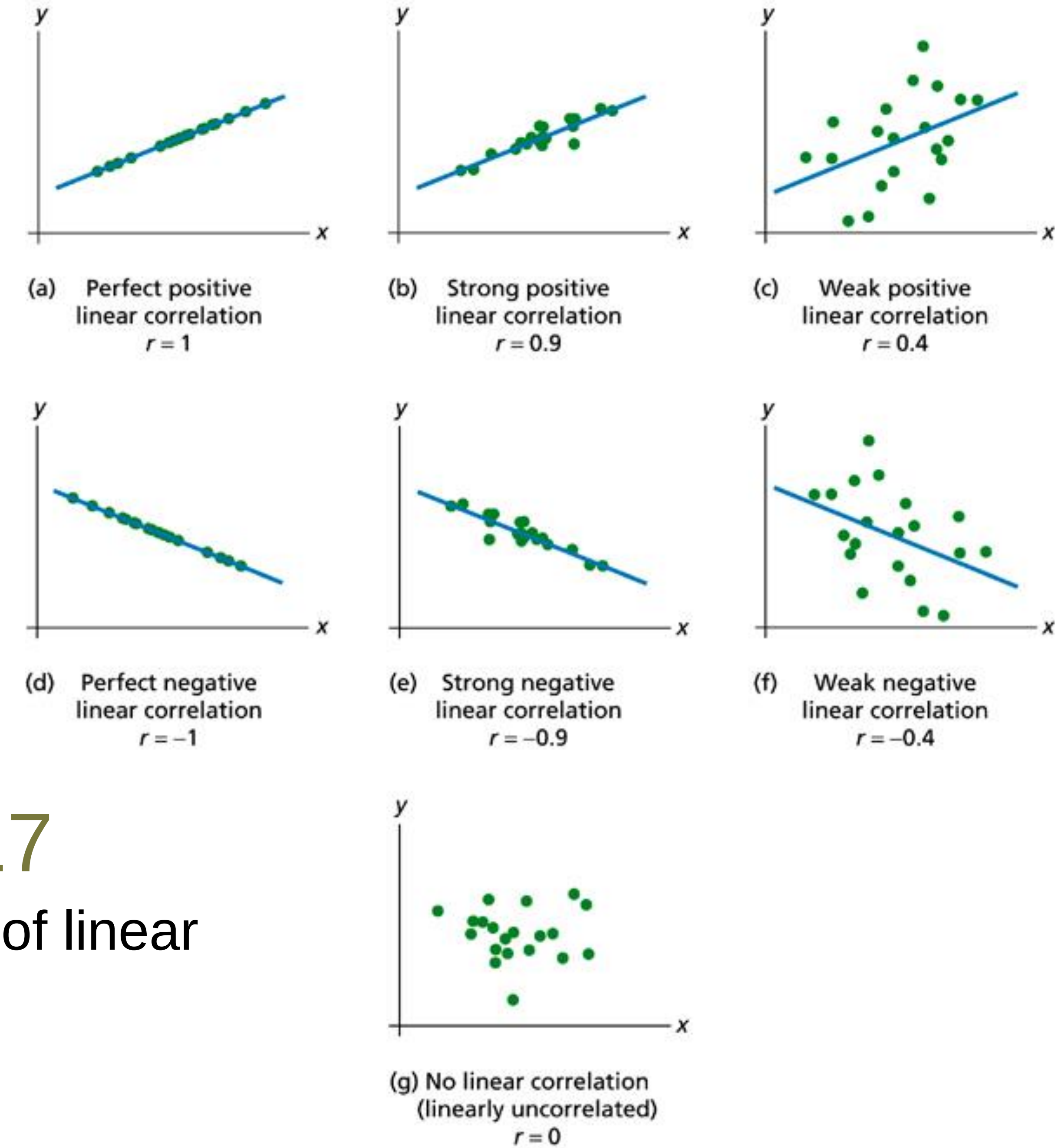


Figure 14.17

Various degrees of linear correlation