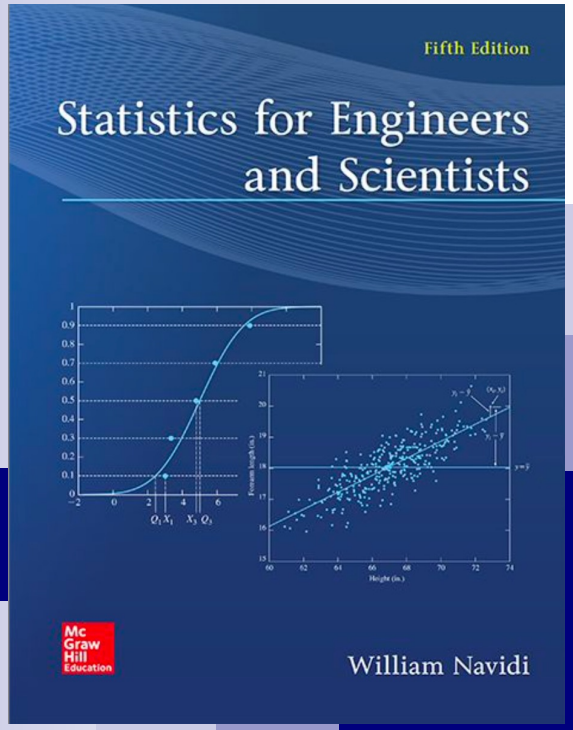


Chapter 7

Correlation and Simple Linear Regression

Chapter 7 Overview

7-1 Scatter Plots and Correlation

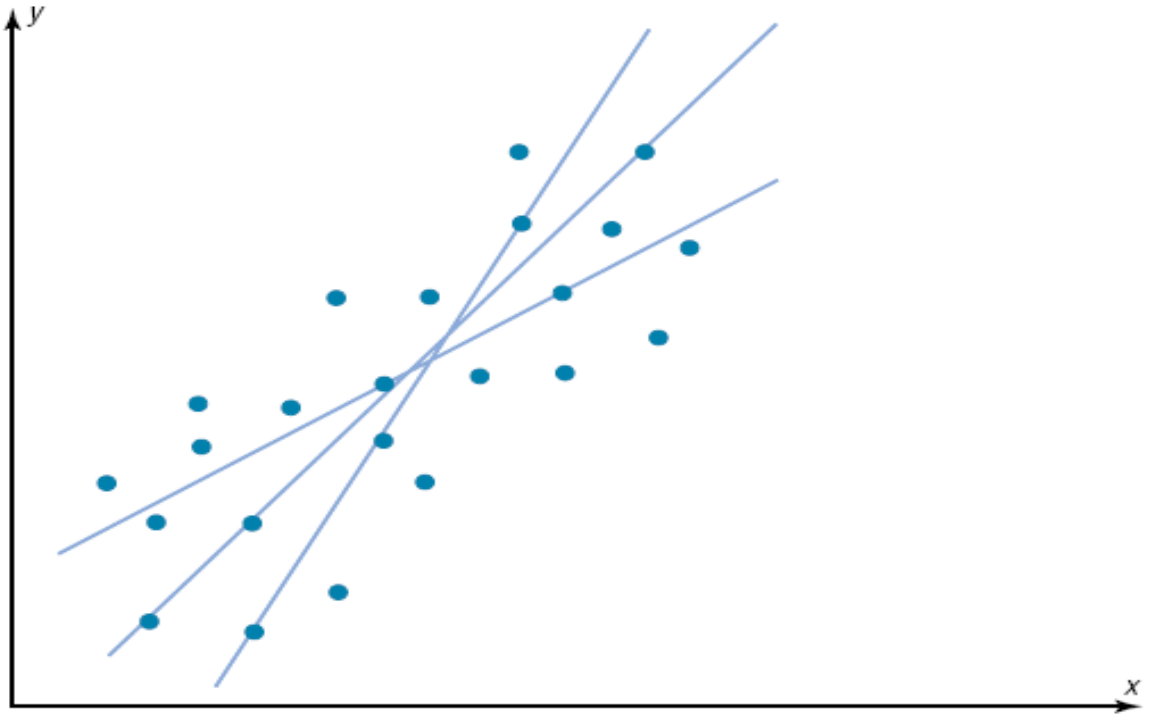
7-2 The Least-Squares Line

7-3 Uncertainties in the Least-Squares Coefficients

7-4 Checking Assumptions and Transforming Data

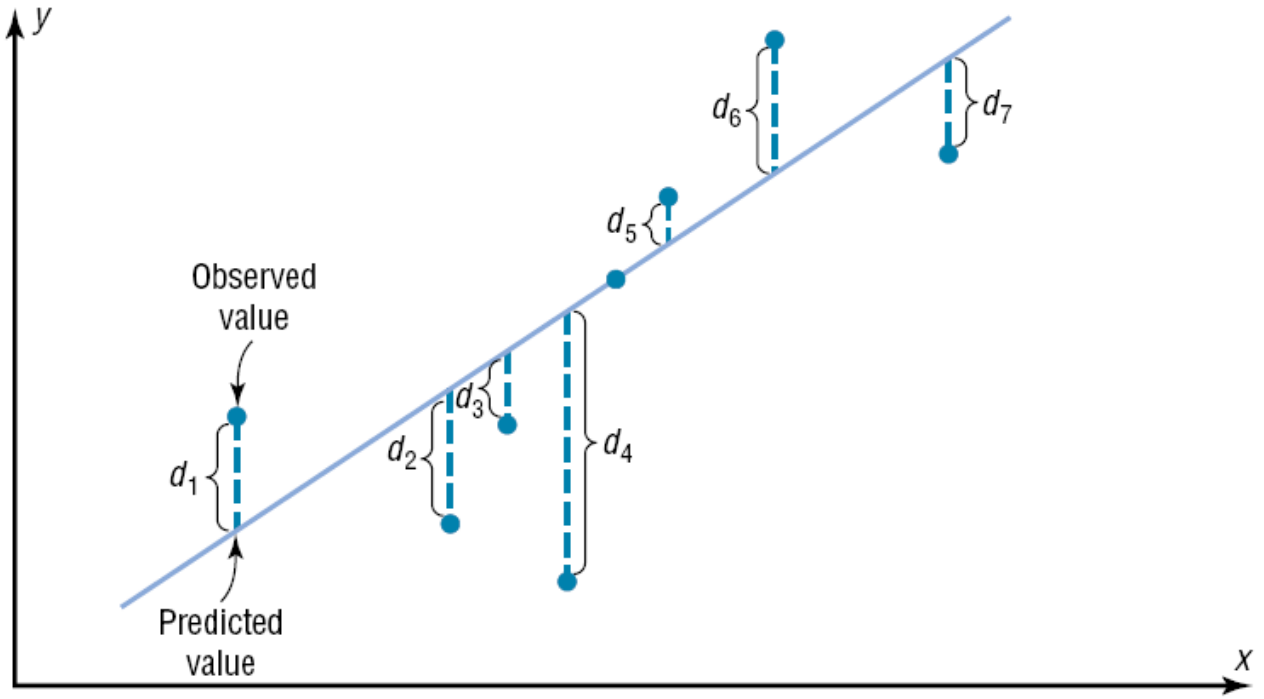
Simple Linear Regression

- 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, also known as ***the least-squares line***.



Simple Linear Regression…

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



Example – Stiffness of a Spring (p.532)

- Springs are used in applications for their ability to extend (stretch) under load.
- The stiffness of a spring is measured by the "spring constant", which is the length that the spring will be extended by one unit of force or load.
- To make sure that a given spring functions appropriately, it is necessary to estimate its spring constant with good accuracy and precision.
- In an experiment, a spring is hung vertically with the top end fixed, and weights are hung one at a time from the other end.

Example – Stiffness of a Spring…

- Let $x_1,\ldots, x_n$ represent the weights, and let l_i represent the length of the spring under the load x_i,

- Hooke's law states that

$$l_i = \beta_0 + \beta_1 x_i$$

where β_0 is the length of the spring when unloaded and β_1 is the spring constant.

Example – Stiffness of a Spring…

- Let $\mathbf{y_i}$ be the measured length of the spring under load $\mathbf{x_i}$.

- Because of measurement error (see Chapter 3), y_i will differ from the true length l_i.

- We write

$$y_i = l_i + \varepsilon_i$$

where ε_i is the error in the *i-th* measurement.

Example – Stiffness of a Spring…

- Combining these two equations we obtain so-called **a linear model**

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

y_i is the dependent variable

x_i is the independent variable

β_0 and β_1 are the regression coefficients

ε_i is the error

Example – Stiffness of a Spring…

TABLE 7.1 Measured lengths of a spring under various loads

Weight (lb) x	Measured Length (in.) y
0.0	5.06
0.2	5.01
0.4	5.12
0.6	5.13
0.8	5.14
1.0	5.16
1.2	5.25
1.4	5.19
1.6	5.24
1.8	5.46
2.0	5.40
2.2	5.57
2.4	5.47
2.6	5.53
2.8	5.61
3.0	5.59
3.2	5.61
3.4	5.75
3.6	5.68
3.8	5.80

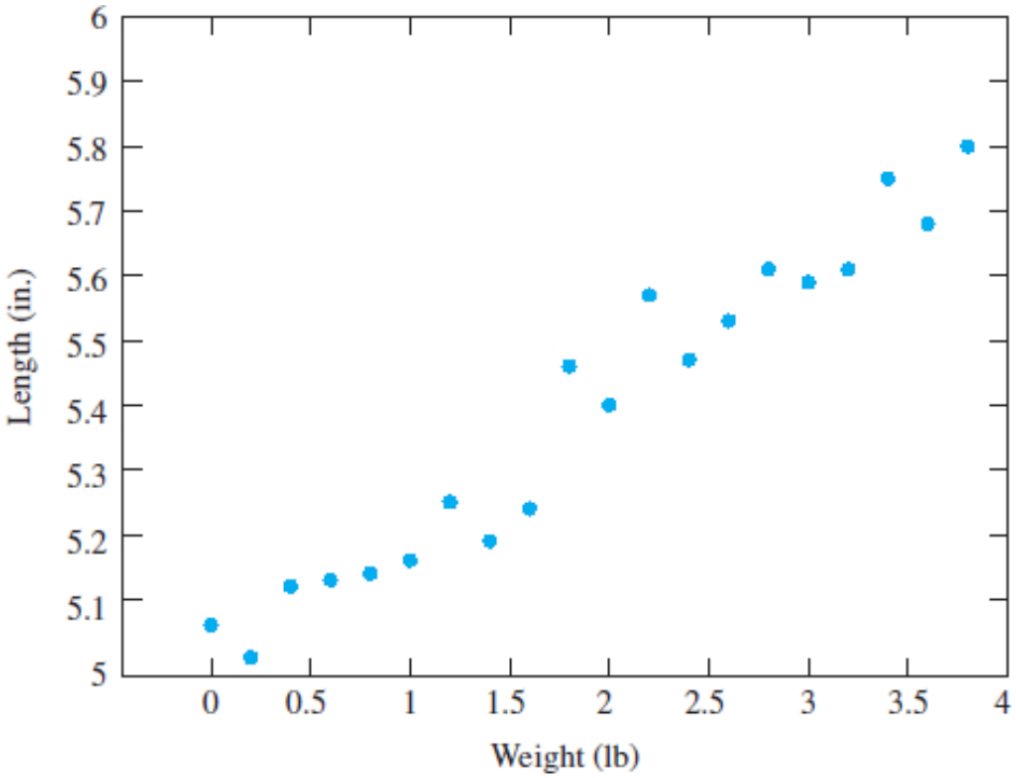


FIGURE 7.9 Plot of measured lengths of a spring versus load.

Example – Stiffness of a Spring…

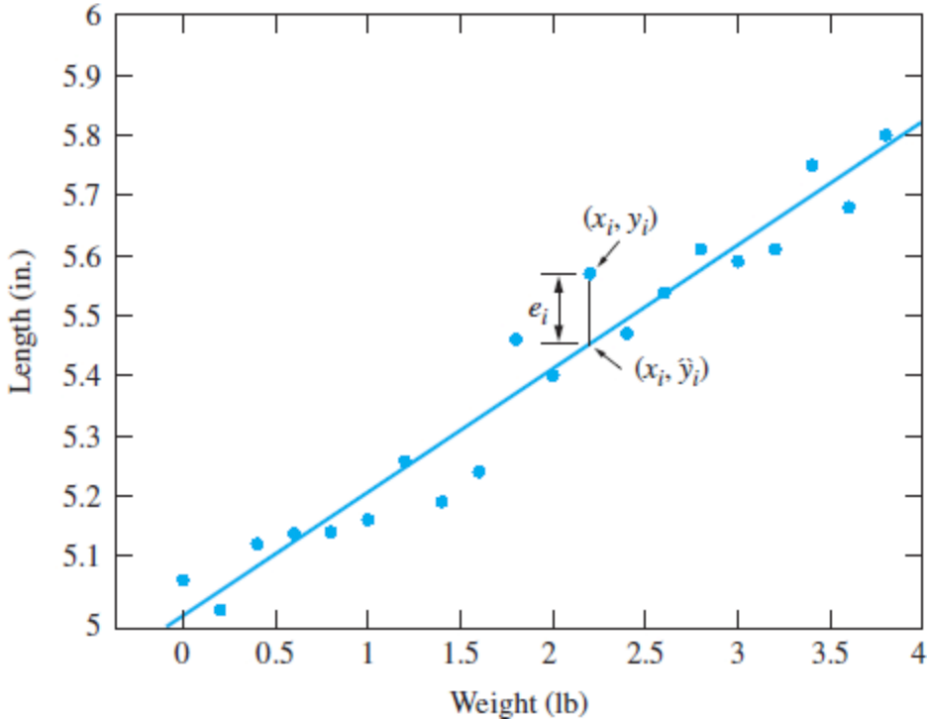


FIGURE 7.10 Plot of measured lengths of a spring versus load. The least-squares line $y = \hat{\beta}_0 + \hat{\beta}_1 x$ is superimposed. The vertical distance from a data point (x_i, y_i) to the point $(x_i, \hat{y}_i)$ on the line is the ith residual e_i. The least-squares line is the line that minimizes the sum of the squared residuals.

Equation of the Least-Squares Line (p.535)

- The equation of the least-squares line is

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

the slope of the line

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

the intercept

Equation of the Least-Squares Line (p.535)

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Computing Formulas

The expressions on the right are equivalent to those on the left, and are often easier to compute:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \qquad (7.16)$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 \qquad (7.17)$$

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \qquad (7.18)$$

Example 7.6 – Stiffness of a Spring…

Using the Hooke's law data in Table 7.1, compute the least-squares estimates of the spring constant and the unloaded length of the spring.

Write the equation of the least-squares line.

Example 7.6 – Stiffness of a Spring…

Solution

The estimate of the spring constant is $\hat{\beta}_1$, and the estimate of the unloaded length is $\hat{\beta}_0$. From Table 7.1 we compute:

$$\bar{x} = 1.9000 \qquad \bar{y} = 5.3885$$

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = 26.6000$$

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = 5.4430$$

Using Equations (7.14) and (7.15), we compute

$$\hat{\beta}_1 = \frac{5.4430}{26.6000} = 0.2046$$

Page 536

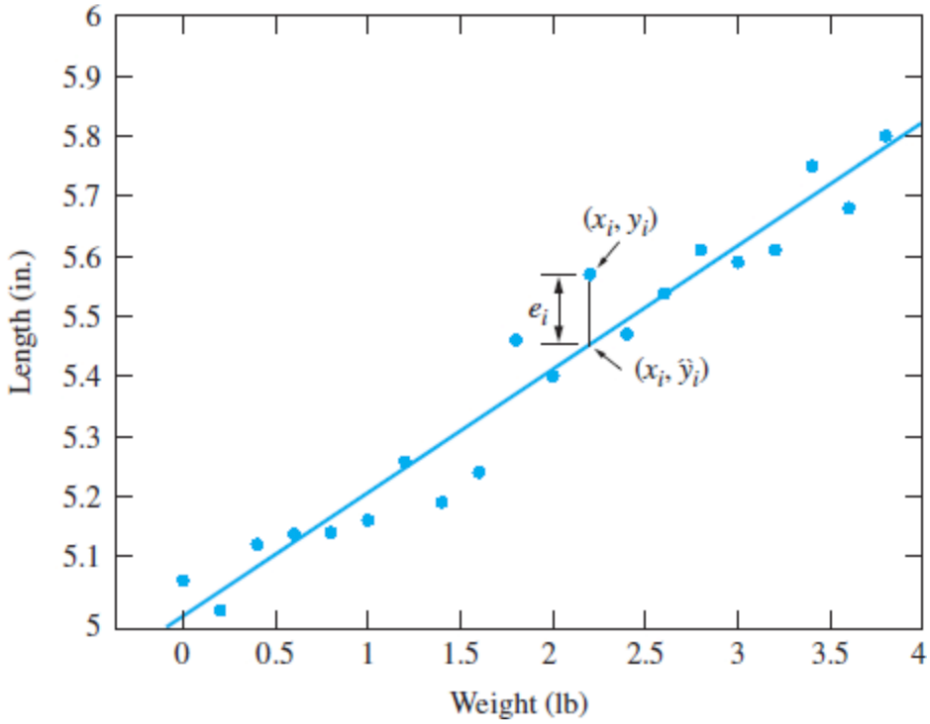
$$\hat{\beta}_0 = 5.3885 - (0.2046)(1.9000) = 4.9997$$

The equation of the least-squares line is $y = \hat{\beta}_0 + \hat{\beta}_1 x$. Substituting the computed values for $\hat{\beta}_0$ and $\hat{\beta}_1$, we obtain

$$y = 4.9997 + 0.2046x$$

Example 7.6 – Stiffness of a Spring…

Using the equation of the least-squares line, we can compute the fitted values $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ and the residuals $e_i = y_i - \hat{y}_i$ for each point (x_i, y_i) in the Hooke's law data set. The results are presented in Table 7.2. The point whose residual is shown in Figure 7.10 (page 534) is the one where $x = 2.2$.



Example 7.7

Using the Hooke's law data, estimate the length of the spring under a load of 1.3 lb.

Solution

In Example 7.6, the equation of the least-squares line was computed to be $y = 4.9997 + 0.2046x$. Using the value $x = 1.3$, we estimate the length of the spring under a load of 1.3 lb to be

$$\hat{y} = 4.9997 + (0.2046)(1.3) = 5.27 \text{ in.}$$

Example 7.8

Using the Hooke's law data, estimate the length of the spring under a load of 1.4 lb.

Solution

The estimate is $\hat{y} = 4.9997 + (0.2046)(1.4) = 5.29$ in.

Example 7.8

Using the Hooke's law data, estimate the length of the spring under a load of 1.4 lb.

Solution

The estimate is $\hat{y} = 4.9997 + (0.2046)(1.4) = 5.29$ in.

- Note that the measured length at a load of 1.4 was 5.19 in. (see Table 7.2).
- The least-squares estimate of 5.29 in. is based on all the data and is more precise (has smaller uncertainty).

TABLE 7.1 Measured lengths of a spring under various loads

Weight (lb) x	Measured Length (in.) y
0.0	5.06
0.2	5.01
0.4	5.12
0.6	5.13
0.8	5.14
1.0	5.16
1.2	5.25
1.4	5.19
1.6	5.24
1.8	5.46
2.0	5.40
2.2	5.57
2.4	5.47
2.6	5.53
2.8	5.61
3.0	5.59
3.2	5.61
3.4	5.75
3.6	5.68
3.8	5.80

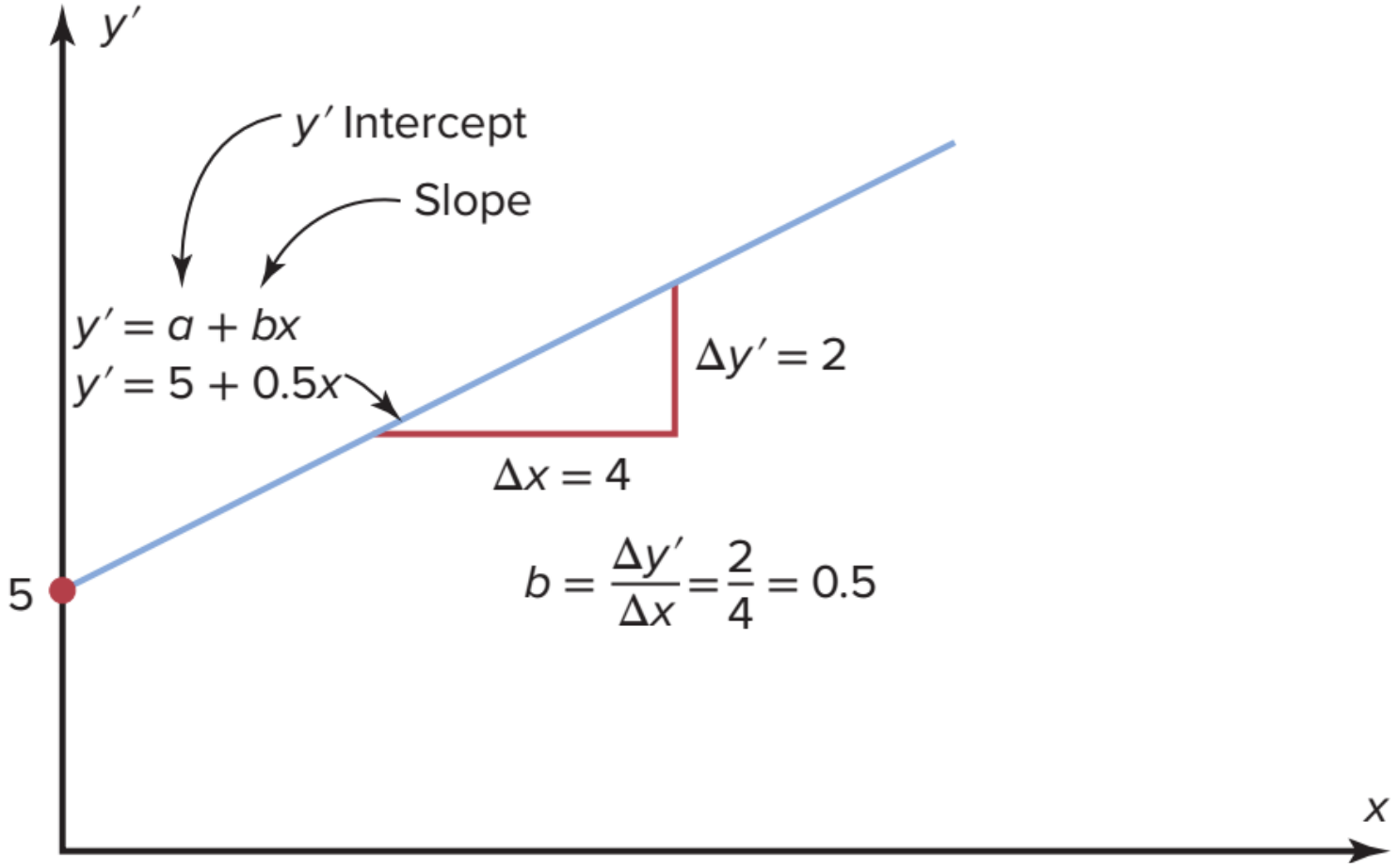
The Least-Squares Line

Summary

Given points $(x_1, y_1), \ldots, (x_n, y_n)$:

- The least-squares line is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.
- $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
- The quantities $\hat{\beta}_0$ and $\hat{\beta}_1$ can be thought of as estimates of a true slope β_1 and a true intercept β_0.
- For any x, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ is an estimate of the quantity $\beta_0 + \beta_1 x$.

Regression Line – Simplified Approach



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.

Procedure Table

Step 1: Make a table with subject, x, y, xy, x^2, and y^2 columns.

Step 2: Find the values of xy, x^2, and y^2. Place them in the appropriate columns and sum each column.

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

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

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: Car Rental Companies

Find the equation of the regression line for the data in Example 1 (in part 1), 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: Car Rental Companies…

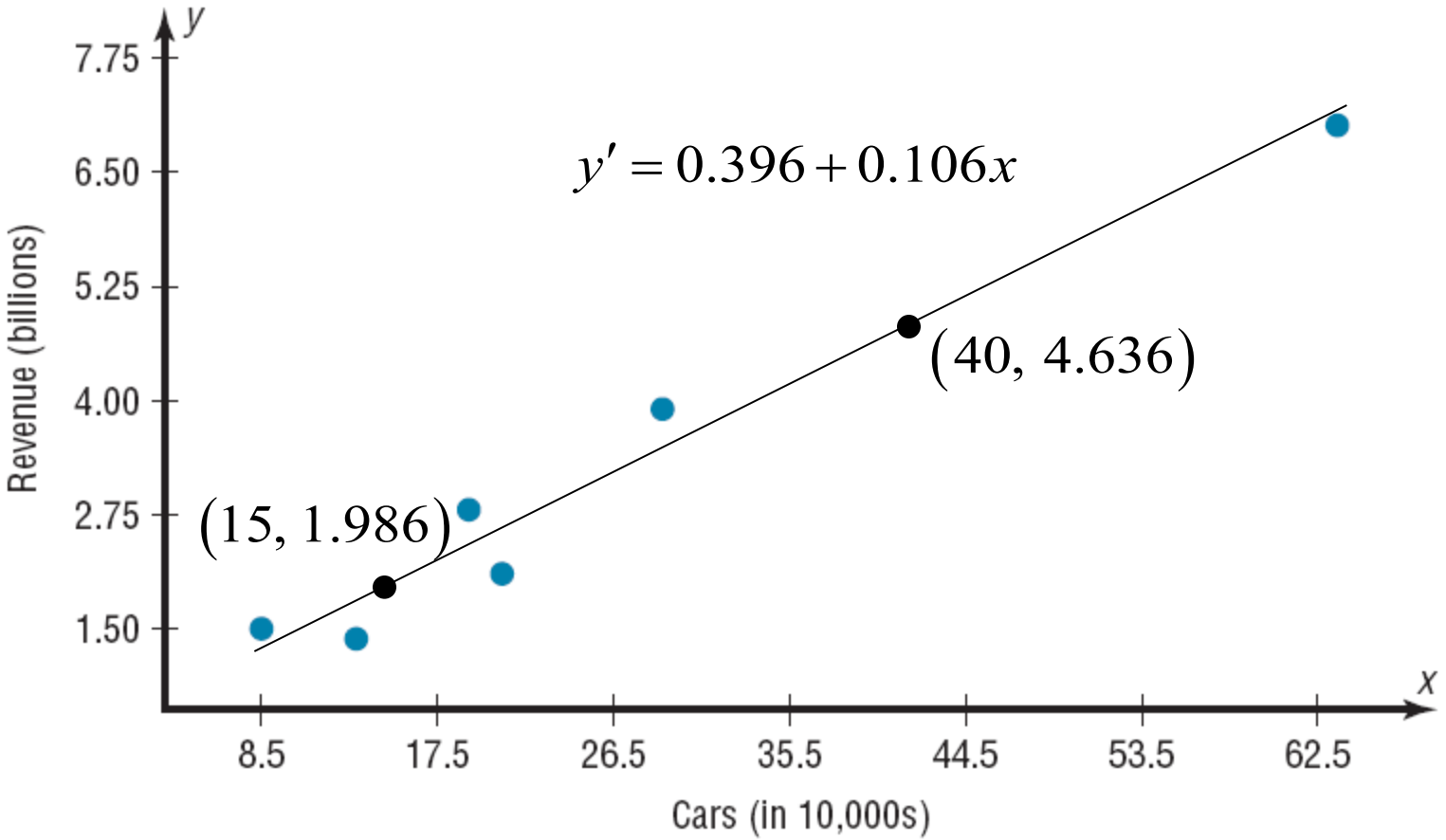
Find two points to sketch the graph of the regression line.

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: Car Rental Companies



Prediction in Regression

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

1. The points of the scatter plot fit the linear regression line **reasonably well**.
2. The value of ***r* is significant**.
3. The value of a specific x is **not much beyond** the observed values (x values) in the original data.
4. If ***r* is not significant**, then the best predicted value for a specific x value is **the mean of the *y* value** in the original data.

Example: Car Rental Companies…

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

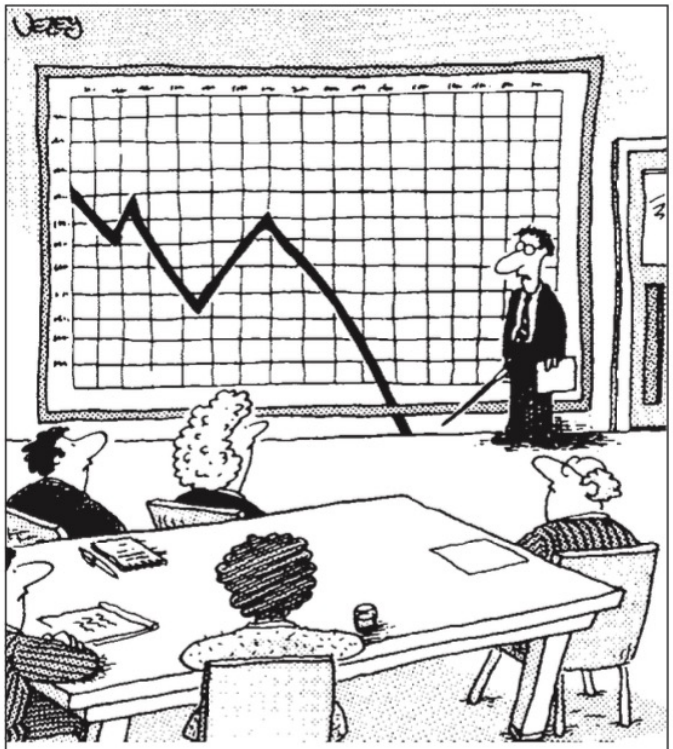
x = 20 corresponds to 200,000 cars.

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

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

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.



"At this point in my report, I'll ask all of you to follow me to the conference room directly below us!"

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

Measuring Goodness-of-Fit (p.541)

- A linear model fits well if there is a strong linear relationship between x and y.
- We mentioned in Section 7.1 that the **correlation coefficient *r*** measures the strength of the linear relationship between x and y.
- Therefore, ***r*** is a **goodness-of-fit statistic** for the linear model.

Measuring Goodness-of-Fit (p.541)

- Figure 7.12 presents Galton's data on forearm lengths versus heights.

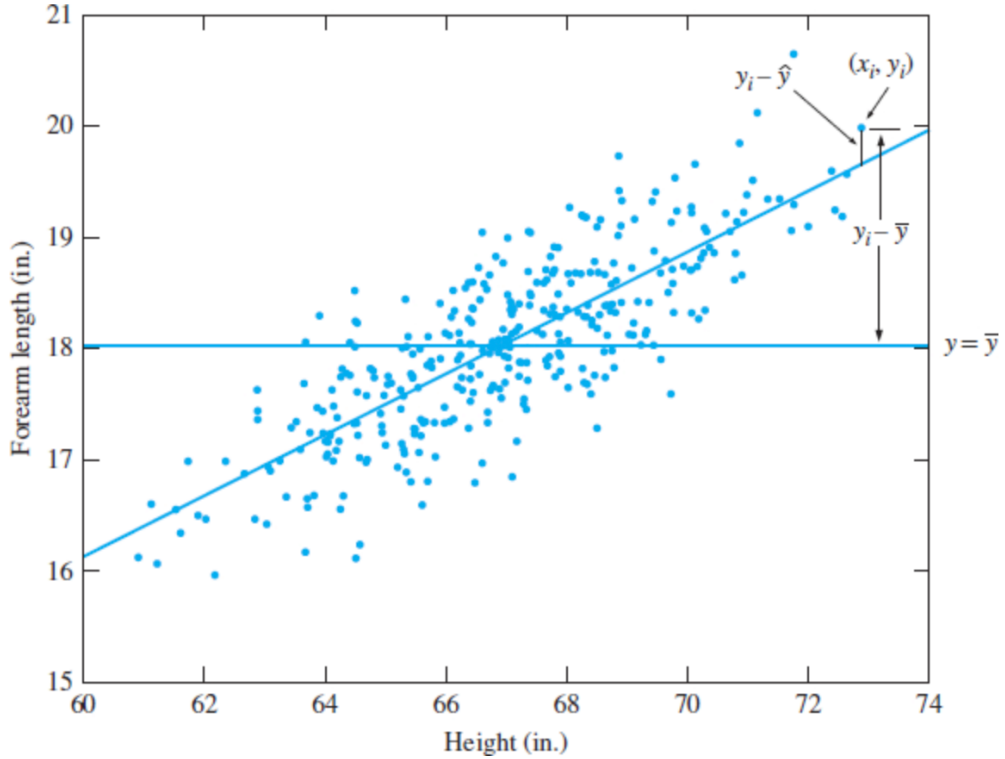


FIGURE 7.12 Heights and forearm lengths of men. The least-squares line and the horizontal line $y = \bar{y}$ are superimposed.

Measuring Goodness-of-Fit (p.541)

- We learned that if the correlation coefficient is significant, the equation of the regression line can be determined.
- Several other methods can be used to measure goodness-of-fit
  - The coefficient of determination
  - The standard error of the estimate
  - The prediction interval.
- Details will be discussed in the next section.

Section 7-2 Summary

- Relationships can be linear or curvilinear. To determine the shape, you draw a scatter plot of the variables. If the relationship is linear, the data can be approximated by a straight line, called the least-squares (regression) line, or the line of best fit.
- The closer the value of r is to +1 or -1, the more closely the points will fit the line.