Chapter 10

Correlation and Regression

Our interest in this chapter is in situations in which we can associate to each element of a population or sample two measurements x and y, particularly in the case that it is of interest to use the value of x to predict the value of y. For example, the population could be the air in automobile garages, x could be the electrical current produced by an electrochemical reaction taking place in a carbon monoxide meter, and y the concentration of carbon monoxide in the air. In this chapter we will learn statistical methods for analyzing the relationship between variables x and y in this context.

A list of all the formulas that appear anywhere in this chapter are collected in the last section for ease of reference.

10.1 Linear Relationships Between Variables

LEARNING OBJECTIVE

1. To learn what it means for two variables to exhibit a relationship that is close to linear but which contains an element of randomness.

The following table gives examples of the kinds of pairs of variables which could be of interest from a statistical point of view.

x	*y*
Predictor or independent variable	Response or dependent variable
Temperature in degrees Celsius	Temperature in degrees Fahrenheit
Area of a house (sq.ft.)	Value of the house
Age of a particular make and model car	Resale value of the car
Amount spent by a business on advertising in a year	Revenue received that year
Height of a 25-year-old man	Weight of the man

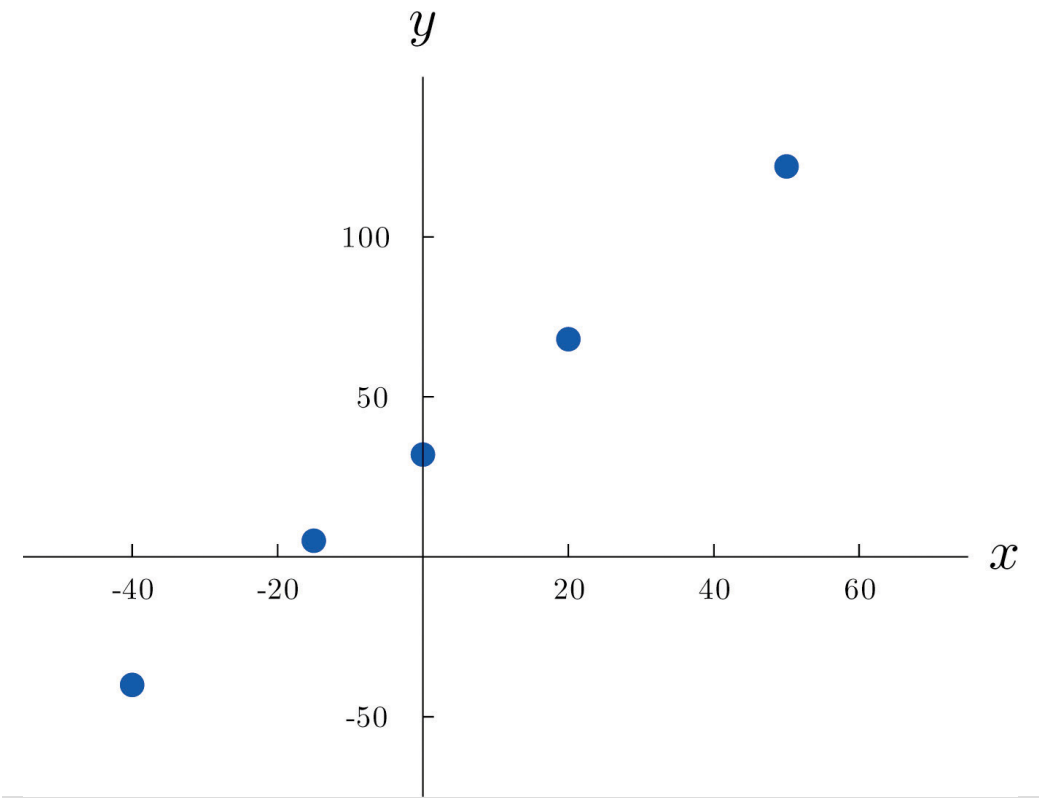
The first line in the table is different from all the rest because in that case and no other the relationship between the variables is **deterministic**: once the value of x is known the value of y is completely determined. In fact there is a formula for y in terms of x: $y = \frac{9}{5}x + 32$. Choosing several values for x and computing the corresponding value for y for each one using the formula gives the table

x	-40	-15	0	20	50
y	-40	5	32	68	122

We can plot these data by choosing a pair of perpendicular lines in the plane, called the coordinate axes, as shown in Figure 10.1 "Plot of Celsius and Fahrenheit Temperature Pairs". Then to each pair of numbers in the table we associate a unique point in the plane, the point that lies x units to the right of the vertical axis (to the left if $x < 0$) and y units above the horizontal axis (below if $y < 0$). The

relationship between x and y is called a **linear relationship** because the points so plotted all lie on a single straight line. The number $\frac{9}{5}$ in the equation $y = \frac{9}{5}x + 32$ is the **slope** of the line, and measures its steepness. It describes how y changes in response to a change in x: if x increases by 1 unit then y increases (since $\frac{9}{5}$ is positive) by $\frac{9}{5}$ unit. If the slope had been negative then y would have decreased in response to an increase in x. The number 32 in the formula $y = \frac{9}{5}x + 32$ is the y-**intercept** of the line; it identifies where the line crosses the y-axis. You may recall from an earlier course that every non-vertical line in the plane is described by an equation of the form $y = mx + b$, where m is the slope of the line and b is its y-intercept.

Figure 10.1 Plot of Celsius and Fahrenheit Temperature Pairs

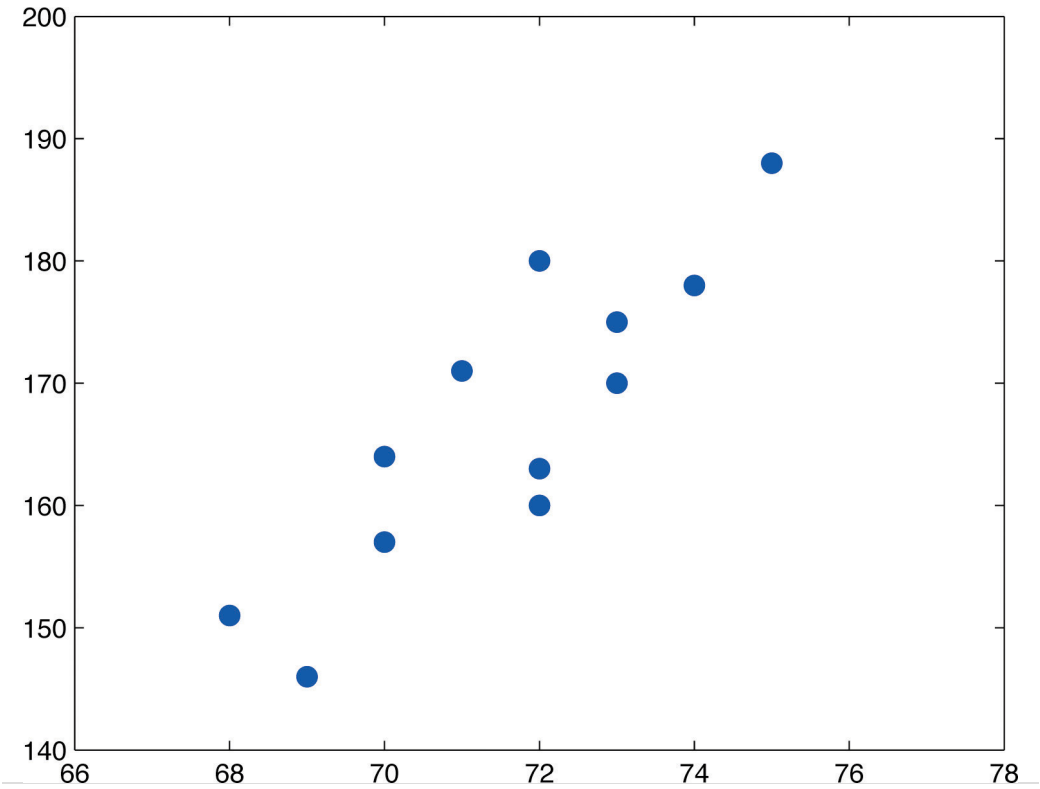


The relationship between x and y in the temperature example is deterministic because once the value of x is known, the value of y is completely determined. In contrast, all the other relationships listed in the table above have an element of randomness in them. Consider the relationship described in the last line of the table, the height x of a man aged 25 and his weight y. If we were to randomly select several 25-year-old men and measure the height and weight of each one, we might obtain a collection of (x, y) pairs something like this:

$$(68,151)\quad (69,146)\quad (70,157)\quad (70,164)\quad (71,171)\quad (72,160)$$
$$(72,163)\quad (72,180)\quad (73,170)\quad (73,175)\quad (74,178)\quad (75,188)$$

A plot of these data is shown in Figure 10.2 "Plot of Height and Weight Pairs". Such a plot is called a **scatter diagram** or **scatter plot**. Looking at the plot it is evident that there exists a linear relationship between height x and weight y, but not a perfect one. The points appear to be following a line, but not exactly. There is an element of randomness present.

Figure 10.2 *Plot of Height and Weight Pairs*



In this chapter we will analyze situations in which variables x and y exhibit such a linear relationship with randomness. The level of randomness will vary from situation to situation. In the introductory example connecting an electric current and the level of carbon monoxide in air, the relationship is almost perfect. In other situations, such as the height and weights of individuals, the connection between the two variables involves a high degree of randomness. In the next section we will see how to quantify the strength of the linear relationship between two variables.

KEY TAKEAWAYS

- Two variables x and y have a deterministic linear relationship if points plotted from (x, y) pairs lie exactly along a single straight line.
- In practice it is common for two variables to exhibit a relationship that is close to linear but which contains an element, possibly large, of randomness.

EXERCISES

BASIC

1. A line has equation $y = 0.5x + 2$.
   a. Pick five distinct x-values, use the equation to compute the corresponding y-values, and plot the five points obtained.
   b. Give the value of the slope of the line; give the value of the y-intercept.
2. A line has equation $y = x - 0.5$.
   a. Pick five distinct x-values, use the equation to compute the corresponding y-values, and plot the five points obtained.
   b. Give the value of the slope of the line; give the value of the y-intercept.
3. A line has equation $y = -2x + 4$.
   a. Pick five distinct x-values, use the equation to compute the corresponding y-values, and plot the five points obtained.
   b. Give the value of the slope of the line; give the value of the y-intercept.
4. A line has equation $y = -1.5x + 1$.
   a. Pick five distinct x-values, use the equation to compute the corresponding y-values, and plot the five points obtained.
   b. Give the value of the slope of the line; give the value of the y-intercept.
5. Based on the information given about a line, determine how y will change (increase, decrease, or stay the same) when x is increased, and explain. In some cases it might be impossible to tell from the information given.
   a. The slope is positive.
   b. The y-intercept is positive.
   c. The slope is zero.
6. Based on the information given about a line, determine how y will change (increase, decrease, or stay the same) when x is increased, and explain. In some cases it might be impossible to tell from the information given.
   a. The y-intercept is negative.
   b. The y-intercept is zero.
   c. The slope is negative.
7. A data set consists of eight (x, y) pairs of numbers:

$$(0,12) \quad (4,16) \quad (8,22) \quad (15,28)$$
$$(2,15) \quad (5,14) \quad (13,24) \quad (20,30)$$

a. Plot the data in a scatter diagram.
b. Based on the plot, explain whether the relationship between x and y appears to be deterministic or to involve randomness.
c. Based on the plot, explain whether the relationship between x and y appears to be linear or not linear.

8. A data set consists of ten (x, y) pairs of numbers:

$$(3,20) \quad (6,9) \quad (11,0) \quad (14,1) \quad (18,9)$$
$$(5,13) \quad (8,4) \quad (12,0) \quad (17,6) \quad (20,16)$$

a. Plot the data in a scatter diagram.
b. Based on the plot, explain whether the relationship between x and y appears to be deterministic or to involve randomness.
c. Based on the plot, explain whether the relationship between x and y appears to be linear or not linear.

9. A data set consists of nine (x, y) pairs of numbers:

$$(8,16) \quad (10,4) \quad (12,0) \quad (14,4) \quad (16,16)$$
$$(9,9) \quad (11,1) \quad (13,1) \quad (15,9)$$

a. Plot the data in a scatter diagram.
b. Based on the plot, explain whether the relationship between x and y appears to be deterministic or to involve randomness.
c. Based on the plot, explain whether the relationship between x and y appears to be linear or not linear.

10. A data set consists of five (x, y) pairs of numbers:

$$(0,1) \quad (2,5) \quad (3,7) \quad (5,11) \quad (8,17)$$

a. Plot the data in a scatter diagram.
b. Based on the plot, explain whether the relationship between x and y appears to be deterministic or to involve randomness.
c. Based on the plot, explain whether the relationship between x and y appears to be linear or not linear.

APPLICATIONS

11. At 60°F a particular blend of automotive gasoline weights 6.17 lb/gal. The weight y of gasoline on a tank truck that is loaded with x gallons of gasoline is given by the linear equation

$$y = 6.17x$$

    a. Explain whether the relationship between the weight y and the amount x of gasoline is deterministic or contains an element of randomness.
    b. Predict the weight of gasoline on a tank truck that has just been loaded with 6,750 gallons of gasoline.

12. The rate for renting a motor scooter for one day at a beach resort area is $25 plus 30 cents for each mile the scooter is driven. The total cost y in dollars for renting a scooter and driving it x miles is

$$y = 0.30x + 25$$

    a. Explain whether the relationship between the cost y of renting the scooter for a day and the distance x that the scooter is driven that day is deterministic or contains an element of randomness.
    b. A person intends to rent a scooter one day for a trip to an attraction 17 miles away. Assuming that the total distance the scooter is driven is 34 miles, predict the cost of the rental.

13. The pricing schedule for labor on a service call by an elevator repair company is $150 plus $50 per hour on site.

    a. Write down the linear equation that relates the labor cost y to the number of hours x that the repairman is on site.
    b. Calculate the labor cost for a service call that lasts 2.5 hours.

14. The cost of a telephone call made through a leased line service is 2.5 cents per minute.

    a. Write down the linear equation that relates the cost y (in cents) of a call to its length x.
    b. Calculate the cost of a call that lasts 23 minutes.

LARGE DATA SET EXERCISES

15. Large Data Set 1 lists the SAT scores and GPAs of 1,000 students. Plot the scatter diagram with SAT score as the independent variable (x) and GPA as the dependent variable (y). Comment on the appearance and strength of any linear trend.

http://www.gone.2012books.lardbucket.org/sites/all/files/data1.xls

16. Large Data Set 12 lists the golf scores on one round of golf for 75 golfers first using their own original clubs, then using clubs of a new, experimental design (after two months of familiarization with the new clubs). Plot the scatter diagram with golf score using the original clubs as the independent variable (x) and golf score using the new clubs as the dependent variable (y). Comment on the appearance and strength of any linear trend.

    http://www.gone.2012books.lardbucket.org/sites/all/files/data12.xls

17. Large Data Set 13 records the number of bidders and sales price of a particular type of antique grandfather clock at 60 auctions. Plot the scatter diagram with the number of bidders at the auction as the independent variable (x) and the sales price as the dependent variable (y). Comment on the appearance and strength of any linear trend.

    http://www.gone.2012books.lardbucket.org/sites/all/files/data13.xls

ANSWERS

1. a. Answers vary.
   b. Slope $m = 0.5$; y-intercept $b = 2$.

3. a. Answers vary.
   b. Slope $m = -2$; y-intercept $b = 4$.

5. a. y increases.
   b. Impossible to tell.
   c. y does not change.

7. a. Scatter diagram needed.
   b. Involves randomness.
   c. Linear.

9. a. Scatter diagram needed.
   b. Deterministic.
   c. Not linear.

11. a. Deterministic.
    b. 41,647.5 pounds.

13. a. $y = 50x + 150$.
    b. b. $275.

15. There appears to a hint of some positive correlation.

17. There appears to be clear positive correlation.

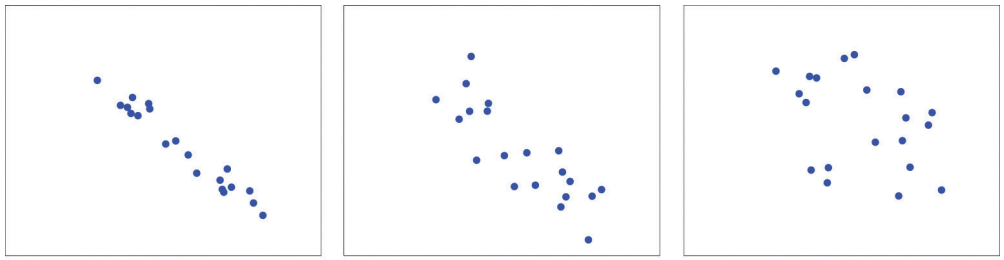
10.2 The Linear Correlation Coefficient

LEARNING OBJECTIVE

1. To learn what the linear correlation coefficient is, how to compute it, and what it tells us about the relationship between two variables x and y.

Figure 10.3 "Linear Relationships of Varying Strengths" illustrates linear relationships between two variables x and y of varying strengths. It is visually apparent that in the situation in panel (a), x could serve as a useful predictor of y, it would be less useful in the situation illustrated in panel (b), and in the situation of panel (c) the linear relationship is so weak as to be practically nonexistent. The *linear correlation coefficient* is a number computed directly from the data that measures the strength of the linear relationship between the two variables x and y.

Figure 10.3 *Linear Relationships of Varying Strengths*



Definition

The **linear correlation coefficient**[1] *for a collection of n pairs* (x, y) *of numbers in a sample is the number r given by the formula*

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} \cdot SS_{yy}}}$$

where

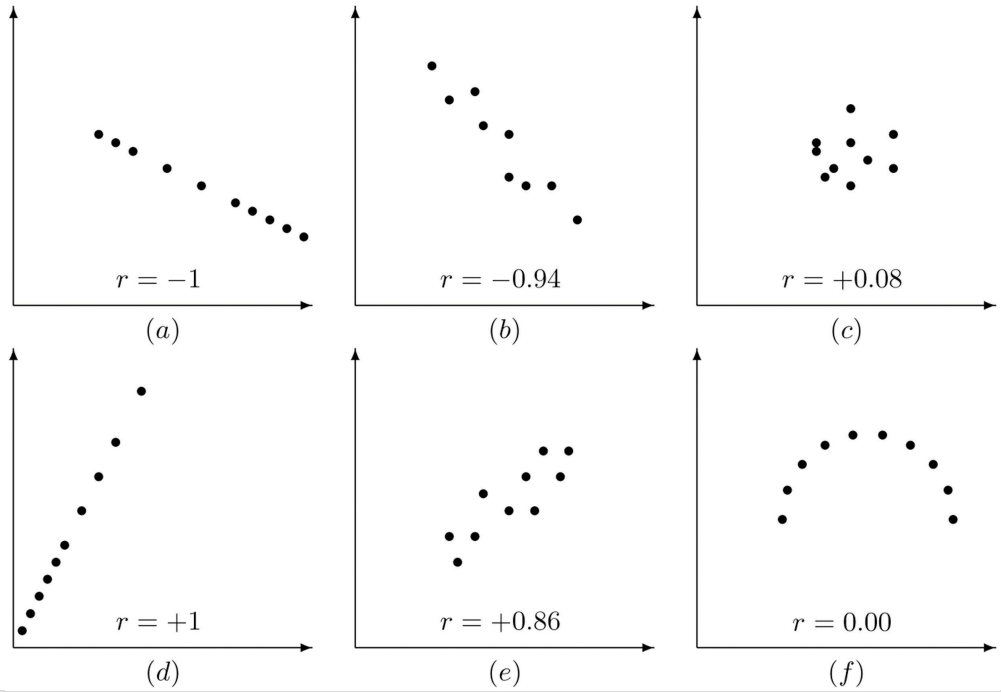
$$SS_{xx} = \Sigma x^2 - \frac{1}{n}(\Sigma x)^2, \quad SS_{xy} = \Sigma xy - \frac{1}{n}(\Sigma x)(\Sigma y), \quad SS_{yy} = \Sigma y^2 -$$

The linear correlation coefficient has the following properties, illustrated in Figure 10.4 "Linear Correlation Coefficient ":

1. The value of r lies between −1 and 1, inclusive.
2. The sign of r indicates the direction of the linear relationship between x and y:
    1. If $r < 0$ then y tends to decrease as x is increased.
    2. If $r > 0$ then y tends to increase as x is increased.
3. The size of $|r|$ indicates the strength of the linear relationship between x and y:
    1. If $|r|$ is near 1 (that is, if r is near either 1 or −1) then the linear relationship between x and y is strong.
    2. If $|r|$ is near 0 (that is, if r is near 0 and of either sign) then the linear relationship between x and y is weak.

1. A number computed directly from the data that measures the strength of the linear relationship between the two variables x and y.

Figure 10.4 *Linear Correlation Coefficient R*



Pay particular attention to panel (f) in Figure 10.4 "Linear Correlation Coefficient ". It shows a perfectly deterministic relationship between x and y, but $r = 0$ because the relationship is not linear. (In this particular case the points lie on the top half of a circle.)

EXAMPLE 1

Compute the linear correlation coefficient for the height and weight pairs plotted in Figure 10.2 "Plot of Height and Weight Pairs".

Solution:

Even for small data sets like this one computations are too long to do completely by hand. In actual practice the data are entered into a calculator or computer and a statistics program is used. In order to clarify the meaning of the formulas we will display the data and related quantities in tabular form. For each (x, y) pair we compute three numbers: x^2, xy, and y^2, as shown in the table provided. In the last line of the table we have the sum of the numbers in each column. Using them we compute:

	x	y	x^2	xy	y^2
	68	151	4624	10268	22801
	69	146	4761	10074	21316
	70	157	4900	10990	24649
	70	164	4900	11480	26896
	71	171	5041	12141	29241
	72	160	5184	11520	25600
	72	163	5184	11736	26569
	72	180	5184	12960	32400
	73	170	5329	12410	28900
	73	175	5329	12775	30625
	74	178	5476	13172	31684
	75	188	5625	14100	35344
Σ	859	2003	61537	143626	336025

$$SS_{xx} = \Sigma x^2 - \frac{1}{n}(\Sigma x)^2 = 61537 - \frac{1}{12}(859)^2 = 46.91\overline{6}$$

$$SS_{xy} = \Sigma xy - \frac{1}{n}(\Sigma x)\left(\Sigma y\right) = 143626 - \frac{1}{12}(859)(2003) = 244.58\overline{3}$$

$$SS_{yy} = \Sigma y^2 - \frac{1}{n}\left(\Sigma y\right)^2 = 336025 - \frac{1}{12}(2003)^2 = 1690.91\overline{6}$$

so that

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} = \frac{244.58\overline{3}}{\sqrt{(46.91\overline{6})(1690.91\overline{6})}} = 0.868$$

The number $r = 0.868$ quantifies what is visually apparent from Figure 10.2 "Plot of Height and Weight Pairs": weights tends to increase linearly with height (r is positive) and although the relationship is not perfect, it is reasonably strong (r is near 1).

KEY TAKEAWAYS

- The linear correlation coefficient measures the strength and direction of the linear relationship between two variables x and y.
- The sign of the linear correlation coefficient indicates the direction of the linear relationship between x and y.
- When r is near 1 or –1 the linear relationship is strong; when it is near 0 the linear relationship is weak.

EXERCISES

BASIC

With the exception of the exercises at the end of Section 10.3 "Modelling Linear Relationships with Randomness Present", the first Basic exercise in each of the following sections through Section 10.7 "Estimation and Prediction" uses the data from the first exercise here, the second Basic exercise uses the data from the second exercise here, and so on, and similarly for the Application exercises. Save your computations done on these exercises so that you do not need to repeat them later.

1. For the sample data

x	0	1	3	5	8
y	2	4	6	5	9

   a. Draw the scatter plot.
   b. Based on the scatter plot, predict the sign of the linear correlation coefficient. Explain your answer.
   c. Compute the linear correlation coefficient and compare its sign to your answer to part (b).

2. For the sample data

x	0	2	3	6	9
y	0	3	3	4	8

   a. Draw the scatter plot.
   b. Based on the scatter plot, predict the sign of the linear correlation coefficient. Explain your answer.
   c. Compute the linear correlation coefficient and compare its sign to your answer to part (b).

3. For the sample data

x	1	3	4	6	8
y	4	1	3	−1	0

   a. Draw the scatter plot.
   b. Based on the scatter plot, predict the sign of the linear correlation coefficient. Explain your answer.
   c. Compute the linear correlation coefficient and compare its sign to your answer to part (b).

4. For the sample data

x	1	2	4	7	9
y	5	5	6	−3	0

   a. Draw the scatter plot.
   b. Based on the scatter plot, predict the sign of the linear correlation coefficient. Explain your answer.
   c. Compute the linear correlation coefficient and compare its sign to your answer to part (b).

5. For the sample data

x	1	1	3	4	5
y	2	1	5	3	4

   a. Draw the scatter plot.
   b. Based on the scatter plot, predict the sign of the linear correlation coefficient. Explain your answer.
   c. Compute the linear correlation coefficient and compare its sign to your answer to part (b).

6. For the sample data

x	1	3	5	5	8
y	5	−2	2	−1	−3

   a. Draw the scatter plot.
   b. Based on the scatter plot, predict the sign of the linear correlation coefficient. Explain your answer.
   c. Compute the linear correlation coefficient and compare its sign to your answer to part (b).

7. Compute the linear correlation coefficient for the sample data summarized by the following information:

$$n = 5 \qquad \Sigma x = 25 \qquad \Sigma x^2 = 165$$

$$\Sigma y = 24 \quad \Sigma y^2 = 134 \quad \Sigma xy = 144$$

$$1 \le x \le 9$$

8. Compute the linear correlation coefficient for the sample data summarized by the following information:

$$n = 5 \quad \Sigma x = 31 \quad \Sigma x^2 = 253$$

$$\Sigma y = 18 \quad \Sigma y^2 = 90 \quad \Sigma xy = 148$$

$$2 \le x \le 12$$

9. Compute the linear correlation coefficient for the sample data summarized by the following information:

$$n = 10 \quad \Sigma x = 0 \quad \Sigma x^2 = 60$$

$$\Sigma y = 24 \quad \Sigma y^2 = 234 \quad \Sigma xy = -87$$

$$-4 \le x \le 4$$

10. Compute the linear correlation coefficient for the sample data summarized by the following information:

$$n = 10 \quad \Sigma x = -3 \quad \Sigma x^2 = 263$$

$$\Sigma y = 55 \quad \Sigma y^2 = 917 \quad \Sigma xy = -355$$

$$-10 \le x \le 10$$

APPLICATIONS

11. The age x in months and vocabulary y were measured for six children, with the results shown in the table.

x	13	14	15	16	16	18
y	8	10	15	20	27	30

Compute the linear correlation coefficient for these sample data and interpret its meaning in the context of the problem.

12. The curb weight x in hundreds of pounds and braking distance y in feet, at 50 miles per hour on dry pavement, were measured for five vehicles, with the results shown in the table.

x	25	27.5	32.5	35	45
y	105	125	140	140	150

Compute the linear correlation coefficient for these sample data and interpret its meaning in the context of the problem.

13. The age x and resting heart rate y were measured for ten men, with the results shown in the table.

x	20	23	30	37	35
y	72	71	73	74	74
x	45	51	55	60	63
y	73	72	79	75	77

Compute the linear correlation coefficient for these sample data and interpret its meaning in the context of the problem.

14. The wind speed x in miles per hour and wave height y in feet were measured under various conditions on an enclosed deep water sea, with the results shown in the table,

x	0	0	2	7	7
y	2.0	0.0	0.3	0.7	3.3
x	9	13	20	22	31
y	4.9	4.9	3.0	6.9	5.9

Compute the linear correlation coefficient for these sample data and interpret its meaning in the context of the problem.

15. The advertising expenditure x and sales y in thousands of dollars for a small retail business in its first eight years in operation are shown in the table.

x	1.4	1.6	1.6	2.0
y	180	184	190	220
x	2.0	2.2	2.4	2.6
y	186	215	205	240

Compute the linear correlation coefficient for these sample data and interpret its meaning in the context of the problem.

16. The height x at age 2 and y at age 20, both in inches, for ten women are tabulated in the table.

x	31.3	31.7	32.5	33.5	34.4
y	60.7	61.0	63.1	64.2	65.9
x	35.2	35.8	32.7	33.6	34.8
y	68.2	67.6	62.3	64.9	66.8

Compute the linear correlation coefficient for these sample data and interpret its meaning in the context of the problem.

17. The course average x just before a final exam and the score y on the final exam were recorded for 15 randomly selected students in a large physics class, with the results shown in the table.

x	69.3	87.7	50.5	51.9	82.7
y	56	89	55	49	61
x	70.5	72.4	91.7	83.3	86.5
y	66	72	83	73	82
x	79.3	78.5	75.7	52.3	62.2
y	92	80	64	18	76

Compute the linear correlation coefficient for these sample data and interpret its meaning in the context of the problem.

18. The table shows the acres x of corn planted and acres y of corn harvested, in millions of acres, in a particular country in ten successive years.

x	75.7	78.9	78.6	80.9	81.8
y	68.8	69.3	70.9	73.6	75.1
x	78.3	93.5	85.9	86.4	88.2
y	70.6	86.5	78.6	79.5	81.4

Compute the linear correlation coefficient for these sample data and interpret its meaning in the context of the problem.

19. Fifty male subjects drank a measured amount x (in ounces) of a medication and the concentration y (in percent) in their blood of the active ingredient was measured 30 minutes later. The sample data are summarized by the following information.

$$n = 50 \quad \Sigma x = 112.5 \quad \Sigma y = 4.83$$
$$\Sigma xy = 15.255 \quad 0 \le x \le 4.5$$
$$\Sigma x^2 = 356.25 \quad \Sigma y^2 = 0.667$$

Compute the linear correlation coefficient for these sample data and interpret its meaning in the context of the problem.

20. In an effort to produce a formula for estimating the age of large free-standing oak trees non-invasively, the girth x (in inches) five feet off the ground of 15 such trees of known age y (in years) was measured. The sample data are summarized by the following information.

$$n = 15 \quad \Sigma x = 3368 \quad \Sigma y = 6496$$
$$\Sigma xy = 1{,}933{,}219 \quad \Sigma x^2 = 917{,}780$$
$$\Sigma y^2 = 4{,}260{,}666 \quad 74 \le x \le 395$$

Compute the linear correlation coefficient for these sample data and interpret its meaning in the context of the problem.

21. Construction standards specify the strength of concrete 28 days after it is poured. For 30 samples of various types of concrete the strength x after 3 days and the strength y after 28 days (both in hundreds of pounds per square inch) were measured. The sample data are summarized by the following information.

$$n = 30 \quad \Sigma x = 501.6 \quad \Sigma y = 1338.8$$
$$\Sigma xy = 23{,}246.55 \quad \Sigma x^2 = 8724.74$$
$$\Sigma y^2 = 61{,}980.14 \quad 11 \le x \le 22$$

Compute the linear correlation coefficient for these sample data and interpret its meaning in the context of the problem.

22. Power-generating facilities used forecasts of temperature to forecast energy demand. The average temperature x (degrees Fahrenheit) and the day's energy demand y (million watt-hours) were recorded on 40 randomly selected winter days in the region served by a power company. The sample data are summarized by the following information.

$$n = 40 \quad \Sigma x = 2000 \quad \Sigma y = 2969$$
$$\Sigma xy = 143{,}042 \quad \Sigma x^2 = 101{,}340$$
$$\Sigma y^2 = 243{,}027 \quad 40 \le x \le 60$$

Compute the linear correlation coefficient for these sample data and interpret its meaning in the context of the problem.

ADDITIONAL EXERCISES

23. In each case state whether you expect the two variables x and y indicated to have positive, negative, or zero correlation.

    a. the number x of pages in a book and the age y of the author
    b. the number x of pages in a book and the age y of the intended reader
    c. the weight x of an automobile and the fuel economy y in miles per gallon
    d. the weight x of an automobile and the reading y on its odometer

e. the amount x of a sedative a person took an hour ago and the time y it takes him to respond to a stimulus

24. In each case state whether you expect the two variables x and y indicated to have positive, negative, or zero correlation.

    a. the length x of time an emergency flare will burn and the length y of time the match used to light it burned
    b. the average length x of time that calls to a retail call center are on hold one day and the number y of calls received that day
    c. the length x of a regularly scheduled commercial flight between two cities and the headwind y encountered by the aircraft
    d. the value x of a house and the its size y in square feet
    e. the average temperature x on a winter day and the energy consumption y of the furnace

25. Changing the units of measurement on two variables x and y should not change the linear correlation coefficient. Moreover, most change of units amount to simply multiplying one unit by the other (for example, 1 foot = 12 inches). Multiply each x value in the table in Exercise 1 by two and compute the linear correlation coefficient for the new data set. Compare the new value of r to the one for the original data.

26. Refer to the previous exercise. Multiply each x value in the table in Exercise 2 by two, multiply each y value by three, and compute the linear correlation coefficient for the new data set. Compare the new value of r to the one for the original data.

27. Reversing the roles of x and y in the data set of Exercise 1 produces the data set

x	2	4	6	5	9
y	0	1	3	5	8

    Compute the linear correlation coefficient of the new set of data and compare it to what you got in Exercise 1.

28. In the context of the previous problem, look at the formula for r and see if you can tell why what you observed there must be true for every data set.

LARGE DATA SET EXERCISES

29. Large Data Set 1 lists the SAT scores and GPAs of 1,000 students. Compute the linear correlation coefficient r. Compare its value to your comments on the appearance and strength of any linear trend in the scatter diagram that you

constructed in the first large data set problem for Section 10.1 "Linear Relationships Between Variables".

http://www.gone.2012books.lardbucket.org/sites/all/files/data1.xls

30. Large Data Set 12 lists the golf scores on one round of golf for 75 golfers first using their own original clubs, then using clubs of a new, experimental design (after two months of familiarization with the new clubs). Compute the linear correlation coefficient *r*. Compare its value to your comments on the appearance and strength of any linear trend in the scatter diagram that you constructed in the second large data set problem for Section 10.1 "Linear Relationships Between Variables".

    http://www.gone.2012books.lardbucket.org/sites/all/files/data12.xls

31. Large Data Set 13 records the number of bidders and sales price of a particular type of antique grandfather clock at 60 auctions. Compute the linear correlation coefficient *r*. Compare its value to your comments on the appearance and strength of any linear trend in the scatter diagram that you constructed in the third large data set problem for Section 10.1 "Linear Relationships Between Variables".

    http://www.gone.2012books.lardbucket.org/sites/all/files/data13.xls

ANSWERS

1. $r = 0.921$
3. $r = -0.794$
5. $r = 0.707$
7. 0.875
9. –0.846
11. 0.948
13. 0.709
15. 0.832
17. 0.751
19. 0.965
21. 0.992
23.
    a. zero
    b. positive
    c. negative
    d. zero
    e. positive
25. same value
27. same value
29. $r = 0.4601$
31. $r = 0.9002$

10.3 Modelling Linear Relationships with Randomness Present

LEARNING OBJECTIVE

1. To learn the framework in which the statistical analysis of the linear relationship between two variables *x* and *y* will be done.

In this chapter we are dealing with a population for which we can associate to each element two measurements, *x* and *y*. We are interested in situations in which the value of *x* can be used to draw conclusions about the value of *y*, such as predicting the resale value *y* of a residential house based on its size *x*. Since the relationship between *x* and *y* is not deterministic, statistical procedures must be applied. For any statistical procedures, given in this book or elsewhere, the associated formulas are valid only under specific assumptions. The set of assumptions in simple linear regression are a mathematical description of the relationship between *x* and *y*. Such a set of assumptions is known as a **model**.

For each fixed value of *x* a sub-population of the full population is determined, such as the collection of all houses with 2,100 square feet of living space. For each element of that sub-population there is a measurement *y*, such as the value of any 2,100-square-foot house. Let $E\left(y\right)$ denote the mean of all the *y*-values for each particular value of *x*. $E\left(y\right)$ can change from *x*-value to *x*-value, such as the mean value of all 2,100-square-foot houses, the (different) mean value for all 2,500-square foot-houses, and so on.

Our first assumption is that the relationship between *x* and the mean of the *y*-values in the sub-population determined by *x* is linear. This means that there exist numbers β_1 and β_0 such that

$$E\left(y\right) = \beta_1 x + \beta_0$$

This linear relationship is the reason for the word “linear” in “simple linear regression” below. (The word “simple” means that *y* depends on only one other variable and not two or more.)

Our next assumption is that for each value of *x* the *y*-values scatter about the mean $E\left(y\right)$ according to a normal distribution centered at $E\left(y\right)$ and with a standard deviation σ that is the same for every value of *x*. This is the same as saying that

there exists a normally distributed random variable ε with mean 0 and standard deviation σ so that the relationship between x and y in the whole population is

$$y = \beta_1 x + \beta_0 + \varepsilon$$

Our last assumption is that the random deviations associated with different observations are independent.

In summary, the model is:

Simple Linear Regression Model

For each point (x, y) in data set the y-value is an independent observation of

$$y = \beta_1 x + \beta_0 + \varepsilon$$

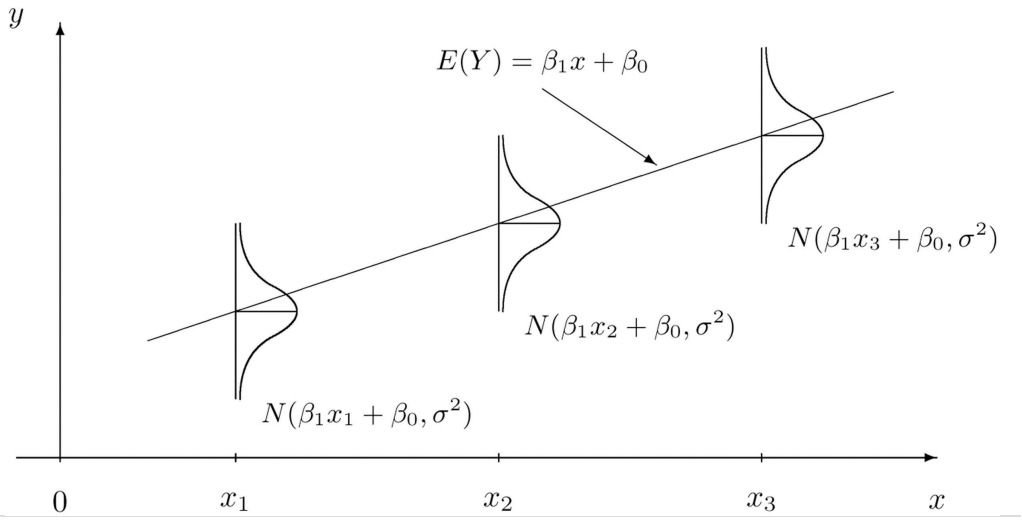
where β_1 and β_0 are fixed parameters and ε is a normally distributed random variable with mean 0 and an unknown standard deviation σ.

The line with equation $y = \beta_1 x + \beta_0$ is called the **population regression line**[2].

Figure 10.5 "The Simple Linear Model Concept" illustrates the model. The symbols $N(\mu, \sigma^2)$ denote a normal distribution with mean μ and variance σ^2, hence standard deviation σ.

2. The line with equation $y = \beta_1 x + \beta_0$ that gives the mean of the variable y over the sub-population determined by x.

Figure 10.5 The Simple Linear Model Concept



It is conceptually important to view the model as a sum of two parts:

$$y = \boxed{\beta_1 x + \beta_0} + \boxed{\varepsilon}$$

1. **Deterministic Part.** The first part $\beta_1 x + \beta_0$ is the equation that describes the trend in y as x increases. The line that we seem to see when we look at the scatter diagram is an approximation of the line $y = \beta_1 x + \beta_0$. There is nothing random in this part, and therefore it is called the *deterministic* part of the model.
2. **Random Part.** The second part ε is a random variable, often called the *error* term or the *noise*. This part explains why the actual observed values of y are not exactly on but fluctuate near a line. Information about this term is important since only when one knows how much noise there is in the data can one know how trustworthy the detected trend is.

There are three parameters in this model: β_0, β_1, and σ. Each has an important interpretation, particularly β_1 and σ. The slope parameter β_1 represents the expected change in y brought about by a unit increase in x. The standard deviation σ represents the magnitude of the noise in the data.

There are procedures for checking the validity of the three assumptions, but for us it will be sufficient to visually verify the linear trend in the data. If the data set is large then the points in the scatter diagram will form a band about an apparent straight line. The normality of ε with a constant standard deviation corresponds

graphically to the band being of roughly constant width, and with most points concentrated near the middle of the band.

Fortunately, the three assumptions do not need to hold exactly in order for the procedures and analysis developed in this chapter to be useful.

KEY TAKEAWAY

- Statistical procedures are valid only when certain assumptions are valid. The assumptions underlying the analyses done in this chapter are graphically summarized in Figure 10.5 "The Simple Linear Model Concept".

EXERCISES

1. State the three assumptions that are the basis for the Simple Linear Regression Model.
2. The Simple Linear Regression Model is summarized by the equation
$$y = \beta_1 x + \beta_0 + \varepsilon$$
Identify the deterministic part and the random part.
3. Is the number β_1 in the equation $y = \beta_1 x + \beta_0$ a statistic or a population parameter? Explain.
4. Is the number σ in the Simple Linear Regression Model a statistic or a population parameter? Explain.
5. Describe what to look for in a scatter diagram in order to check that the assumptions of the Simple Linear Regression Model are true.
6. True or false: the assumptions of the Simple Linear Regression Model must hold exactly in order for the procedures and analysis developed in this chapter to be useful.

ANSWERS

1. a. The mean of y is linearly related to x.
   b. For each given x, y is a normal random variable with mean $\beta_1 x + \beta_0$ and standard deviation σ.
   c. All the observations of y in the sample are independent.

3. β_1 is a population parameter.

5. A linear trend.

10.4 The Least Squares Regression Line

LEARNING OBJECTIVES

1. To learn how to measure how well a straight line fits a collection of data.
2. To learn how to construct the least squares regression line, the straight line that best fits a collection of data.
3. To learn the meaning of the slope of the least squares regression line.
4. To learn how to use the least squares regression line to estimate the response variable *y* in terms of the predictor variable *x*.

Goodness of Fit of a Straight Line to Data

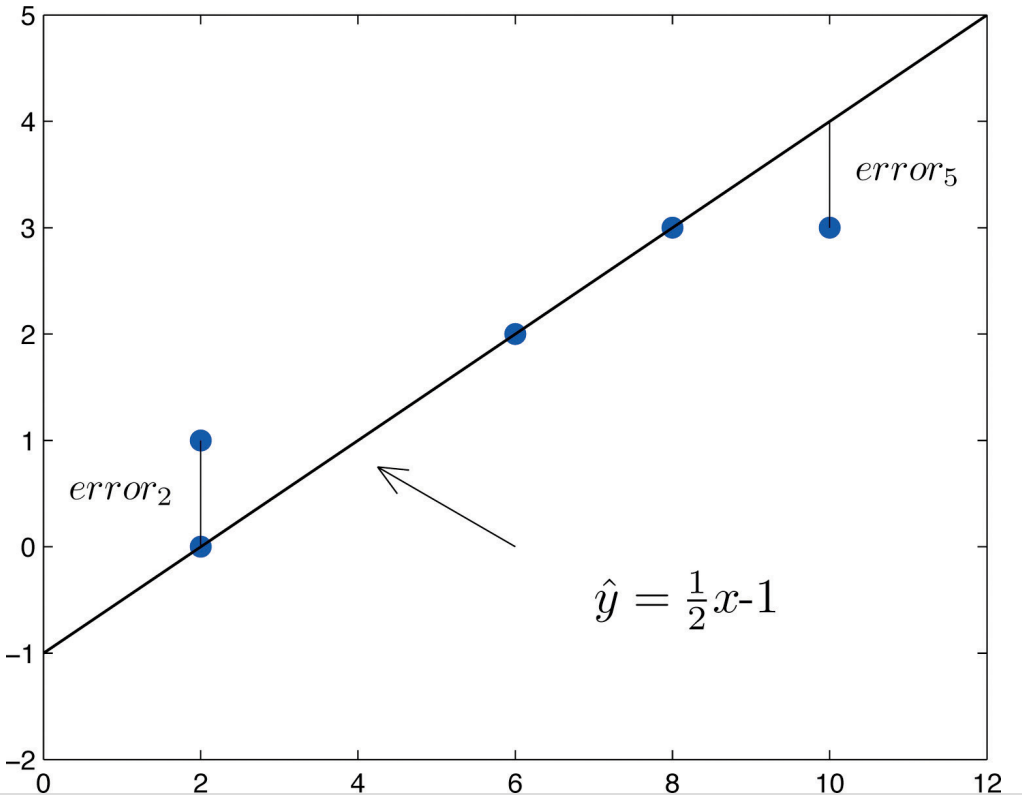
Once the scatter diagram of the data has been drawn and the model assumptions described in the previous sections at least visually verified (and perhaps the correlation coefficient *r* computed to quantitatively verify the linear trend), the next step in the analysis is to find the straight line that best fits the data. We will explain how to measure how well a straight line fits a collection of points by examining how well the line $y = \frac{1}{2}x - 1$ fits the data set

x	2	2	6	8	10
y	0	1	2	3	3

(which will be used as a running example for the next three sections). We will write the equation of this line as $\hat{y} = \frac{1}{2}x - 1$ with an accent on the *y* to indicate that the *y*-values computed using this equation are not from the data. We will do this with all lines approximating data sets. The line $\hat{y} = \frac{1}{2}x - 1$ was selected as one that seems to fit the data reasonably well.

The idea for measuring the goodness of fit of a straight line to data is illustrated in Figure 10.6 "Plot of the Five-Point Data and the Line ", in which the graph of the line $\hat{y} = \frac{1}{2}x - 1$ has been superimposed on the scatter plot for the sample data set.

Figure 10.6 Plot of the Five-Point Data and the Line $\hat{y} = \frac{1}{2}x - 1$



To each point in the data set there is associated an "**error**[3]," the positive or negative vertical distance from the point to the line: positive if the point is above the line and negative if it is below the line. The error can be computed as the actual y-value of the point minus the y-value $\hat{y}$ that is "predicted" by inserting the x-value of the data point into the formula for the line:

$$\text{error at data point } (x, y) = (\text{true } y) - (\text{predicted } y) = y - \hat{y}$$

The computation of the error for each of the five points in the data set is shown in Table 10.1 "The Errors in Fitting Data with a Straight Line".

Table 10.1 The Errors in Fitting Data with a Straight Line

	x	y	$\hat{y} = \frac{1}{2}x - 1$	$y - \hat{y}$	$(y - \hat{y})^2$
	2	0	0	0	0
	2	1	0	1	1

3. Using $y - \hat{y}$, the actual y-value of a data point minus the y-value that is computed from the equation of the line fitting the data.

	x	y	$\hat{y} = \frac{1}{2}x - 1$	$y - \hat{y}$	$(y - \hat{y})^2$
	6	2	2	0	0
	8	3	3	0	0
	10	3	4	−1	1
Σ	-	-	-	0	2

A first thought for a measure of the goodness of fit of the line to the data would be simply to add the errors at every point, but the example shows that this cannot work well in general. The line does not fit the data perfectly (no line can), yet because of cancellation of positive and negative errors the sum of the errors (the fourth column of numbers) is zero. Instead goodness of fit is measured by the sum of the squares of the errors. Squaring eliminates the minus signs, so no cancellation can occur. For the data and line in Figure 10.6 "Plot of the Five-Point Data and the Line " the sum of the squared errors (the last column of numbers) is 2. This number measures the goodness of fit of the line to the data.

Definition

The **goodness of fit** *of a line* $\hat{y} = mx + b$ *to a set of n pairs* (x, y) *of numbers in a sample is the sum of the squared errors*

$$\Sigma(y - \hat{y})^2$$

(n terms in the sum, one for each data pair).

The Least Squares Regression Line

Given any collection of pairs of numbers (except when all the x-values are the same) and the corresponding scatter diagram, there always exists exactly one straight line that fits the data better than any other, in the sense of minimizing the sum of the squared errors. It is called the *least squares regression line*. Moreover there are formulas for its slope and y-intercept.

Definition

Given a collection of pairs (x, y) *of numbers (in which not all the x-values are the same), there is a line* $\widehat{y} = \widehat{\beta}_1 x + \widehat{\beta}_0$ *that best fits the data in the sense of minimizing the sum of the squared errors. It is called the* **least squares regression line**[4]. *Its slope* $\widehat{\beta}_1$ *and y-intercept* $\widehat{\beta}_0$ *are computed using the formulas*

$$\widehat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \quad \text{and} \quad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$

where

$$SS_{xx} = \Sigma x^2 - \frac{1}{n}(\Sigma x)^2, \quad SS_{xy} = \Sigma xy - \frac{1}{n}(\Sigma x)(\Sigma y)$$

$\overline{x}$ *is the mean of all the x-values,* $\overline{y}$ *is the mean of all the y-values, and n is the number of pairs in the data set.*

The equation $\widehat{y} = \widehat{\beta}_1 x + \widehat{\beta}_0$ *specifying the least squares regression line is called the* **least squares regression equation**[5].

Remember from Section 10.3 "Modelling Linear Relationships with Randomness Present" that the line with the equation $y = \beta_1 x + \beta_0$ is called the population regression line. The numbers $\widehat{\beta}_1$ and $\widehat{\beta}_0$ are statistics that estimate the population parameters β_1 and β_0.

We will compute the least squares regression line for the five-point data set, then for a more practical example that will be another running example for the introduction of new concepts in this and the next three sections.

4. The line that best fits a set of sample data in the sense of minimizing the sum of the squared errors.

5. The equation $\widehat{y} = \widehat{\beta}_1 x + \widehat{\beta}_0$ of the least squares regression line.

EXAMPLE 2

Find the least squares regression line for the five-point data set

x	2	2	6	8	10
y	0	1	2	3	3

and verify that it fits the data better than the line $\hat{y} = \frac{1}{2}x - 1$ considered in Section 10.4.1 "Goodness of Fit of a Straight Line to Data".

Solution:

In actual practice computation of the regression line is done using a statistical computation package. In order to clarify the meaning of the formulas we display the computations in tabular form.

	x	y	x^2	xy
	2	0	4	0
	2	1	4	2
	6	2	36	12
	8	3	64	24
	10	3	100	30
Σ	28	9	208	68

In the last line of the table we have the sum of the numbers in each column. Using them we compute:

$$SS_{xx} = \Sigma x^2 - \frac{1}{n}(\Sigma x)^2 = 208 - \frac{1}{5}(28)^2 = 51.2$$

$$SS_{xy} = \Sigma xy - \frac{1}{n}(\Sigma x)(\Sigma y) = 68 - \frac{1}{5}(28)(9) = 17.6$$

$$\overline{x} = \frac{\Sigma x}{n} = \frac{28}{5} = 5.6$$

$$\overline{y} = \frac{\Sigma y}{n} = \frac{9}{5} = 1.8$$

so that

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{17.6}{51.2} = 0.34375 \quad \text{and} \quad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 1.8 - (0.34375)(5$$

The least squares regression line for these data is

$$\hat{y} = 0.34375x - 0.125$$

The computations for measuring how well it fits the sample data are given in Table 10.2 "The Errors in Fitting Data with the Least Squares Regression Line". The sum of the squared errors is the sum of the numbers in the last column, which is 0.75. It is less than 2, the sum of the squared errors for the fit of the line $\hat{y} = \frac{1}{2}x - 1$ to this data set.

TABLE 10.2 THE ERRORS IN FITTING DATA WITH THE LEAST SQUARES REGRESSION LINE

x	y	$\hat{y} = 0.34375x - 0.125$	$y - \hat{y}$	$(y - \hat{y})^2$
2	0	0.5625	−0.5625	0.31640625
2	1	0.5625	0.4375	0.19140625
6	2	1.9375	0.0625	0.00390625
8	3	2.6250	0.3750	0.14062500
10	3	3.3125	−0.3125	0.09765625

EXAMPLE 3

Table 10.3 "Data on Age and Value of Used Automobiles of a Specific Make and Model" shows the age in years and the retail value in thousands of dollars of a random sample of ten automobiles of the same make and model.

a. Construct the scatter diagram.
b. Compute the linear correlation coefficient r. Interpret its value in the context of the problem.
c. Compute the least squares regression line. Plot it on the scatter diagram.
d. Interpret the meaning of the slope of the least squares regression line in the context of the problem.
e. Suppose a four-year-old automobile of this make and model is selected at random. Use the regression equation to predict its retail value.
f. Suppose a 20-year-old automobile of this make and model is selected at random. Use the regression equation to predict its retail value. Interpret the result.
g. Comment on the validity of using the regression equation to predict the price of a brand new automobile of this make and model.

TABLE 10.3 DATA ON AGE AND VALUE OF USED AUTOMOBILES OF A SPECIFIC MAKE AND MODEL

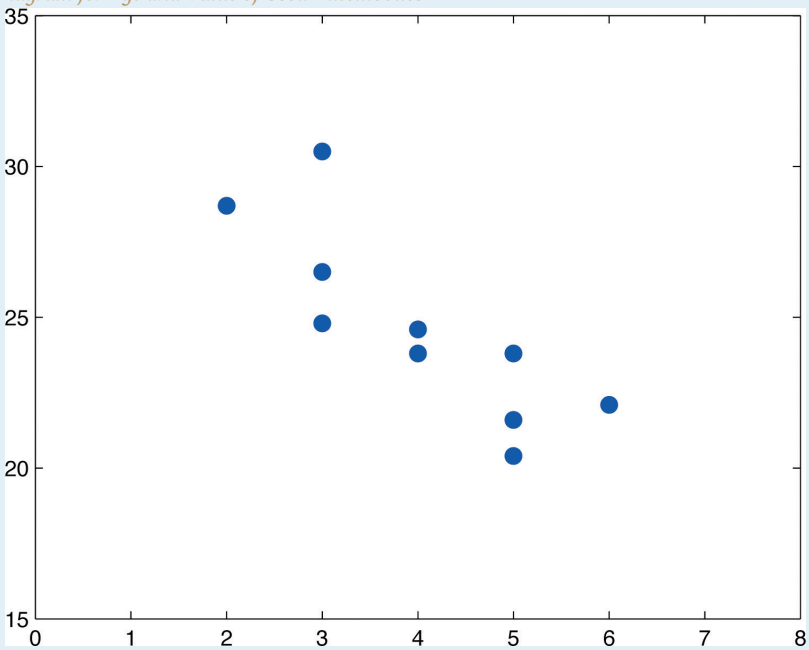
x	2	3	3	3	4	4	5	5	5	6
y	28.7	24.8	26.0	30.5	23.8	24.6	23.8	20.4	21.6	22.1

Solution:

a. The scatter diagram is shown in Figure 10.7 "Scatter Diagram for Age and Value of Used Automobiles".

Figure 10.7

Scatter Diagram for Age and Value of Used Automobiles



a. We must first compute SS_{xx}, SS_{xy}, SS_{yy}, which means computing Σx, Σy, Σx^2, Σy^2, and Σxy. Using a computing device we obtain

$$\Sigma x = 40 \quad \Sigma y = 246.3 \quad \Sigma x^2 = 174 \quad \Sigma y^2 = 6154.15 \quad \Sigma xy = 956.5$$

Thus

$$SS_{xx} = \Sigma x^2 - \frac{1}{n}(\Sigma x)^2 = 174 - \frac{1}{10}(40)^2 = 14$$

$$SS_{xy} = \Sigma xy - \frac{1}{n}(\Sigma x)(\Sigma y) = 956.5 - \frac{1}{10}(40)(246.3) = -28.7$$

$$SS_{yy} = \Sigma y^2 - \frac{1}{n}(\Sigma y)^2 = 6154.15 - \frac{1}{10}(246.3)^2 = 87.781$$

so that

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} \cdot SS_{yy}}} = \frac{-28.7}{\sqrt{(14)(87.781)}} = -0.819$$

The age and value of this make and model automobile are moderately strongly negatively correlated. As the age increases, the value of the automobile tends to decrease.

b. Using the values of Σx and Σy computed in part (b),

$$\overline{x} = \frac{\Sigma x}{n} = \frac{40}{10} = 4 \quad \text{and} \quad \overline{y} = \frac{\Sigma y}{n} = \frac{246.3}{10} = 24.63$$

Thus using the values of SS_{xx} and SS_{xy} from part (b),

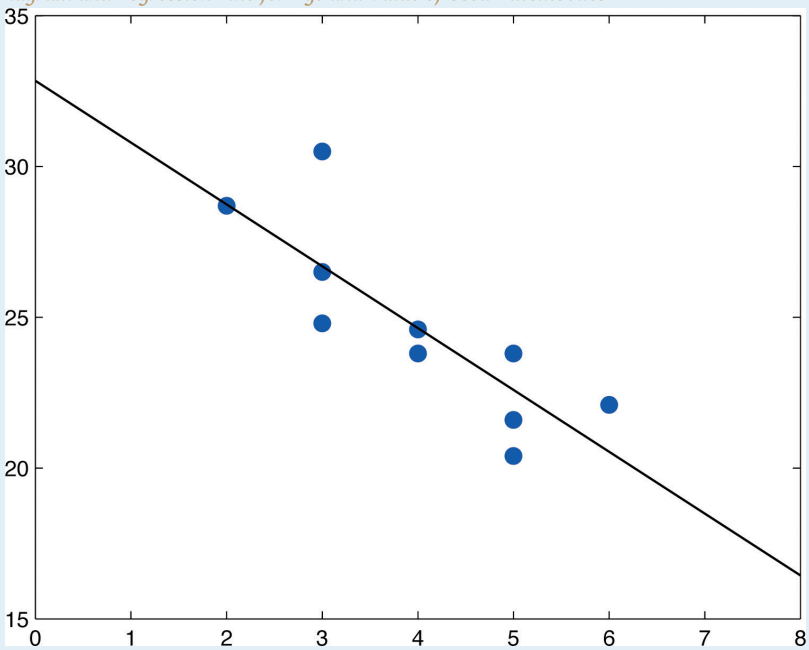
$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{-28.7}{14} = -2.05 \quad \text{and} \quad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 24.63 - (-2$$

The equation $\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0$ of the least squares regression line for these sample data is

$$\hat{y} = -2.05x + 32.83$$

Figure 10.8 "Scatter Diagram and Regression Line for Age and Value of Used Automobiles" shows the scatter diagram with the graph of the least squares regression line superimposed.

Figure 10.8
Scatter Diagram and Regression Line for Age and Value of Used Automobiles



a. The slope –2.05 means that for each unit increase in x (additional year of age) the average value of this make and model vehicle decreases by about 2.05 units (about $2,050).

b. Since we know nothing about the automobile other than its age, we assume that it is of about average value and use the average value of all four-year-old vehicles of this make and model as our estimate. The average value is simply the value of $\hat{y}$ obtained when the number 4 is inserted for x in the least squares regression equation:

$$\hat{y} = -2.05\,(4) + 32.83 = 24.63$$

which corresponds to $24,630.

c. Now we insert $x = 20$ into the least squares regression equation, to obtain

$$\hat{y} = -2.05\,(20) + 32.83 = -8.17$$

which corresponds to –$8,170. Something is wrong here, since a negative makes no sense. The error arose from applying the

regression equation to a value of x not in the range of x-values in the original data, from two to six years.

Applying the regression equation $\widehat{y} = \widehat{\beta}_1 x + \widehat{\beta}_0$ to a value of x outside the range of x-values in the data set is called *extrapolation*. It is an invalid use of the regression equation and should be avoided.

d. The price of a brand new vehicle of this make and model is the value of the automobile at age 0. If the value $x = 0$ is inserted into the regression equation the result is always $\widehat{\beta}_0$, the y-intercept, in this case 32.83, which corresponds to \$32,830. But this is a case of extrapolation, just as part (f) was, hence this result is invalid, although not obviously so. In the context of the problem, since automobiles tend to lose value much more quickly immediately after they are purchased than they do after they are several years old, the number \$32,830 is probably an underestimate of the price of a new automobile of this make and model.

For emphasis we highlight the points raised by parts (f) and (g) of the example.

Definition

The process of using the least squares regression equation to estimate the value of y at a value of x that does not lie in the range of the x-values in the data set that was used to form the regression line is called **extrapolation**[6]*. It is an invalid use of the regression equation that can lead to errors, hence should be avoided.*

The Sum of the Squared Errors SSE

In general, in order to measure the goodness of fit of a line to a set of data, we must compute the predicted y-value $\widehat{y}$ at every point in the data set, compute each error, square it, and then add up all the squares. In the case of the least squares regression line, however, the line that best fits the data, the sum of the squared errors can be computed directly from the data using the following formula.

6. The process of using the least squares regression equation to estimate the value of y at an x value not in the proper range.

The sum of the squared errors for the least squares regression line is denoted by SSE. It can be computed using the formula

$$SSE = SS_{yy} - \widehat{\beta}_1 SS_{xy}$$

EXAMPLE 4

Find the sum of the squared errors SSE for the least squares regression line for the five-point data set

x	2	2	6	8	10
y	0	1	2	3	3

Do so in two ways:

a. using the definition $\Sigma\left(y-\hat{y}\right)^2$;
b. using the formula $SSE = SS_{yy} - \hat{\beta}_1 SS_{xy}$.

Solution:

a. The least squares regression line was computed in Note 10.18 "Example 2" and is $\hat{y} = 0.34375x - 0.125$. SSE was found at the end of that example using the definition $\Sigma\left(y-\hat{y}\right)^2$. The computations were tabulated in Table 10.2 "The Errors in Fitting Data with the Least Squares Regression Line". SSE is the sum of the numbers in the last column, which is 0.75.

b. The numbers SS_{xy} and $\hat{\beta}_1$ were already computed in Note 10.18 "Example 2" in the process of finding the least squares regression line. So was the number $\Sigma y = 9$. We must compute SS_{yy}. To do so it is necessary to first compute $\Sigma y^2 = 0 + 1^2 + 2^2 + 3^2 + 3^2 = 23$. Then

$$SS_{yy} = \Sigma y^2 - \frac{1}{n}\left(\Sigma y\right)^2 = 23 - \frac{1}{5}(9)^2 = 6.8$$

so that

$$SSE = SS_{yy} - \hat{\beta}_1 SS_{xy} = 6.8 - (0.34375)(17.6) = 0.75$$

EXAMPLE 5

Find the sum of the squared errors SSE for the least squares regression line for the data set, presented in Table 10.3 "Data on Age and Value of Used Automobiles of a Specific Make and Model", on age and values of used vehicles in Note 10.19 "Example 3".

Solution:

From Note 10.19 "Example 3" we already know that

$$SS_{xy} = -28.7, \quad \widehat{\beta}_1 = -2.05, \quad \text{and} \quad \Sigma y = 246.3$$

To compute SS_{yy} we first compute

$$\Sigma y^2 = 28.7^2 + 24.8^2 + 26.0^2 + 30.5^2 + 23.8^2 + 24.6^2 + 23.8^2 + 20.4^2$$

Then

$$SS_{yy} = \Sigma y^2 - \frac{1}{n}(\Sigma y)^2 = 6154.15 - \frac{1}{10}(246.3)^2 = 87.781$$

Therefore

$$SSE = SS_{yy} - \widehat{\beta}_1 SS_{xy} = 87.781 - (-2.05)(-28.7) = 28.946$$

KEY TAKEAWAYS

- How well a straight line fits a data set is measured by the sum of the squared errors.
- The least squares regression line is the line that best fits the data. Its slope and y-intercept are computed from the data using formulas.
- The slope $\widehat{\beta}_1$ of the least squares regression line estimates the size and direction of the mean change in the dependent variable y when the independent variable x is increased by one unit.
- The sum of the squared errors SSE of the least squares regression line can be computed using a formula, without having to compute all the individual errors.

EXERCISES

BASIC

For the Basic and Application exercises in this section use the computations that were done for the exercises with the same number in Section 10.2 "The Linear Correlation Coefficient".

1. Compute the least squares regression line for the data in Exercise 1 of Section 10.2 "The Linear Correlation Coefficient".
2. Compute the least squares regression line for the data in Exercise 2 of Section 10.2 "The Linear Correlation Coefficient".
3. Compute the least squares regression line for the data in Exercise 3 of Section 10.2 "The Linear Correlation Coefficient".
4. Compute the least squares regression line for the data in Exercise 4 of Section 10.2 "The Linear Correlation Coefficient".
5. For the data in Exercise 5 of Section 10.2 "The Linear Correlation Coefficient"
   a. Compute the least squares regression line.
   b. Compute the sum of the squared errors SSE using the definition $\Sigma(y - \hat{y})^2$.
   c. Compute the sum of the squared errors SSE using the formula $SSE = SS_{yy} - \hat{\beta}_1 SS_{xy}$.
6. For the data in Exercise 6 of Section 10.2 "The Linear Correlation Coefficient"
   a. Compute the least squares regression line.
   b. Compute the sum of the squared errors SSE using the definition $\Sigma(y - \hat{y})^2$.
   c. Compute the sum of the squared errors SSE using the formula $SSE = SS_{yy} - \hat{\beta}_1 SS_{xy}$.
7. Compute the least squares regression line for the data in Exercise 7 of Section 10.2 "The Linear Correlation Coefficient".
8. Compute the least squares regression line for the data in Exercise 8 of Section 10.2 "The Linear Correlation Coefficient".
9. For the data in Exercise 9 of Section 10.2 "The Linear Correlation Coefficient"
   a. Compute the least squares regression line.

b. Can you compute the sum of the squared errors SSE using the definition $\Sigma(y - \hat{y})^2$? Explain.
c. Compute the sum of the squared errors SSE using the formula $SSE = SS_{yy} - \hat{\beta}_1 SS_{xy}$.

10. For the data in Exercise 10 of Section 10.2 "The Linear Correlation Coefficient"

a. Compute the least squares regression line.
b. Can you compute the sum of the squared errors SSE using the definition $\Sigma(y - \hat{y})^2$? Explain.
c. Compute the sum of the squared errors SSE using the formula $SSE = SS_{yy} - \hat{\beta}_1 SS_{xy}$.

APPLICATIONS

11. For the data in Exercise 11 of Section 10.2 "The Linear Correlation Coefficient"

a. Compute the least squares regression line.
b. On average, how many new words does a child from 13 to 18 months old learn each month? Explain.
c. Estimate the average vocabulary of all 16-month-old children.

12. For the data in Exercise 12 of Section 10.2 "The Linear Correlation Coefficient"

a. Compute the least squares regression line.
b. On average, how many additional feet are added to the braking distance for each additional 100 pounds of weight? Explain.
c. Estimate the average braking distance of all cars weighing 3,000 pounds.

13. For the data in Exercise 13 of Section 10.2 "The Linear Correlation Coefficient"

a. Compute the least squares regression line.
b. Estimate the average resting heart rate of all 40-year-old men.
c. Estimate the average resting heart rate of all newborn baby boys. Comment on the validity of the estimate.

14. For the data in Exercise 14 of Section 10.2 "The Linear Correlation Coefficient"

a. Compute the least squares regression line.
b. Estimate the average wave height when the wind is blowing at 10 miles per hour.
c. Estimate the average wave height when there is no wind blowing. Comment on the validity of the estimate.

15. For the data in Exercise 15 of Section 10.2 "The Linear Correlation Coefficient"

a. Compute the least squares regression line.
b. On average, for each additional thousand dollars spent on advertising, how does revenue change? Explain.
c. Estimate the revenue if $2,500 is spent on advertising next year.

16. For the data in Exercise 16 of Section 10.2 "The Linear Correlation Coefficient"

a. Compute the least squares regression line.
b. On average, for each additional inch of height of two-year-old girl, what is the change in the adult height? Explain.
c. Predict the adult height of a two-year-old girl who is 33 inches tall.

17. For the data in Exercise 17 of Section 10.2 "The Linear Correlation Coefficient"

a. Compute the least squares regression line.
b. Compute SSE using the formula $SSE = SS_{yy} - \widehat{\beta}_1 SS_{xy}$.
c. Estimate the average final exam score of all students whose course average just before the exam is 85.

18. For the data in Exercise 18 of Section 10.2 "The Linear Correlation Coefficient"

a. Compute the least squares regression line.
b. Compute SSE using the formula $SSE = SS_{yy} - \widehat{\beta}_1 SS_{xy}$.
c. Estimate the number of acres that would be harvested if 90 million acres of corn were planted.

19. For the data in Exercise 19 of Section 10.2 "The Linear Correlation Coefficient"

a. Compute the least squares regression line.
b. Interpret the value of the slope of the least squares regression line in the context of the problem.
c. Estimate the average concentration of the active ingredient in the blood in men after consuming 1 ounce of the medication.

20. For the data in Exercise 20 of Section 10.2 "The Linear Correlation Coefficient"

a. Compute the least squares regression line.
b. Interpret the value of the slope of the least squares regression line in the context of the problem.
c. Estimate the age of an oak tree whose girth five feet off the ground is 92 inches.

21. For the data in Exercise 21 of Section 10.2 "The Linear Correlation Coefficient"

a. Compute the least squares regression line.

b. The 28-day strength of concrete used on a certain job must be at least 3,200 psi. If the 3-day strength is 1,300 psi, would we anticipate that the concrete will be sufficiently strong on the 28th day? Explain fully.

22. For the data in Exercise 22 of Section 10.2 "The Linear Correlation Coefficient"

a. Compute the least squares regression line.
b. If the power facility is called upon to provide more than 95 million watt-hours tomorrow then energy will have to be purchased from elsewhere at a premium. The forecast is for an average temperature of 42 degrees. Should the company plan on purchasing power at a premium?

ADDITIONAL EXERCISES

23. Verify that no matter what the data are, the least squares regression line always passes through the point with coordinates $(\overline{x}, \overline{y})$. Hint: Find the predicted value of y when $x = \overline{x}$.

24. In Exercise 1 you computed the least squares regression line for the data in Exercise 1 of Section 10.2 "The Linear Correlation Coefficient".

a. Reverse the roles of x and y and compute the least squares regression line for the new data set

x	2	4	6	5	9
y	0	1	3	5	8

b. Interchanging x and y corresponds geometrically to reflecting the scatter plot in a 45-degree line. Reflecting the regression line for the original data the same way gives a line with the equation $\hat{y} = 1.346x - 3.600$. Is this the equation that you got in part (a)? Can you figure out why not? Hint: Think about how x and y are treated differently geometrically in the computation of the goodness of fit.
c. Compute SSE for each line and see if they fit the same, or if one fits the data better than the other.

LARGE DATA SET EXERCISES

25. Large Data Set 1 lists the SAT scores and GPAs of 1,000 students.

http://www.gone.2012books.lardbucket.org/sites/all/files/data1.xls

a. Compute the least squares regression line with SAT score as the independent variable (x) and GPA as the dependent variable (y).
b. Interpret the meaning of the slope $\widehat{\beta}_1$ of regression line in the context of problem.
c. Compute SSE, the measure of the goodness of fit of the regression line to the sample data.
d. Estimate the GPA of a student whose SAT score is 1350.

26. Large Data Set 12 lists the golf scores on one round of golf for 75 golfers first using their own original clubs, then using clubs of a new, experimental design (after two months of familiarization with the new clubs).

    http://www.gone.2012books.lardbucket.org/sites/all/files/data12.xls

    a. Compute the least squares regression line with scores using the original clubs as the independent variable (x) and scores using the new clubs as the dependent variable (y).
    b. Interpret the meaning of the slope $\widehat{\beta}_1$ of regression line in the context of problem.
    c. Compute SSE, the measure of the goodness of fit of the regression line to the sample data.
    d. Estimate the score with the new clubs of a golfer whose score with the old clubs is 73.

27. Large Data Set 13 records the number of bidders and sales price of a particular type of antique grandfather clock at 60 auctions.

    http://www.gone.2012books.lardbucket.org/sites/all/files/data13.xls

    a. Compute the least squares regression line with the number of bidders present at the auction as the independent variable (x) and sales price as the dependent variable (y).
    b. Interpret the meaning of the slope $\widehat{\beta}_1$ of regression line in the context of problem.
    c. Compute SSE, the measure of the goodness of fit of the regression line to the sample data.
    d. Estimate the sales price of a clock at an auction at which the number of bidders is seven.

ANSWERS

1. $\hat{y} = 0.743x + 2.675$
3. $\hat{y} = -0.610x + 4.082$
5. $\hat{y} = 0.625x + 1.25$, $SSE = 5$
7. $\hat{y} = 0.6x + 1.8$
9. $\hat{y} = -1.45x + 2.4$, $SSE = 50.25$ (cannot use the definition to compute)
11. a. $\hat{y} = 4.848x - 56$,
    b. 4.8,
    c. 21.6
13. a. $\hat{y} = 0.114x + 69.222$,
    b. 73.8,
    c. 69.2, invalid extrapolation
15. a. $\hat{y} = 42.024x + 119.502$,
    b. increases by $42,024,
    c. $224,562
17. a. $\hat{y} = 1.045x - 8.527$,
    b. 2151.93367,
    c. 80.3
19. a. $\hat{y} = 0.043x + 0.001$,
    b. For each additional ounce of medication consumed blood concentration of the active ingredient increases by 0.043 %,
    c. 0.044%
21. a. $\hat{y} = 2.550x + 1.993$,
    b. Predicted 28-day strength is 3,514 psi; sufficiently strong
25. a. $\hat{y} = 0.0016x + 0.022$
    b. On average, every 100 point increase in SAT score adds 0.16 point to the GPA.
    c. $SSE = 432.10$
    d. $\hat{y} = 2.182$
27. a. $\hat{y} = 116.62x + 6955.1$

b. On average, every 1 additional bidder at an auction raises the price by 116.62 dollars.
c. $SSE = 1850314.08$
d. $\widehat{y} = 7771.44$

10.5 Statistical Inferences About β_1

LEARNING OBJECTIVES

1. To learn how to construct a confidence interval for β_1, the slope of the population regression line.
2. To learn how to test hypotheses regarding β_1.

The parameter β_1, the slope of the population regression line, is of primary importance in regression analysis because it gives the true rate of change in the mean $E(y)$ in response to a unit increase in the predictor variable *x*. For every unit increase in *x* the mean of the response variable *y* changes by β_1 units, increasing if $\beta_1 > 0$ and decreasing if $\beta_1 < 0$. We wish to construct confidence intervals for β_1 and test hypotheses about it.

Confidence Intervals for β_1

The slope $\hat{\beta}_1$ of the least squares regression line is a point estimate of β_1. A confidence interval for β_1 is given by the following formula.

$100(1-\alpha)\%$ Confidence Interval for the Slope β_1 of the Population Regression Line

$$\hat{\beta}_1 \pm t_{\alpha/2} \frac{s_\varepsilon}{\sqrt{SS_{xx}}}$$

where $s_\varepsilon = \sqrt{\frac{SSE}{n-2}}$ and the number of degrees of freedom is $df = n-2$.

The assumptions listed in Section 10.3 "Modelling Linear Relationships with Randomness Present" must hold.

Definition

The statistic s_ε is called the **sample standard deviation of errors**[7]*. It estimates the standard deviation σ of the errors in the population of y-values for each fixed value of x (see Figure 10.5 "The Simple Linear Model Concept" in Section 10.3 "Modelling Linear Relationships with Randomness Present").*

7. The statistic s_ε.

EXAMPLE 6

Construct the 95% confidence interval for the slope β_1 of the population regression line based on the five-point sample data set

x	2	2	6	8	10
y	0	1	2	3	3

Solution:

The point estimate $\widehat{\beta}_1$ of β_1 was computed in Note 10.18 "Example 2" in Section 10.4 "The Least Squares Regression Line" as $\widehat{\beta}_1 = 0.34375$. In the same example SS_{xx} was found to be $SS_{xx} = 51.2$. The sum of the squared errors SSE was computed in Note 10.23 "Example 4" in Section 10.4 "The Least Squares Regression Line" as $SSE = 0.75$. Thus

$$s_\varepsilon = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{0.75}{3}} = 0.50$$

Confidence level 95% means $\alpha = 1 - 0.95 = 0.05$ so $\alpha / 2 = 0.025$. From the row labeled $df = 3$ in Figure 12.3 "Critical Values of " we obtain $t_{0.025} = 3.182$. Therefore

$$\widehat{\beta}_1 \pm t_{\alpha/2} \frac{s_\varepsilon}{\sqrt{SS_{xx}}} = 0.34375 \pm 3.182 \left(\frac{0.50}{\sqrt{51.2}} \right) = 0.34375 \pm 0.2223$$

which gives the interval $(0.1215, 0.5661)$. We are 95% confident that the slope β_1 of the population regression line is between 0.1215 and 0.5661.

EXAMPLE 7

Using the sample data in Table 10.3 "Data on Age and Value of Used Automobiles of a Specific Make and Model" construct a 90% confidence interval for the slope β_1 of the population regression line relating age and value of the automobiles of Note 10.19 "Example 3" in Section 10.4 "The Least Squares Regression Line". Interpret the result in the context of the problem.

Solution:

The point estimate $\widehat{\beta}_1$ of β_1 was computed in Note 10.19 "Example 3", as was SS_{xx}. Their values are $\widehat{\beta}_1 = -2.05$ and $SS_{xx} = 14.$ The sum of the squared errors SSE was computed in Note 10.24 "Example 5" in Section 10.4 "The Least Squares Regression Line" as $SSE = 28.946.$ Thus

$$s_\varepsilon = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{28.946}{8}} = 1.902169814$$

Confidence level 90% means $\alpha = 1 - 0.90 = 0.10$ so $\alpha / 2 = 0.05.$ From the row labeled $df = 8$ in Figure 12.3 "Critical Values of " we obtain $t_{0.05} = 1.860.$ Therefore

$$\widehat{\beta}_1 \pm t_{\alpha/2} \frac{s_\varepsilon}{\sqrt{SS_{xx}}} = -2.05 \pm 1.860 \left(\frac{1.902169814}{\sqrt{14}} \right) = -2.05 \pm 0.95$$

which gives the interval $(-3.00, -1.10)$. We are 90% confident that the slope β_1 of the population regression line is between -3.00 and -1.10. In the context of the problem this means that for vehicles of this make and model between two and six years old we are 90% confident that for each additional year of age the average value of such a vehicle decreases by between $1,100 and $3,000.

Testing Hypotheses About β_1

Hypotheses regarding β_1 can be tested using the same five-step procedures, either the critical value approach or the *p*-value approach, that were introduced in Section 8.1 "The Elements of Hypothesis Testing" and Section 8.3 "The Observed

Significance of a Test" of Chapter 8 "Testing Hypotheses". The null hypothesis always has the form $H_0 : \beta_1 = B_0$ where B_0 is a number determined from the statement of the problem. The three forms of the alternative hypothesis, with the terminology for each case, are:

Form of H_a	Terminology
$H_a : \beta_1 < B_0$	Left-tailed
$H_a : \beta_1 > B_0$	Right-tailed
$H_a : \beta_1 \neq B_0$	Two-tailed

The value zero for B_0 is of particular importance since in that case the null hypothesis is $H_0 : \beta_1 = 0$, which corresponds to the situation in which x is not useful for predicting y. For if $\beta_1 = 0$ then the population regression line is horizontal, so the mean $E(y)$ is the same for every value of x and we are just as well off in ignoring x completely and approximating y by its average value. Given two variables x and y, the burden of proof is that x is useful for predicting y, not that it is not. Thus the phrase "test whether x is useful for prediction of y," or words to that effect, means to perform the test

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_a : \beta_1 \neq 0$$

Standardized Test Statistic for Hypothesis Tests Concerning the Slope β_1 of the Population Regression Line

$$T = \frac{\hat{\beta}_1 - B_0}{s_\varepsilon \Big/ \sqrt{SS_{xx}}}$$

The test statistic has Student's t-distribution with $df = n-2$ degrees of freedom.

The assumptions listed in Section 10.3 "Modelling Linear Relationships with Randomness Present" must hold.

EXAMPLE 8

Test, at the 2% level of significance, whether the variable x is useful for predicting y based on the information in the five-point data set

x	2	2	6	8	10
y	0	1	2	3	3

Solution:

We will perform the test using the critical value approach.

- Step 1. Since x is useful for prediction of y precisely when the slope β_1 of the population regression line is nonzero, the relevant test is

$$H_0 : \beta_1 = 0$$
$$\text{vs. } H_a : \beta_1 \neq 0 \quad @\, \alpha = 0.02$$

- Step 2. The test statistic is

$$T = \frac{\widehat{\beta}_1}{s_\varepsilon \Big/ \sqrt{SS_{xx}}}$$

  and has Student's t-distribution with $n-2 = 5 - 2 = 3$ degrees of freedom.

- Step 3. From Note 10.18 "Example 2", $\widehat{\beta}_1 = 0.34375$ and $SS_{xx} = 51.2$. From Note 10.30 "Example 6", $s_\varepsilon = 0.50$. The value of the test statistic is therefore

$$T = \frac{\widehat{\beta}_1 - B_0}{s_\varepsilon \Big/ \sqrt{SS_{xx}}} = \frac{0.34375}{0.50 \Big/ \sqrt{51.2}} = 4.919$$

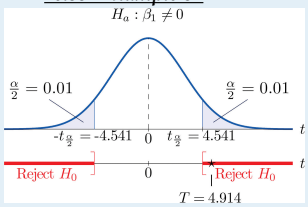
- Step 4. Since the symbol in H_a is "≠" this is a two-tailed test, so there are two critical values $\pm t_{\alpha/2} = \pm t_{0.01}$. Reading from the line in Figure

12.3 "Critical Values of " labeled $df = 3, t_{0.01} = 4.541$. The rejection region is $(-\infty, -4.541] \cup [4.541, \infty)$.

- Step 5. As shown in Figure 10.9 "Rejection Region and Test Statistic for " the test statistic falls in the rejection region. The decision is to reject H_0. In the context of the problem our conclusion is:

  The data provide sufficient evidence, at the 2% level of significance, to conclude that the slope of the population regression line is nonzero, so that x is useful as a predictor of y.

Figure 10.9
Rejection Region and Test Statistic for Note 10.33 "Example 8"



EXAMPLE 9

A car salesman claims that automobiles between two and six years old of the make and model discussed in Note 10.19 "Example 3" in Section 10.4 "The Least Squares Regression Line" lose more than $1,100 in value each year. Test this claim at the 5% level of significance.

Solution:

We will perform the test using the critical value approach.

- Step 1. In terms of the variables x and y, the salesman's claim is that if x is increased by 1 unit (one additional year in age), then y decreases by more than 1.1 units (more than $1,100). Thus his assertion is that the slope of the population regression line is negative, and that it is more negative than -1.1. In symbols, $\beta_1 < -1.1$. Since it contains an inequality, this has to be the alternative hypotheses. The null hypothesis has to be an equality and have the same number on the right hand side, so the relevant test is

$$H_0 : \beta_1 = -1.1$$
$$\text{vs. } H_\alpha : \beta_1 < -1.1 \quad @\, \alpha = 0.05$$

- Step 2. The test statistic is

$$T = \frac{\widehat{\beta}_1 - B_0}{s_\varepsilon \Big/ \sqrt{SS_{xx}}}$$

and has Student's t-distribution with 8 degrees of freedom.

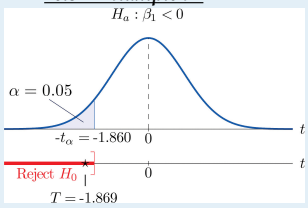
- Step 3. From Note 10.19 "Example 3", $\widehat{\beta}_1 = -2.05$ and $SS_{xx} = 14$. From Note 10.31 "Example 7", $s_\varepsilon = 1.902169814$. The value of the test statistic is therefore

$$T = \frac{\hat{\beta}_1 - B_0}{s_\varepsilon \Big/ \sqrt{SS_{xx}}} = \frac{-2.05 - (-1.1)}{1.902169814 \Big/ \sqrt{14}} = -1.869$$

- Step 4. Since the symbol in H_a is "<" this is a left-tailed test, so there is a single critical value $-t_\alpha = -t_{0.05}$. Reading from the line in Figure 12.3 "Critical Values of " labeled $df = 8, t_{0.05} = 1.860.$ The rejection region is $\left(-\infty, -1.860\right]$.
    - Step 5. As shown in Figure 10.10 "Rejection Region and Test Statistic for " the test statistic falls in the rejection region. The decision is to reject H_0. In the context of the problem our conclusion is:

      The data provide sufficient evidence, at the 5% level of significance, to conclude that vehicles of this make and model and in this age range lose more than $1,100 per year in value, on average.

Figure 10.10 Rejection Region and Test Statistic for Note 10.34 "Example 9"



KEY TAKEAWAYS

- The parameter β_1, the slope of the population regression line, is of primary interest because it describes the average change in *y* with respect to unit increase in *x*.
- The statistic $\widehat{\beta}_1$, the slope of the least squares regression line, is a point estimate of β_1. Confidence intervals for β_1 can be computed using a formula.
- Hypotheses regarding β_1 are tested using the same five-step procedures introduced in Chapter 8 "Testing Hypotheses".

EXERCISES

BASIC

For the Basic and Application exercises in this section use the computations that were done for the exercises with the same number in Section 10.2 "The Linear Correlation Coefficient" and Section 10.4 "The Least Squares Regression Line".

1. Construct the 95% confidence interval for the slope β_1 of the population regression line based on the sample data set of Exercise 1 of Section 10.2 "The Linear Correlation Coefficient".
2. Construct the 90% confidence interval for the slope β_1 of the population regression line based on the sample data set of Exercise 2 of Section 10.2 "The Linear Correlation Coefficient".
3. Construct the 90% confidence interval for the slope β_1 of the population regression line based on the sample data set of Exercise 3 of Section 10.2 "The Linear Correlation Coefficient".
4. Construct the 99% confidence interval for the slope β_1 of the population regression Exercise 4 of Section 10.2 "The Linear Correlation Coefficient".
5. For the data in Exercise 5 of Section 10.2 "The Linear Correlation Coefficient" test, at the 10% level of significance, whether x is useful for predicting y (that is, whether $\beta_1 \neq 0$).
6. For the data in Exercise 6 of Section 10.2 "The Linear Correlation Coefficient" test, at the 5% level of significance, whether x is useful for predicting y (that is, whether $\beta_1 \neq 0$).
7. Construct the 90% confidence interval for the slope β_1 of the population regression line based on the sample data set of Exercise 7 of Section 10.2 "The Linear Correlation Coefficient".
8. Construct the 95% confidence interval for the slope β_1 of the population regression line based on the sample data set of Exercise 8 of Section 10.2 "The Linear Correlation Coefficient".
9. For the data in Exercise 9 of Section 10.2 "The Linear Correlation Coefficient" test, at the 1% level of significance, whether x is useful for predicting y (that is, whether $\beta_1 \neq 0$).

10. For the data in Exercise 10 of Section 10.2 "The Linear Correlation Coefficient" test, at the 1% level of significance, whether x is useful for predicting y (that is, whether $\beta_1 \neq 0$).

APPLICATIONS

11. For the data in Exercise 11 of Section 10.2 "The Linear Correlation Coefficient" construct a 90% confidence interval for the mean number of new words acquired per month by children between 13 and 18 months of age.

12. For the data in Exercise 12 of Section 10.2 "The Linear Correlation Coefficient" construct a 90% confidence interval for the mean increased braking distance for each additional 100 pounds of vehicle weight.

13. For the data in Exercise 13 of Section 10.2 "The Linear Correlation Coefficient" test, at the 10% level of significance, whether age is useful for predicting resting heart rate.

14. For the data in Exercise 14 of Section 10.2 "The Linear Correlation Coefficient" test, at the 10% level of significance, whether wind speed is useful for predicting wave height.

15. For the situation described in Exercise 15 of Section 10.2 "The Linear Correlation Coefficient"

    a. Construct the 95% confidence interval for the mean increase in revenue per additional thousand dollars spent on advertising.
    b. An advertising agency tells the business owner that for every additional thousand dollars spent on advertising, revenue will increase by over $25,000. Test this claim (which is the alternative hypothesis) at the 5% level of significance.
    c. Perform the test of part (b) at the 10% level of significance.
    d. Based on the results in (b) and (c), how believable is the ad agency's claim? (This is a subjective judgement.)

16. For the situation described in Exercise 16 of Section 10.2 "The Linear Correlation Coefficient"

    a. Construct the 90% confidence interval for the mean increase in height per additional inch of length at age two.
    b. It is claimed that for girls each additional inch of length at age two means more than an additional inch of height at maturity. Test this claim (which is the alternative hypothesis) at the 10% level of significance.

17. For the data in Exercise 17 of Section 10.2 "The Linear Correlation Coefficient" test, at the 10% level of significance, whether course average before the final exam is useful for predicting the final exam grade.

18. For the situation described in Exercise 18 of Section 10.2 "The Linear Correlation Coefficient", an agronomist claims that each additional million acres planted results in more than 750,000 additional acres harvested. Test this claim at the 1% level of significance.

19. For the data in Exercise 19 of Section 10.2 "The Linear Correlation Coefficient" test, at the 1/10th of 1% level of significance, whether, ignoring all other facts such as age and body mass, the amount of the medication consumed is a useful predictor of blood concentration of the active ingredient.

20. For the data in Exercise 20 of Section 10.2 "The Linear Correlation Coefficient" test, at the 1% level of significance, whether for each additional inch of girth the age of the tree increases by at least two and one-half years.

21. For the data in Exercise 21 of Section 10.2 "The Linear Correlation Coefficient"

    a. Construct the 95% confidence interval for the mean increase in strength at 28 days for each additional hundred psi increase in strength at 3 days.
    b. Test, at the 1/10th of 1% level of significance, whether the 3-day strength is useful for predicting 28-day strength.

22. For the situation described in Exercise 22 of Section 10.2 "The Linear Correlation Coefficient"

    a. Construct the 99% confidence interval for the mean decrease in energy demand for each one-degree drop in temperature.
    b. An engineer with the power company believes that for each one-degree increase in temperature, daily energy demand will decrease by more than 3.6 million watt-hours. Test this claim at the 1% level of significance.

LARGE DATA SET EXERCISES

23. Large Data Set 1 lists the SAT scores and GPAs of 1,000 students.

    http://www.gone.2012books.lardbucket.org/sites/all/files/data1.xls

    a. Compute the 90% confidence interval for the slope β_1 of the population regression line with SAT score as the independent variable (x) and GPA as the dependent variable (y).

b. Test, at the 10% level of significance, the hypothesis that the slope of the population regression line is greater than 0.001, against the null hypothesis that it is exactly 0.001.

24. Large Data Set 12 lists the golf scores on one round of golf for 75 golfers first using their own original clubs, then using clubs of a new, experimental design (after two months of familiarization with the new clubs).

    http://www.gone.2012books.lardbucket.org/sites/all/files/data12.xls

    a. Compute the 95% confidence interval for the slope β_1 of the population regression line with scores using the original clubs as the independent variable (x) and scores using the new clubs as the dependent variable (y).
    b. Test, at the 10% level of significance, the hypothesis that the slope of the population regression line is different from 1, against the null hypothesis that it is exactly 1.

25. Large Data Set 13 records the number of bidders and sales price of a particular type of antique grandfather clock at 60 auctions.

    http://www.gone.2012books.lardbucket.org/sites/all/files/data13.xls

    a. Compute the 95% confidence interval for the slope β_1 of the population regression line with the number of bidders present at the auction as the independent variable (x) and sales price as the dependent variable (y).
    b. Test, at the 10% level of significance, the hypothesis that the average sales price increases by more than $90 for each additional bidder at an auction, against the default that it increases by exactly $90.

ANSWERS

1. 0.743 ± 0.578
3. -0.610 ± 0.633
5. $T = 1.732$, $\pm t_{0.05} = \pm 2.353$, do not reject H_0
7. 0.6 ± 0.451
9. $T = -4.481$, $\pm t_{0.005} = \pm 3.355$, reject H_0
11. 4.8 ± 1.7 words
13. $T = 2.843$, $\pm t_{0.05} = \pm 1.860$, reject H_0
15. a. 42.024 ± 28.011 thousand dollars,
    b. $T = 1.487$, $t_{0.05} = 1.943$, do not reject H_0;
    c. $t_{0.10} = 1.440$, reject H_0
17. $T = 4.096$, $\pm t_{0.05} = \pm 1.771$, reject H_0
19. $T = 25.524$, $\pm t_{0.0005} = \pm 3.505$, reject H_0
21. a. 2.550 ± 0.127 hundred psi,
    b. $T = 41.072$, $\pm t_{0.005} = \pm 3.674$, reject H_0
23. a. $(0.0014, 0.0018)$
    b. $H_0 : \beta_1 = 0.001$ vs. $H_a : \beta_1 > 0.001$. Test Statistic: $Z = 6.1625$. Rejection Region: $[1.28, +\infty)$. Decision: Reject H_0.
25. a. $(101.789, 131.4435)$
    b. $H_0 : \beta_1 = 90$ vs. $H_a : \beta_1 > 90$. Test Statistic: $T = 3.5938$. $d.f. = 58$. Rejection Region: $[1.296, +\infty)$. Decision: Reject H_0.

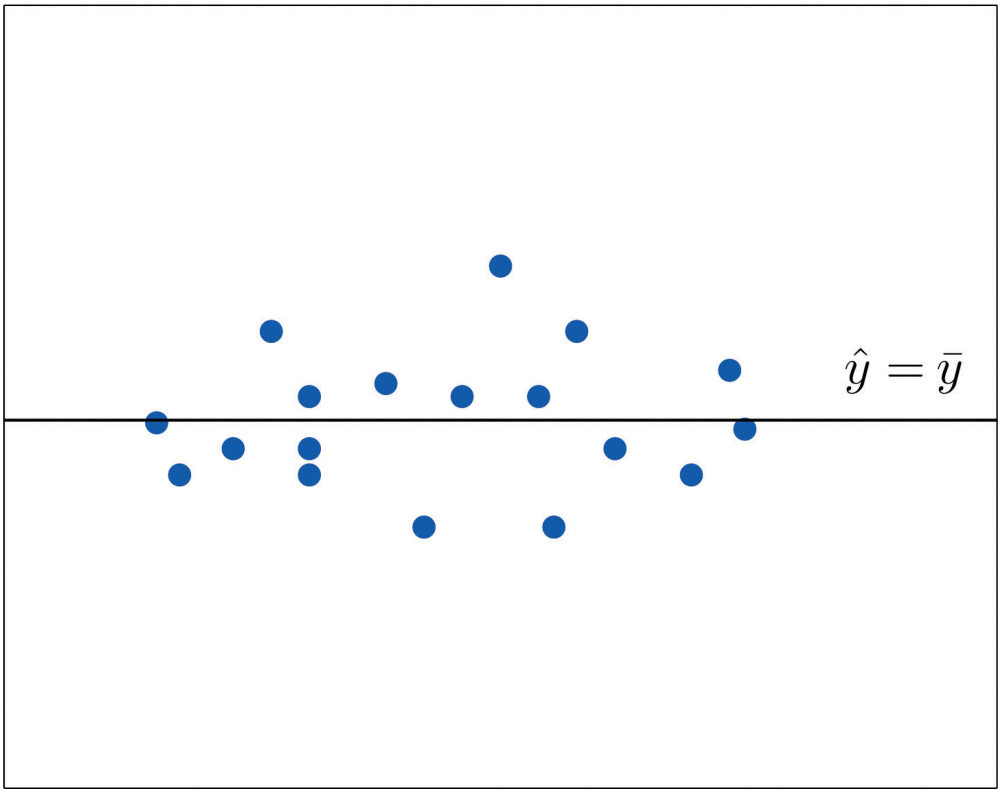
10.6 The Coefficient of Determination

LEARNING OBJECTIVE

1. To learn what the coefficient of determination is, how to compute it, and what it tells us about the relationship between two variables x and y.

If the scatter diagram of a set of (x, y) pairs shows neither an upward or downward trend, then the horizontal line $\hat{y} = \bar{y}$ fits it well, as illustrated in Figure 10.11. The lack of any upward or downward trend means that when an element of the population is selected at random, knowing the value of the measurement x for that element is not helpful in predicting the value of the measurement y.

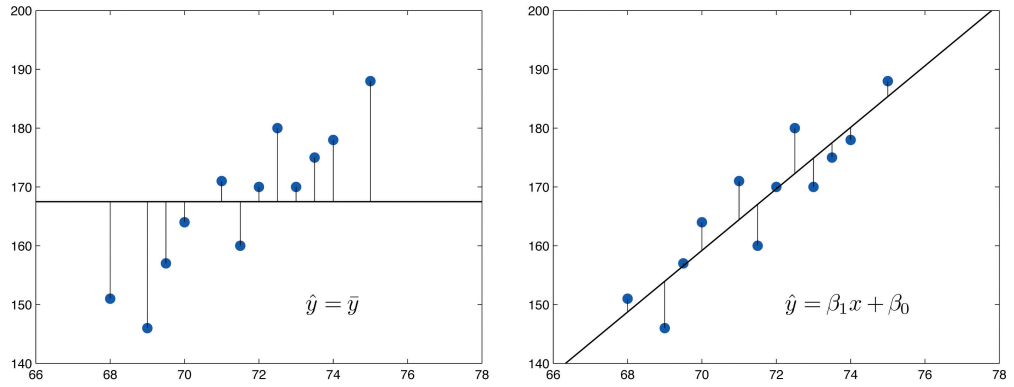
Figure 10.11



The line $\hat{y} = \bar{y}$ fits the scatter diagram well.

If the scatter diagram shows a linear trend upward or downward then it is useful to compute the least squares regression line $\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0$ and use it in predicting y. Figure 10.12 "Same Scatter Diagram with Two Approximating Lines" illustrates this. In each panel we have plotted the height and weight data of Section 10.1 "Linear Relationships Between Variables". This is the same scatter plot as Figure 10.2 "Plot of Height and Weight Pairs", with the average value line $\hat{y} = \bar{y}$ superimposed on it in the left panel and the least squares regression line imposed on it in the right panel. The errors are indicated graphically by the vertical line segments.

Figure 10.12 *Same Scatter Diagram with Two Approximating Lines*



(a): Set I (b): Set II

The sum of the squared errors computed for the regression line, SSE, is smaller than the sum of the squared errors computed for any other line. In particular it is less than the sum of the squared errors computed using the line $\hat{y} = \bar{y}$, which sum is actually the number SS_{yy} that we have seen several times already. A measure of how useful it is to use the regression equation for prediction of y is how much smaller SSE is than SS_{yy}. In particular, the *proportion* of the sum of the squared errors for the line $\hat{y} = \bar{y}$ that is eliminated by going over to the least squares regression line is

$$\frac{SS_{yy} - SSE}{SS_{yy}} = \frac{SS_{yy}}{SS_{yy}} - \frac{SSE}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}}$$

We can think of $SSE \,/\, SS_{yy}$ as the proportion of the variability in y that cannot be accounted for by the linear relationship between x and y, since it is still there even when x is taken into account in the best way possible (using the least squares regression line; remember that SSE is the smallest the sum of the squared errors can be for any line). Seen in this light, the coefficient of determination, the complementary proportion of the variability in y, is the proportion of the

variability in all the y measurements that is accounted for by the linear relationship between x and y.

In the context of linear regression the coefficient of determination is always the square of the correlation coefficient r discussed in Section 10.2 "The Linear Correlation Coefficient". Thus the coefficient of determination is denoted r^2, and we have two additional formulas for computing it.

Definition

The **coefficient of determination**[8] *of a collection of* (x, y) *pairs is the number* r^2 *computed by any of the following three expressions:*

$$r^2 = \frac{SS_{yy} - SSE}{SS_{yy}} = \frac{SS_{xy}^2}{SS_{xx}\,SS_{yy}} = \hat{\beta}_1 \frac{SS_{xy}}{SS_{yy}}$$

It measures the proportion of the variability in y that is accounted for by the linear relationship between x and y.

If the correlation coefficient r is already known then the coefficient of determination can be computed simply by squaring r, as the notation indicates, $r^2 = (r)^2$.

8. A number that measures the proportion of the variability in y that is explained by x.

EXAMPLE 10

The value of used vehicles of the make and model discussed in Note 10.19 "Example 3" in Section 10.4 "The Least Squares Regression Line" varies widely. The most expensive automobile in the sample in Table 10.3 "Data on Age and Value of Used Automobiles of a Specific Make and Model" has value $30,500, which is nearly half again as much as the least expensive one, which is worth $20,400. Find the proportion of the variability in value that is accounted for by the linear relationship between age and value.

Solution:

The proportion of the variability in value y that is accounted for by the linear relationship between it and age x is given by the coefficient of determination, r^2. Since the correlation coefficient r was already computed in Note 10.19 "Example 3" as $r = -0.819, r^2 = (-0.819)^2 = 0.671$. About 67% of the variability in the value of this vehicle can be explained by its age.

EXAMPLE 11

Use each of the three formulas for the coefficient of determination to compute its value for the example of ages and values of vehicles.

Solution:

In Note 10.19 "Example 3" in Section 10.4 "The Least Squares Regression Line" we computed the exact values

$$SS_{xx} = 14 \quad SS_{xy} = -28.7 \quad SS_{yy} = 87.781 \quad \widehat{\beta}_1 = -2.05$$

In Note 10.24 "Example 5" in Section 10.4 "The Least Squares Regression Line" we computed the exact value

$$SSE = 28.946$$

Inserting these values into the formulas in the definition, one after the other, gives

$$r^2 = \frac{SS_{yy} - SSE}{SS_{yy}} = \frac{87.781 - 28.946}{87.781} = 0.6702475479$$

$$r^2 = \frac{SS_{xy}^2}{SS_{xx}\,SS_{yy}} = \frac{(-28.7)^2}{(14)\ (87.781)} = 0.6702475479$$

$$r^2 = \widehat{\beta}_1 \frac{SS_{xy}}{SS_{yy}} = -2.05\ \frac{-28.7}{87.781} = 0.6702475479$$

which rounds to 0.670. The discrepancy between the value here and in the previous example is because a rounded value of r from Note 10.19 "Example 3" was used there. The actual value of r before rounding is 0.8186864772, which when squared gives the value for r^2 obtained here.

The coefficient of determination r^2 can always be computed by squaring the correlation coefficient r if it is known. Any one of the defining formulas can also be used. Typically one would make the choice based on which quantities have already been computed. What should be avoided is trying to compute r by taking the square root of r^2, if it is already known, since it is easy to make a sign error this way. To see what can go wrong, suppose $r^2 = 0.64$. Taking the square root of a positive

number with any calculating device will always return a positive result. The square root of 0.64 is 0.8. However, the actual value of r might be the negative number −0.8.

KEY TAKEAWAYS

- The coefficient of determination r^2 estimates the proportion of the variability in the variable y that is explained by the linear relationship between y and the variable x.
- There are several formulas for computing r^2. The choice of which one to use can be based on which quantities have already been computed so far.

EXERCISES

BASIC

For the Basic and Application exercises in this section use the computations that were done for the exercises with the same number in Section 10.2 "The Linear Correlation Coefficient", Section 10.4 "The Least Squares Regression Line", and Section 10.5 "Statistical Inferences About ".

1. For the sample data set of Exercise 1 of Section 10.2 "The Linear Correlation Coefficient" find the coefficient of determination using the formula $r^2 = \hat{\beta}_1 SS_{xy} \ / \ SS_{yy}$. Confirm your answer by squaring r as computed in that exercise.
2. For the sample data set of Exercise 2 of Section 10.2 "The Linear Correlation Coefficient" find the coefficient of determination using the formula $r^2 = \hat{\beta}_1 SS_{xy} \ / \ SS_{yy}$. Confirm your answer by squaring r as computed in that exercise.
3. For the sample data set of Exercise 3 of Section 10.2 "The Linear Correlation Coefficient" find the coefficient of determination using the formula $r^2 = \hat{\beta}_1 SS_{xy} \ / \ SS_{yy}$. Confirm your answer by squaring r as computed in that exercise.
4. For the sample data set of Exercise 4 of Section 10.2 "The Linear Correlation Coefficient" find the coefficient of determination using the formula $r^2 = \hat{\beta}_1 SS_{xy} \ / \ SS_{yy}$. Confirm your answer by squaring r as computed in that exercise.
5. For the sample data set of Exercise 5 of Section 10.2 "The Linear Correlation Coefficient" find the coefficient of determination using the formula $r^2 = \hat{\beta}_1 SS_{xy} \ / \ SS_{yy}$. Confirm your answer by squaring r as computed in that exercise.
6. For the sample data set of Exercise 6 of Section 10.2 "The Linear Correlation Coefficient" find the coefficient of determination using the formula $r^2 = \hat{\beta}_1 SS_{xy} \ / \ SS_{yy}$. Confirm your answer by squaring r as computed in that exercise.
7. For the sample data set of Exercise 7 of Section 10.2 "The Linear Correlation Coefficient" find the coefficient of determination using the formula

$r^2 = \left(SS_{yy} - SSE\right) / SS_{yy}$. Confirm your answer by squaring r as computed in that exercise.

8. For the sample data set of Exercise 8 of Section 10.2 "The Linear Correlation Coefficient" find the coefficient of determination using the formula $r^2 = \left(SS_{yy} - SSE\right) / SS_{yy}$. Confirm your answer by squaring r as computed in that exercise.

9. For the sample data set of Exercise 9 of Section 10.2 "The Linear Correlation Coefficient" find the coefficient of determination using the formula $r^2 = \left(SS_{yy} - SSE\right) / SS_{yy}$. Confirm your answer by squaring r as computed in that exercise.

10. For the sample data set of Exercise 9 of Section 10.2 "The Linear Correlation Coefficient" find the coefficient of determination using the formula $r^2 = \left(SS_{yy} - SSE\right) / SS_{yy}$. Confirm your answer by squaring r as computed in that exercise.

APPLICATIONS

11. For the data in Exercise 11 of Section 10.2 "The Linear Correlation Coefficient" compute the coefficient of determination and interpret its value in the context of age and vocabulary.

12. For the data in Exercise 12 of Section 10.2 "The Linear Correlation Coefficient" compute the coefficient of determination and interpret its value in the context of vehicle weight and braking distance.

13. For the data in Exercise 13 of Section 10.2 "The Linear Correlation Coefficient" compute the coefficient of determination and interpret its value in the context of age and resting heart rate. In the age range of the data, does age seem to be a very important factor with regard to heart rate?

14. For the data in Exercise 14 of Section 10.2 "The Linear Correlation Coefficient" compute the coefficient of determination and interpret its value in the context of wind speed and wave height. Does wind speed seem to be a very important factor with regard to wave height?

15. For the data in Exercise 15 of Section 10.2 "The Linear Correlation Coefficient" find the proportion of the variability in revenue that is explained by level of advertising.

16. For the data in Exercise 16 of Section 10.2 "The Linear Correlation Coefficient" find the proportion of the variability in adult height that is explained by the variation in length at age two.
17. For the data in Exercise 17 of Section 10.2 "The Linear Correlation Coefficient" compute the coefficient of determination and interpret its value in the context of course average before the final exam and score on the final exam.
18. For the data in Exercise 18 of Section 10.2 "The Linear Correlation Coefficient" compute the coefficient of determination and interpret its value in the context of acres planted and acres harvested.
19. For the data in Exercise 19 of Section 10.2 "The Linear Correlation Coefficient" compute the coefficient of determination and interpret its value in the context of the amount of the medication consumed and blood concentration of the active ingredient.
20. For the data in Exercise 20 of Section 10.2 "The Linear Correlation Coefficient" compute the coefficient of determination and interpret its value in the context of tree size and age.
21. For the data in Exercise 21 of Section 10.2 "The Linear Correlation Coefficient" find the proportion of the variability in 28-day strength of concrete that is accounted for by variation in 3-day strength.
22. For the data in Exercise 22 of Section 10.2 "The Linear Correlation Coefficient" find the proportion of the variability in energy demand that is accounted for by variation in average temperature.

LARGE DATA SET EXERCISES

23. Large Data Set 1 lists the SAT scores and GPAs of 1,000 students. Compute the coefficient of determination and interpret its value in the context of SAT scores and GPAs.

    http://www.gone.2012books.lardbucket.org/sites/all/files/data1.xls
24. Large Data Set 12 lists the golf scores on one round of golf for 75 golfers first using their own original clubs, then using clubs of a new, experimental design (after two months of familiarization with the new clubs). Compute the coefficient of determination and interpret its value in the context of golf scores with the two kinds of golf clubs.

    http://www.gone.2012books.lardbucket.org/sites/all/files/data12.xls

25. Large Data Set 13 records the number of bidders and sales price of a particular type of antique grandfather clock at 60 auctions. Compute the coefficient of determination and interpret its value in the context of the number of bidders at an auction and the price of this type of antique grandfather clock.

    http://www.gone.2012books.lardbucket.org/sites/all/files/data13.xls

ANSWERS

1. 0.848
3. 0.631
5. 0.5
7. 0.766
9. 0.715
11. 0.898; about 90% of the variability in vocabulary is explained by age
13. 0.503; about 50% of the variability in heart rate is explained by age. Age is a significant but not dominant factor in explaining heart rate.
15. The proportion is $r^2 = 0.692$.
17. 0.563; about 56% of the variability in final exam scores is explained by course average before the final exam
19. 0.931; about 93% of the variability in the blood concentration of the active ingredient is explained by the amount of the medication consumed
21. The proportion is $r^2 = 0.984$.
23. $r^2 = 21.17\%$.
25. $r^2 = 81.04\%$.

10.7 Estimation and Prediction

LEARNING OBJECTIVES

1. To learn the distinction between estimation and prediction.
2. To learn the distinction between a confidence interval and a prediction interval.
3. To learn how to implement formulas for computing confidence intervals and prediction intervals.

Consider the following pairs of problems, in the context of Note 10.19 "Example 3" in Section 10.4 "The Least Squares Regression Line", the automobile age and value example.

1.
    1. Estimate the average value of all four-year-old automobiles of this make and model.
    2. Construct a 95% confidence interval for the average value of all four-year-old automobiles of this make and model.
2.
    1. Shylock intends to buy a four-year-old automobile of this make and model next week. Predict the value of the first such automobile that he encounters.
    2. Construct a 95% confidence interval for the value of the first such automobile that he encounters.

The method of solution and answer to the first question in each pair, (1a) and (2a), are the same. When we set x equal to 4 in the least squares regression equation $\hat{y} = -2.05x + 32.83$ that was computed in part (c) of Note 10.19 "Example 3" in Section 10.4 "The Least Squares Regression Line", the number returned,

$$\hat{y} = -2.05\,(4) + 32.83 = 24.63$$

which corresponds to value \$24,630, is an estimate of precisely the number sought in question (1a): the mean $E\left(y\right)$ of all y values when x = 4. Since nothing is known about the first four-year-old automobile of this make and model that Shylock will

encounter, our best guess as to its value is the mean value $E\left(y\right)$ of all such automobiles, the number 24.63 or $24,630, computed in the same way.

The answers to the second part of each question differ. In question (1b) we are trying to estimate a population parameter: the mean of the all the *y*-values in the sub-population picked out by the value $x = 4$, that is, the average value of all four-year-old automobiles. In question (2b), however, we are not trying to capture a fixed parameter, but the value of the random variable *y* in one trial of an experiment: examine the first four-year-old car Shylock encounters. In the first case we seek to construct a confidence interval in the same sense that we have done before. In the second case the situation is different, and the interval constructed has a different name, prediction interval. In the second case we are trying to "predict" where a the value of a random variable will take its value.

$100\left(1-\alpha\right)\%$ Confidence Interval for the Mean Value of *y* at $x = x_p$

$$\hat{y}_p \pm t_{\alpha/2}\, s_\varepsilon \sqrt{\frac{1}{n} + \frac{\left(x_p - \overline{x}\right)^2}{SS_{xx}}}$$

where

a. x_p is a particular value of *x* that lies in the range of *x*-values in the sample data set used to construct the least squares regression line;
b. $\hat{y}_p$ is the numerical value obtained when the least square regression equation is evaluated at $x = x_p$; and
c. the number of degrees of freedom for $t_{\alpha/2}$ is $df = n-2$.

The assumptions listed in Section 10.3 "Modelling Linear Relationships with Randomness Present" must hold.

The formula for the prediction interval is identical except for the presence of the number 1 underneath the square root sign. This means that the prediction interval is always wider than the confidence interval at the same confidence level and value of *x*. In practice the presence of the number 1 tends to make it much wider.

$100\,(1-\alpha)\,\%$ Prediction Interval for an Individual New Value of y at $x = x_p$

$$\hat{y}_p \pm t_{\alpha/2}\, s_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{\left(x_p - \overline{x}\right)^2}{SS_{xx}}}$$

where

a. x_p is a particular value of x that lies in the range of x-values in the data set used to construct the least squares regression line;
b. $\hat{y}_p$ is the numerical value obtained when the least square regression equation is evaluated at $x = x_p$; and
c. the number of degrees of freedom for $t_{\alpha/2}$ is $df = n-2$.

The assumptions listed in Section 10.3 "Modelling Linear Relationships with Randomness Present" must hold.

EXAMPLE 12

Using the sample data of Note 10.19 "Example 3" in Section 10.4 "The Least Squares Regression Line", recorded in Table 10.3 "Data on Age and Value of Used Automobiles of a Specific Make and Model", construct a 95% confidence interval for the average value of all three-and-one-half-year-old automobiles of this make and model.

Solution:

Solving this problem is merely a matter of finding the values of $\hat{y}_p$, α and $t_{\alpha/2}$, s_ε, $\overline{x}$, and SS_{xx} and inserting them into the confidence interval formula given just above. Most of these quantities are already known. From Note 10.19 "Example 3" in Section 10.4 "The Least Squares Regression Line", $SS_{xx} = 14$ and $\overline{x} = 4$. From Note 10.31 "Example 7" in Section 10.5 "Statistical Inferences About ", $s_\varepsilon = 1.902169814$.

From the statement of the problem $x_p = 3.5$, the value of x of interest. The value of $\hat{y}_p$ is the number given by the regression equation, which by Note 10.19 "Example 3" is $\hat{y} = -2.05x + 32.83$, when $x = x_p$, that is, when $x = 3.5$. Thus here $\hat{y}_p = -2.05\left(3.5\right) + 32.83 = 25.655$.

Lastly, confidence level 95% means that $\alpha = 1 - 0.95 = 0.05$ so $\alpha / 2 = 0.025$. Since the sample size is $n = 10$, there are $n-2 = 8$ degrees of freedom. By Figure 12.3 "Critical Values of ", $t_{0.025} = 2.306$. Thus

$$\hat{y}_p \pm t_{\alpha/2}\, s_\varepsilon \sqrt{\frac{1}{n} + \frac{\left(x_p - \overline{x}\right)^2}{SS_{xx}}} = 25.655 \pm \left(2.306\right)\left(1.902169814\right)\sqrt{-}$$

$$= 25.655 \pm 4.386403591\sqrt{0.117857142}$$

$$= 25.655 \pm 1.506$$

which gives the interval $\left(24.149, 27.161\right)$.

We are 95% confident that the average value of all three-and-one-half-year-old vehicles of this make and model is between $24,149 and $27,161.

EXAMPLE 13

Using the sample data of Note 10.19 "Example 3" in Section 10.4 "The Least Squares Regression Line", recorded in Table 10.3 "Data on Age and Value of Used Automobiles of a Specific Make and Model", construct a 95% prediction interval for the predicted value of a randomly selected three-and-one-half-year-old automobile of this make and model.

Solution:

The computations for this example are identical to those of the previous example, except that now there is the extra number 1 beneath the square root sign. Since we were careful to record the intermediate results of that computation, we have immediately that the 95% prediction interval is

$$\hat{y}_p \pm t_{\alpha/2}\, s_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{SS_{xx}}} = 25.655 \pm 4.386403591\sqrt{1.1178571}$$

which gives the interval $(21.017, 30.293)$.

We are 95% confident that the value of a randomly selected three-and-one-half-year-old vehicle of this make and model is between $21,017 and $30,293.

Note what an enormous difference the presence of the extra number 1 under the square root sign made. The prediction interval is about two-and-one-half times wider than the confidence interval at the same level of confidence.

KEY TAKEAWAYS

- A confidence interval is used to estimate the mean value of y in the sub-population determined by the condition that x have some specific value x_p.
- The prediction interval is used to predict the value that the random variable y will take when x has some specific value x_p.

EXERCISES

BASIC

For the Basic and Application exercises in this section use the computations that were done for the exercises with the same number in previous sections.

1. For the sample data set of Exercise 1 of Section 10.2 "The Linear Correlation Coefficient"
    a. Give a point estimate for the mean value of y in the sub-population determined by the condition $x = 4$.
    b. Construct the 90% confidence interval for that mean value.
2. For the sample data set of Exercise 2 of Section 10.2 "The Linear Correlation Coefficient"
    a. Give a point estimate for the mean value of y in the sub-population determined by the condition $x = 4$.
    b. Construct the 90% confidence interval for that mean value.
3. For the sample data set of Exercise 3 of Section 10.2 "The Linear Correlation Coefficient"
    a. Give a point estimate for the mean value of y in the sub-population determined by the condition $x = 7$.
    b. Construct the 95% confidence interval for that mean value.
4. For the sample data set of Exercise 4 of Section 10.2 "The Linear Correlation Coefficient"
    a. Give a point estimate for the mean value of y in the sub-population determined by the condition $x = 2$.
    b. Construct the 80% confidence interval for that mean value.
5. For the sample data set of Exercise 5 of Section 10.2 "The Linear Correlation Coefficient"
    a. Give a point estimate for the mean value of y in the sub-population determined by the condition $x = 1$.
    b. Construct the 80% confidence interval for that mean value.
6. For the sample data set of Exercise 6 of Section 10.2 "The Linear Correlation Coefficient"

a. Give a point estimate for the mean value of y in the sub-population determined by the condition $x = 5$.
b. Construct the 95% confidence interval for that mean value.

7. For the sample data set of Exercise 7 of Section 10.2 "The Linear Correlation Coefficient"

   a. Give a point estimate for the mean value of y in the sub-population determined by the condition $x = 6$.
   b. Construct the 99% confidence interval for that mean value.
   c. Is it valid to make the same estimates for $x = 12$? Explain.

8. For the sample data set of Exercise 8 of Section 10.2 "The Linear Correlation Coefficient"

   a. Give a point estimate for the mean value of y in the sub-population determined by the condition $x = 12$.
   b. Construct the 80% confidence interval for that mean value.
   c. Is it valid to make the same estimates for $x = 0$? Explain.

9. For the sample data set of Exercise 9 of Section 10.2 "The Linear Correlation Coefficient"

   a. Give a point estimate for the mean value of y in the sub-population determined by the condition $x = 0$.
   b. Construct the 90% confidence interval for that mean value.
   c. Is it valid to make the same estimates for $x = -1$? Explain.

10. For the sample data set of Exercise 9 of Section 10.2 "The Linear Correlation Coefficient"

    a. Give a point estimate for the mean value of y in the sub-population determined by the condition $x = 8$.
    b. Construct the 95% confidence interval for that mean value.
    c. Is it valid to make the same estimates for $x = 0$? Explain.

APPLICATIONS

11. For the data in Exercise 11 of Section 10.2 "The Linear Correlation Coefficient"

    a. Give a point estimate for the average number of words in the vocabulary of 18-month-old children.
    b. Construct the 95% confidence interval for that mean value.
    c. Is it valid to make the same estimates for two-year-olds? Explain.

12. For the data in Exercise 12 of Section 10.2 "The Linear Correlation Coefficient"

a. Give a point estimate for the average braking distance of automobiles that weigh 3,250 pounds.
b. Construct the 80% confidence interval for that mean value.
c. Is it valid to make the same estimates for 5,000-pound automobiles? Explain.

13. For the data in Exercise 13 of Section 10.2 "The Linear Correlation Coefficient"

a. Give a point estimate for the resting heart rate of a man who is 35 years old.
b. One of the men in the sample is 35 years old, but his resting heart rate is not what you computed in part (a). Explain why this is not a contradiction.
c. Construct the 90% confidence interval for the mean resting heart rate of all 35-year-old men.

14. For the data in Exercise 14 of Section 10.2 "The Linear Correlation Coefficient"

a. Give a point estimate for the wave height when the wind speed is 13 miles per hour.
b. One of the wind speeds in the sample is 13 miles per hour, but the height of waves that day is not what you computed in part (a). Explain why this is not a contradiction.
c. Construct the 90% confidence interval for the mean wave height on days when the wind speed is 13 miles per hour.

15. For the data in Exercise 15 of Section 10.2 "The Linear Correlation Coefficient"

a. The business owner intends to spend $2,500 on advertising next year. Give an estimate of next year's revenue based on this fact.
b. Construct the 90% prediction interval for next year's revenue, based on the intent to spend $2,500 on advertising.

16. For the data in Exercise 16 of Section 10.2 "The Linear Correlation Coefficient"

a. A two-year-old girl is 32.3 inches long. Predict her adult height.
b. Construct the 95% prediction interval for the girl's adult height.

17. For the data in Exercise 17 of Section 10.2 "The Linear Correlation Coefficient"

a. Lodovico has a 78.6 average in his physics class just before the final. Give a point estimate of what his final exam grade will be.
b. Explain whether an interval estimate for this problem is a confidence interval or a prediction interval.
c. Based on your answer to (b), construct an interval estimate for Lodovico's final exam grade at the 90% level of confidence.

18. For the data in Exercise 18 of Section 10.2 "The Linear Correlation Coefficient"

a. This year 86.2 million acres of corn were planted. Give a point estimate of the number of acres that will be harvested this year.
b. Explain whether an interval estimate for this problem is a confidence interval or a prediction interval.
c. Based on your answer to (b), construct an interval estimate for the number of acres that will be harvested this year, at the 99% level of confidence.

19. For the data in Exercise 19 of Section 10.2 "The Linear Correlation Coefficient"

a. Give a point estimate for the blood concentration of the active ingredient of this medication in a man who has consumed 1.5 ounces of the medication just recently.
b. Gratiano just consumed 1.5 ounces of this medication 30 minutes ago. Construct a 95% prediction interval for the concentration of the active ingredient in his blood right now.

20. For the data in Exercise 20 of Section 10.2 "The Linear Correlation Coefficient"

a. You measure the girth of a free-standing oak tree five feet off the ground and obtain the value 127 inches. How old do you estimate the tree to be?
b. Construct a 90% prediction interval for the age of this tree.

21. For the data in Exercise 21 of Section 10.2 "The Linear Correlation Coefficient"

a. A test cylinder of concrete three days old fails at 1,750 psi. Predict what the 28-day strength of the concrete will be.
b. Construct a 99% prediction interval for the 28-day strength of this concrete.
c. Based on your answer to (b), what would be the minimum 28-day strength you could expect this concrete to exhibit?

22. For the data in Exercise 22 of Section 10.2 "The Linear Correlation Coefficient"

a. Tomorrow's average temperature is forecast to be 53 degrees. Estimate the energy demand tomorrow.
b. Construct a 99% prediction interval for the energy demand tomorrow.
c. Based on your answer to (b), what would be the minimum demand you could expect?

LARGE DATA SET EXERCISES

23. Large Data Set 1 lists the SAT scores and GPAs of 1,000 students.

http://www.gone.2012books.lardbucket.org/sites/all/files/data1.xls

a. Give a point estimate of the mean GPA of all students who score 1350 on the SAT.
b. Construct a 90% confidence interval for the mean GPA of all students who score 1350 on the SAT.

24. Large Data Set 12 lists the golf scores on one round of golf for 75 golfers first using their own original clubs, then using clubs of a new, experimental design (after two months of familiarization with the new clubs).

http://www.gone.2012books.lardbucket.org/sites/all/files/data12.xls

a. Thurio averages 72 strokes per round with his own clubs. Give a point estimate for his score on one round if he switches to the new clubs.
b. Explain whether an interval estimate for this problem is a confidence interval or a prediction interval.
c. Based on your answer to (b), construct an interval estimate for Thurio's score on one round if he switches to the new clubs, at 90% confidence.

25. Large Data Set 13 records the number of bidders and sales price of a particular type of antique grandfather clock at 60 auctions.

http://www.gone.2012books.lardbucket.org/sites/all/files/data13.xls

a. There are seven likely bidders at the Verona auction today. Give a point estimate for the price of such a clock at today's auction.
b. Explain whether an interval estimate for this problem is a confidence interval or a prediction interval.
c. Based on your answer to (b), construct an interval estimate for the likely sale price of such a clock at today's sale, at 95% confidence.

ANSWERS

1. a. 5.647,
   b. 5.647 ± 1.253
3. a. -0.188,
   b. -0.188 ± 3.041
5. a. 1.875,
   b. 1.875 ± 1.423
7. a. 5.4,
   b. 5.4 ± 3.355 ,
   c. invalid (extrapolation)
9. a. 2.4,
   b. 2.4 ± 1.474 ,
   c. valid (-1 is in the range of the x-values in the data set)
11. a. 31.3 words,
    b. 31.3 ± 7.1 words,
    c. not valid, since two years is 24 months, hence this is extrapolation
13. a. 73.2 beats/min,
    b. The man's heart rate is not the predicted average for all men his age. c. 73.2 ± 1.2 beats/min
15. a. $224,562,
    b. $224,562 ± $28,699
17. a. 74,
    b. Prediction (one person, not an average for all who have average 78.6 before the final exam),
    c. 74 ± 24
19. a. 0.066%,
    b. $0.066 \pm 0.034\%$
21. a. 4,656 psi,
    b. $4{,}656 \pm 321$ psi,
    c. $4{,}656 - 321 = 4{,}335$ psi
23. a. 2.19
    b. $(2.1421, 2.2316)$
25. a. 7771.39
    b. A prediction interval.

c. (7410. 41,8132. 38)

10.8 A Complete Example

LEARNING OBJECTIVE

1. To see a complete linear correlation and regression analysis, in a practical setting, as a cohesive whole.

In the preceding sections numerous concepts were introduced and illustrated, but the analysis was broken into disjoint pieces by sections. In this section we will go through a complete example of the use of correlation and regression analysis of data from start to finish, touching on all the topics of this chapter in sequence.

In general educators are convinced that, all other factors being equal, class attendance has a significant bearing on course performance. To investigate the relationship between attendance and performance, an education researcher selects for study a multiple section introductory statistics course at a large university. Instructors in the course agree to keep an accurate record of attendance throughout one semester. At the end of the semester 26 students are selected a random. For each student in the sample two measurements are taken: x, the number of days the student was absent, and y, the student's score on the common final exam in the course. The data are summarized in Table 10.4 "Absence and Score Data".

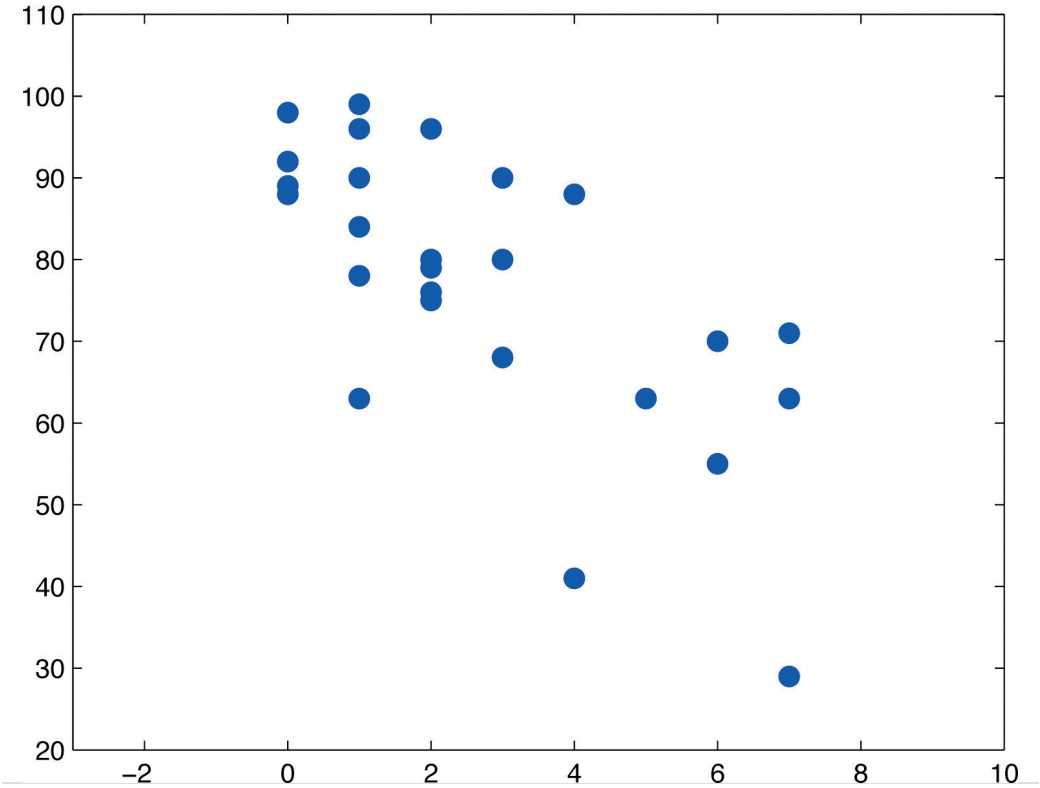
Table 10.4 Absence and Score Data

Absences	Score	Absences	Score
x	y	x	y
2	76	4	41
7	29	5	63
2	96	4	88
7	63	0	98
2	79	1	99
7	71	0	89
0	88	1	96

Absences	Score	Absences	Score
x	*y*	*x*	*y*
0	92	3	90
6	55	1	90
6	70	3	68
2	80	1	84
2	75	3	80
1	63	1	78

A scatter plot of the data is given in Figure 10.13 "Plot of the Absence and Exam Score Pairs". There is a downward trend in the plot which indicates that on average students with more absences tend to do worse on the final examination.

Figure 10.13 *Plot of the Absence and Exam Score Pairs*



The trend observed in Figure 10.13 "Plot of the Absence and Exam Score Pairs" as well as the fairly constant width of the apparent band of points in the plot makes it reasonable to assume a relationship between x and y of the form

$$y = \beta_1 x + \beta_0 + \varepsilon$$

where β_1 and β_0 are unknown parameters and ε is a normal random variable with mean zero and unknown standard deviation σ. Note carefully that this model is being proposed for the population of all students taking this course, not just those taking it this semester, and certainly not just those in the sample. The numbers β_1, β_0, and σ are parameters relating to this large population.

First we perform preliminary computations that will be needed later. The data are processed in Table 10.5 "Processed Absence and Score Data".

Table 10.5 Processed Absence and Score Data

x	y	x^2	xy	y^2	x	y	x^2	xy	y^2
2	76	4	152	5776	4	41	16	164	1681
7	29	49	203	841	5	63	25	315	3969
2	96	4	192	9216	4	88	16	352	7744
7	63	49	441	3969	0	98	0	0	9604
2	79	4	158	6241	1	99	1	99	9801
7	71	49	497	5041	0	89	0	0	7921
0	88	0	0	7744	1	96	1	96	9216
0	92	0	0	8464	3	90	9	270	8100
6	55	36	330	3025	1	90	1	90	8100
6	70	36	420	4900	3	68	9	204	4624
2	80	4	160	6400	1	84	1	84	7056
2	75	4	150	5625	3	80	9	240	6400
1	63	1	63	3969	1	78	1	78	6084

Adding up the numbers in each column in Table 10.5 "Processed Absence and Score Data" gives

$$\Sigma x = 71, \quad \Sigma y = 2001, \quad \Sigma x^2 = 329, \quad \Sigma xy = 4758, \quad \text{and} \quad \Sigma y^2 = 16151$$

Then

$$SS_{xx} = \Sigma x^2 - \frac{1}{n}(\Sigma x)^2 = 329 - \frac{1}{26}(71)^2 = 135.1153846$$
$$SS_{xy} = \Sigma xy - \frac{1}{n}(\Sigma x)\left(\Sigma y\right) = 4758 - \frac{1}{26}(71)(2001) = -706.269230$$
$$SS_{yy} = \Sigma y^2 - \frac{1}{n}\left(\Sigma y\right)^2 = 161511 - \frac{1}{26}(2001)^2 = 7510.961538$$

and

$$\overline{x} = \frac{\Sigma x}{n} = \frac{71}{26} = 2.730769231 \quad \text{and} \quad \overline{y} = \frac{\Sigma y}{n} = \frac{2001}{26} = 76.96153846$$

We begin the actual modelling by finding the least squares regression line, the line that best fits the data. Its slope and y-intercept are

$$\widehat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{-706.2692308}{135.1153846} = -5.227156278$$
$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1\overline{x} = 76.96153846 - \left(-5.227156278\right)\left(2.730769231\right) = 9$$

Rounding these numbers to two decimal places, the least squares regression line for these data is

$$\widehat{y} = -5.23\,x + 91.24.$$

The goodness of fit of this line to the scatter plot, the sum of its squared errors, is

$$SSE = SS_{yy} - \widehat{\beta}_1 SS_{xy} = 7510.961538 - \left(-5.227156278\right)\left(-706.26923\right.$$

This number is not particularly informative in itself, but we use it to compute the important statistic

$$s_\varepsilon = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{3819.181894}{24}} = 12.11988495$$

The statistic s_ε estimates the standard deviation σ of the normal random variable ε in the model. Its meaning is that among all students with the same number of absences, the standard deviation of their scores on the final exam is about 12.1 points. Such a large value on a 100-point exam means that the final exam scores of

each sub-population of students, based on the number of absences, are highly variable.

The size and sign of the slope $\widehat{\beta}_1 = -5.23$ indicate that, for every class missed, students tend to score about 5.23 fewer points lower on the final exam on average. Similarly for every two classes missed students tend to score on average $2 \times 5.23 = 10.46$ fewer points on the final exam, or about a letter grade worse on average.

Since 0 is in the range of *x*-values in the data set, the *y*-intercept also has meaning in this problem. It is an estimate of the average grade on the final exam of all students who have perfect attendance. The predicted average of such students is $\widehat{\beta}_0 = 91.24$.

Before we use the regression equation further, or perform other analyses, it would be a good idea to examine the utility of the linear regression model. We can do this in two ways: 1) by computing the correlation coefficient *r* to see how strongly the number of absences *x* and the score *y* on the final exam are correlated, and 2) by testing the null hypothesis $H_0 : \beta_1 = 0$ (the slope of the *population* regression line is zero, so *x* is not a good predictor of *y*) against the natural alternative $H_a : \beta_1 < 0$ (the slope of the population regression line is negative, so final exam scores *y* go down as absences *x* go up).

The correlation coefficient *r* is

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}\,SS_{yy}}} = \frac{-706.2692308}{\sqrt{(135.1153846)\ (7510.961538)}} = -0.70108409$$

a moderate negative correlation.

Turning to the test of hypotheses, let us test at the commonly used 5% level of significance. The test is

$$H_0 : \beta_1 = 0$$
$$\text{vs. } H_a : \beta_1 < 0 \quad @\ \alpha = 0.05$$

From Figure 12.3 "Critical Values of ", with $df = 26 - 2 = 24$ degrees of freedom $t_{0.05} = 1.711$, so the rejection region is $\left(-\infty, -1.711\right]$. The value of the standardized test statistic is

$$t = \frac{\widehat{\beta}_1 - B_0}{s_\varepsilon \Big/ \sqrt{SS_{xx}}} = \frac{-5.227156278 - 0}{12.11988495 \Big/ \sqrt{135.1153846}} = -5.013$$

which falls in the rejection region. We reject H_0 in favor of H_a. The data provide sufficient evidence, at the 5% level of significance, to conclude that β_1 is negative, meaning that as the number of absences increases average score on the final exam decreases.

As already noted, the value $\beta_1 = -5.23$ gives a point estimate of how much one additional absence is reflected in the average score on the final exam. For each additional absence the average drops by about 5.23 points. We can widen this point estimate to a confidence interval for β_1. At the 95% confidence level, from Figure 12.3 "Critical Values of " with $df = 26 - 2 = 24$ degrees of freedom, $t_{\alpha/2} = t_{0.025} = 2.064$. The 95% confidence interval for β_1 based on our sample data is

$$\widehat{\beta}_1 \pm t_{\alpha/2} \frac{s_\varepsilon}{\sqrt{SS_{xx}}} = -5.23 \pm 2.064 \frac{12.11988495}{\sqrt{135.1153846}} = -5.23 \pm 2.15$$

or $(-7.38, -3.08)$. We are 95% confident that, among all students who ever take this course, for each additional class missed the average score on the final exam goes down by between 3.08 and 7.38 points.

If we restrict attention to the sub-population of all students who have exactly five absences, say, then using the least squares regression equation $\widehat{y} = -5.23x + 91.24$ we estimate that the average score on the final exam for those students is

$$\widehat{y} = -5.23\left(5\right) + 91.24 = 65.09$$

This is also our best guess as to the score on the final exam of any particular student who is absent five times. A 95% confidence interval for the average score on the final exam for all students with five absences is

$$\hat{y}_p \pm t_{\alpha/2} s_\varepsilon \sqrt{\frac{1}{n} + \frac{\left(x_p - \overline{x}\right)^2}{SS_{xx}}} = 65.09 \pm \left(2.064\right)\left(12.11988495\right)\sqrt{\frac{1}{26}}$$

$$= 65.09 \pm 25.01544254\sqrt{0.0765727299}$$

$$= 65.09 \pm 6.92$$

which is the interval $\left(58.17, 72.01\right)$. This confidence interval suggests that the true mean score on the final exam for all students who are absent from class exactly five times during the semester is likely to be between 58.17 and 72.01.

If a particular student misses exactly five classes during the semester, his score on the final exam is predicted with 95% confidence to be in the interval

$$\hat{y}_p \pm t_{\alpha/2} s_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{\left(x_p - \overline{x}\right)^2}{SS_{xx}}} = 65.09 \pm 25.01544254\sqrt{1.0765727}$$

$$= 65.09 \pm 25.96$$

which is the interval $\left(39.13, 91.05\right)$. This prediction interval suggests that this individual student's final exam score is likely to be between 39.13 and 91.05. Whereas the 95% confidence interval for the average score of all student with five absences gave real information, this interval is so wide that it says practically nothing about what the individual student's final exam score might be. This is an example of the dramatic effect that the presence of the extra summand 1 under the square sign in the prediction interval can have.

Finally, the proportion of the variability in the scores of students on the final exam that is explained by the linear relationship between that score and the number of absences is estimated by the coefficient of determination, r^2. Since we have already computed r above we easily find that

$$r^2 = (-0.7010840977)^2 = 0.491518912$$

or about 49%. Thus although there is a significant correlation between attendance and performance on the final exam, and we can estimate with fair accuracy the average score of students who miss a certain number of classes, nevertheless less than half the total variation of the exam scores in the sample is explained by the number of absences. This should not come as a surprise, since there are many factors besides attendance that bear on student performance on exams.

KEY TAKEAWAY

- It is a good idea to attend class.

EXERCISES

The exercises in this section are unrelated to those in previous sections.

1. The data give the amount x of silicofluoride in the water (mg/L) and the amount y of lead in the bloodstream (μg/dL) of ten children in various communities with and without municipal water. Perform a complete analysis of the data, in analogy with the discussion in this section (that is, make a scatter plot, do preliminary computations, find the least squares regression line, find SSE, s_ε, and r, and so on). In the hypothesis test use as the alternative hypothesis $\beta_1 > 0$, and test at the 5% level of significance. Use confidence level 95% for the confidence interval for β_1. Construct 95% confidence and predictions intervals at $x_p = 2$ at the end.

x	0.0	0.0	1.1	1.4	1.6
y	0.3	0.1	4.7	3.2	5.1
x	1.7	2.0	2.0	2.2	2.2
y	7.0	5.0	6.1	8.6	9.5

2. The table gives the weight x (thousands of pounds) and available heat energy y (million BTU) of a standard cord of various species of wood typically used for heating. Perform a complete analysis of the data, in analogy with the discussion in this section (that is, make a scatter plot, do preliminary computations, find the least squares regression line, find SSE, s_ε, and r, and so on). In the hypothesis test use as the alternative hypothesis $\beta_1 > 0$, and test at the 5% level of significance. Use confidence level 95% for the confidence interval for β_1. Construct 95% confidence and predictions intervals at $x_p = 5$ at the end.

x	3.37	3.50	4.29	4.00	4.64
y	23.6	17.5	20.1	21.6	28.1
x	4.99	4.94	5.48	3.26	4.16
y	25.3	27.0	30.7	18.9	20.7

LARGE DATA SET EXERCISES

3. Large Data Sets 3 and 3A list the shoe sizes and heights of 174 customers entering a shoe store. The gender of the customer is not indicated in Large Data Set 3. However, men's and women's shoes are not measured on the same scale; for example, a size 8 shoe for men is not the same size as a size 8 shoe for

women. Thus it would not be meaningful to apply regression analysis to Large Data Set 3. Nevertheless, compute the scatter diagrams, with shoe size as the independent variable (x) and height as the dependent variable (y), for (i) just the data on men, (ii) just the data on women, and (iii) the full mixed data set with both men and women. Does the third, invalid scatter diagram look markedly different from the other two?

http://www.gone.2012books.lardbucket.org/sites/all/files/data3.xls

http://www.gone.2012books.lardbucket.org/sites/all/files/data3A.xls

4. Separate out from Large Data Set 3A just the data on men and do a complete analysis, with shoe size as the independent variable (x) and height as the dependent variable (y). Use $\alpha = 0.05$ and $x_p = 10$ whenever appropriate.

   http://www.gone.2012books.lardbucket.org/sites/all/files/data3A.xls

5. Separate out from Large Data Set 3A just the data on women and do a complete analysis, with shoe size as the independent variable (x) and height as the dependent variable (y). Use $\alpha = 0.05$ and $x_p = 10$ whenever appropriate.

   http://www.gone.2012books.lardbucket.org/sites/all/files/data3A.xls

ANSWERS

1. $\Sigma x = 14.2, \Sigma y = 49.6, \Sigma xy = 91.73, \Sigma x^2 = 26.3,$
   $\Sigma y^2 = 333.86.$

   $SS_{xx} = 6.136, SS_{xy} = 21.298, SS_{yy} = 87.844.$

   $\overline{x} = 1.42, \overline{y} = 4.96.$

   $\widehat{\beta}_1 = 3.47, \widehat{\beta}_0 = 0.03.$

   $SSE = 13.92.$

   $s_\varepsilon = 1.32.$

   $r = 0.9174$, $r^2 = 0.8416.$

   $df = 8$, $T = 6.518.$

   The 95% confidence interval for β_1 is: $(2.24, 4.70)$.

   At $x_p = 2$, the 95% confidence interval for $E(y)$ is $(5.77, 8.17)$.

   At $x_p = 2$, the 95% prediction interval for y is $(3.73, 10.21)$.

3. The positively correlated trend seems less profound than that in each of the previous plots.

5. The regression line: $\widehat{y} = 3.3426x + 138.7692.$ Coefficient of Correlation: $r = 0.9431$. Coefficient of Determination: $r^2 = 0.8894.$ $SSE = 283.2473.$ $s_e = 1.9305.$ A 95% confidence interval for β_1: $(3.0733, 3.6120)$. Test Statistic for $H_0 : \beta_1 = 0$ $T = 24.7209$. At $x_p = 10, \widehat{y} = 172.1956$; a 95% confidence interval for the mean value of y is: $(171.5577, 172.8335)$; and a 95% prediction interval for an individual value of y is: $(168.2974, 176.0938)$.

10.9 Formula List

$$SS_{xx} = \Sigma x^2 - \frac{1}{n}(\Sigma x)^2 \qquad SS_{xy} = \Sigma xy - \frac{1}{n}(\Sigma x)(\Sigma y) \qquad SS_{yy} = \Sigma y^2 -$$

Correlation coefficient:

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} \cdot SS_{yy}}}$$

Least squares regression equation (equation of the least squares regression line):

$$\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0 \quad \text{where} \quad \hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \quad \text{and} \quad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

Sum of the squared errors for the least squares regression line:

$$SSE = SS_{yy} - \hat{\beta}_1 SS_{xy}.$$

Sample standard deviation of errors:

$$s_\varepsilon = \sqrt{\frac{SSE}{n-2}}$$

$100(1-\alpha)\,\%$ confidence interval for β_1:

$$\hat{\beta}_1 \pm t_{\alpha/2} \frac{s_\varepsilon}{\sqrt{SS_{xx}}} \qquad (df = n-2)$$

Standardized test statistic for hypothesis tests concerning β_1:

$$T = \frac{\hat{\beta}_1 - B_0}{s_\varepsilon \Big/ \sqrt{SS_{xx}}} \qquad (df = n-2)$$

Coefficient of determination:

$$r^2 = \frac{SS_{yy} - SSE}{SS_{yy}} = \frac{SS_{xy}^2}{SS_{xx}\,SS_{yy}} = \widehat{\beta}_1 \frac{SS_{xy}}{SS_{yy}}$$

$100\,(1-\alpha)\,\%$ confidence interval for the mean value of y at $x = x_p$:

$$\widehat{y}_p \pm t_{\alpha/2}\, s_\varepsilon \sqrt{\frac{1}{n} + \frac{\left(x_p - \overline{x}\right)^2}{SS_{xx}}} \quad \left(df = n{-}2\right)$$

$100\,(1-\alpha)\,\%$ prediction interval for an individual new value of y at $x = x_p$:

$$\widehat{y}_p \pm t_{\alpha/2}\, s_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{\left(x_p - \overline{x}\right)^2}{SS_{xx}}} \quad \left(df = n{-}2\right)$$