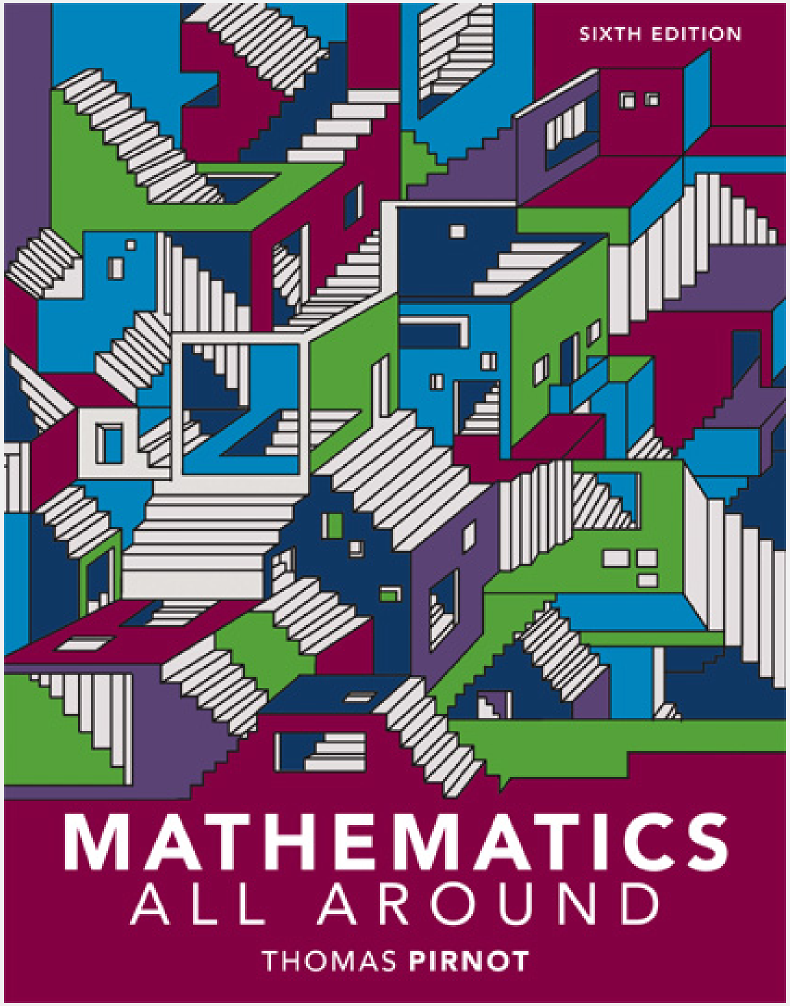
14.5

Linear Correlation and Regression Analysis



14.5 Linear Correlation

- Construct a scatterplot to show the relationship between two variables.
- Explain the properties of the linear correlation coefficient.
- Use linear regression to find the line of best fit for a set of data points.

Scatterplots

To determine whether there is a *correlation* between two variables, we obtain pairs of data, called *data points*, relating the first variable to the second.

To understand such data, we plot data points in a graph called a *scatterplot*.

Example: Constructing a Scatterplot

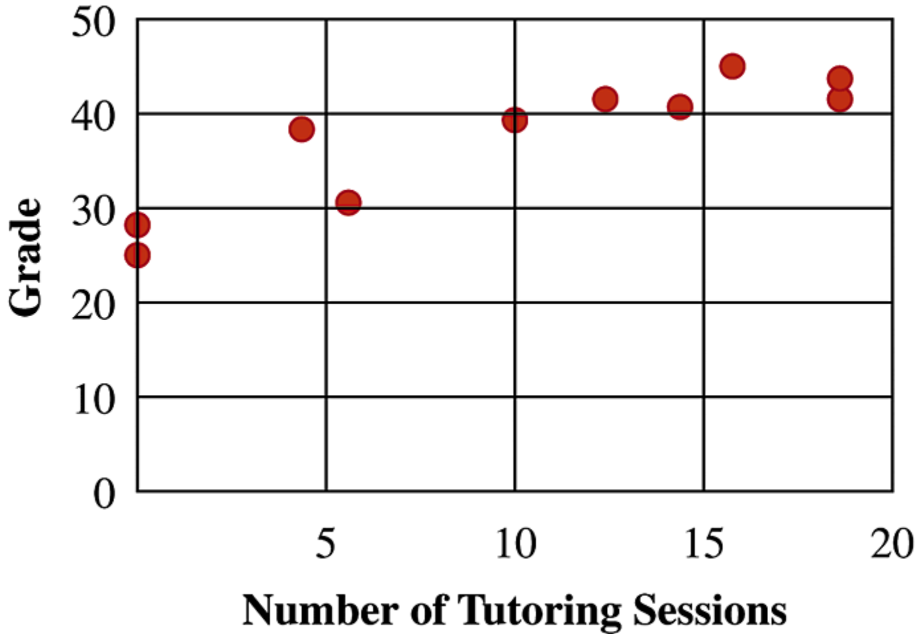
An instructor wants to know whether there is a correlation between the number of times students attended tutoring sessions during the semester and their grades on a 50-point examination. The instructor has collected data for 10 students, as shown in the table. Represent these data points by a scatterplot and interpret the graph.

Number of Tutoring Session, *x*	18	6	16	14	0	4	0	10	12	18
Exam Grade, *y*	42	31	46	41	25	38	28	39	42	44

Example: Constructing a Scatterplot (cont)

Solution

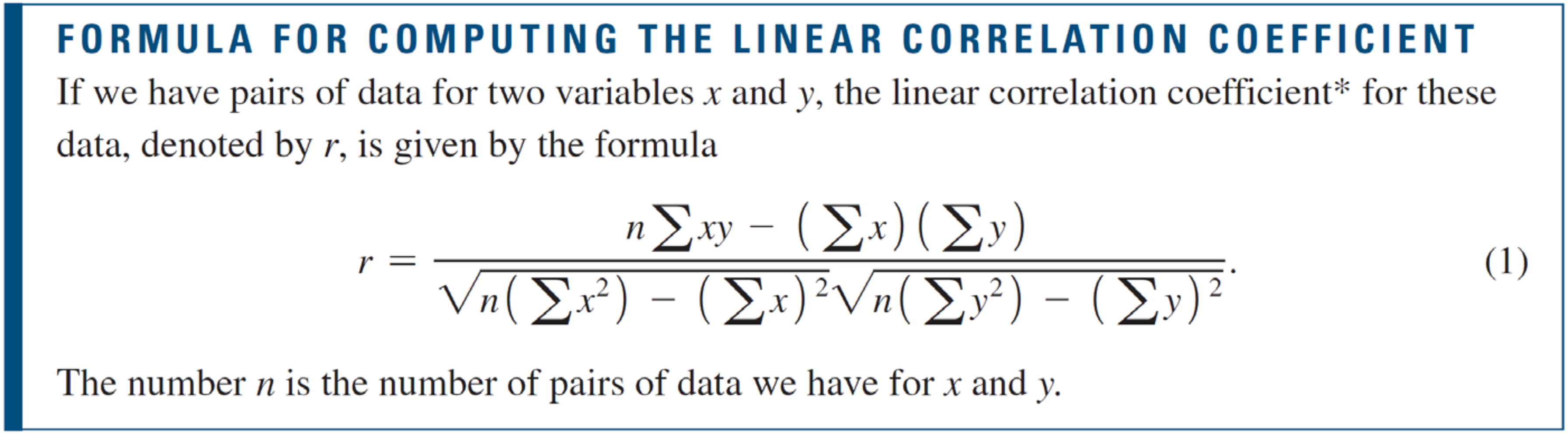
We plot the points (18, 42), (6, 31), (16, 46), and so on.



As the number of tutoring sessions increases, the grades also generally increase.

Linear Correlation

Linear correlation exists between two variables if, when graphed, the points in the graph tend to lie in a straight line. The *linear correlation coefficient* allows us to compute to what degree the points of a scatterplot lie along a straight line.



FORMULA FOR COMPUTING THE LINEAR CORRELATION COEFFICIENT

If we have pairs of data for two variables x and y, the linear correlation coefficient* for these data, denoted by r, is given by the formula

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}}. \qquad (1)$$

The number n is the number of pairs of data we have for x and y.

Example: Computing the Linear Correlation Coefficient

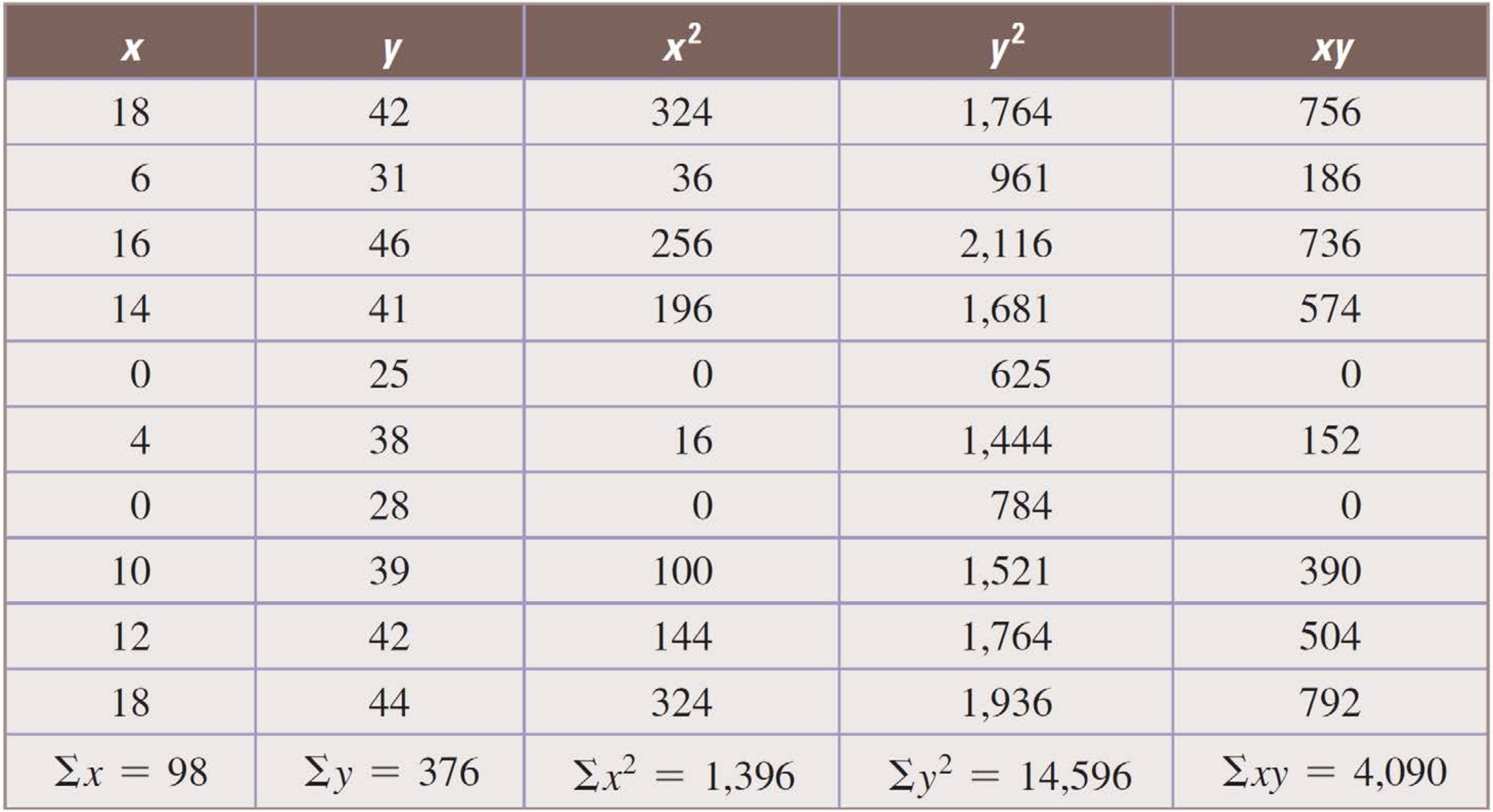
Compute the linear correlation coefficient for the data.

Number of Tutoring Session, *x*	18	6	16	14	0	4	0	10	12	18
Exam Grade, *y*	42	31	46	41	25	38	28	39	42	44

Solution

We begin by creating a table with the necessary components of the linear correlation coefficient.

Example: Computing the Linear Correlation Coefficient (cont)



x	y	x^2	y^2	xy
18	42	324	1,764	756
6	31	36	961	186
16	46	256	2,116	736
14	41	196	1,681	574
0	25	0	625	0
4	38	16	1,444	152
0	28	0	784	0
10	39	100	1,521	390
12	42	144	1,764	504
18	44	324	1,936	792
$\Sigma x = 98$	$\Sigma y = 376$	$\Sigma x^2 = 1,396$	$\Sigma y^2 = 14,596$	$\Sigma xy = 4,090$

Example: Computing the Linear Correlation Coefficient (cont)

The linear correlation coefficient is

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

$$= \frac{10 \cdot 4,090 - (98)(376)}{\sqrt{10 \cdot (1,396) - 98^2}\sqrt{10 \cdot (14,596) - 376^2}}$$

$$= \frac{4,052}{\sqrt{4,356}\sqrt{4,584}} \approx 0.9068.$$

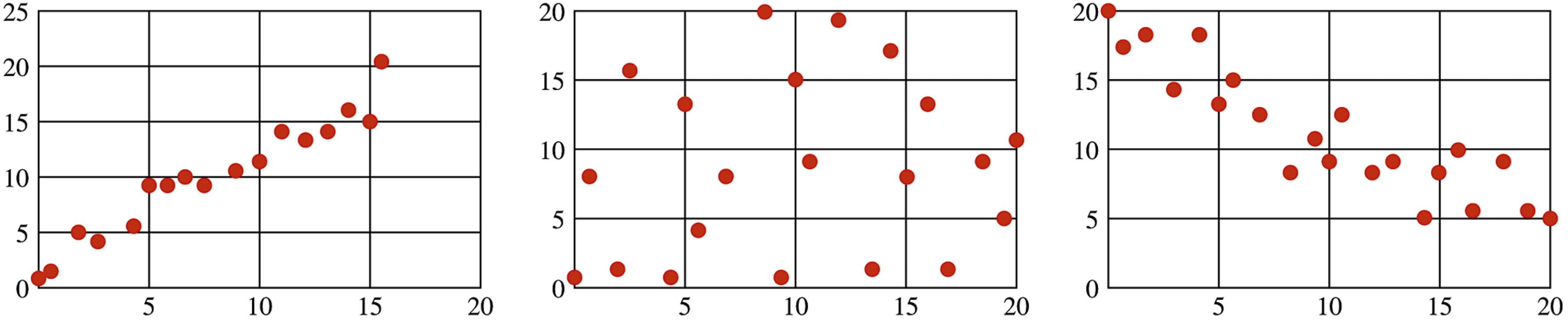
Properties of the Linear Correlation Coefficient

1. *r* is a number between –1 and 1, inclusive. If *r* = 1 or –1, then all of the points lie on a straight line.
2. If *r* is positive, there is positive correlation between the variables. If *r* is negative, there is negative correlation between the variables.
3. If *r* is close to 1, then there is a significant positive linear correlation between the variables and the scatterplot is close to lying along a straight line that rises from left to right.

Properties of the Linear Correlation Coefficient

4. If *r* is close to –1, then there is a significant negative linear correlation between the variables and the scatterplot is close to lying along a straight line that falls from left to right.
5. If *r* is close to 0, then there is little linear correlation between the variables.

Linear Correlation



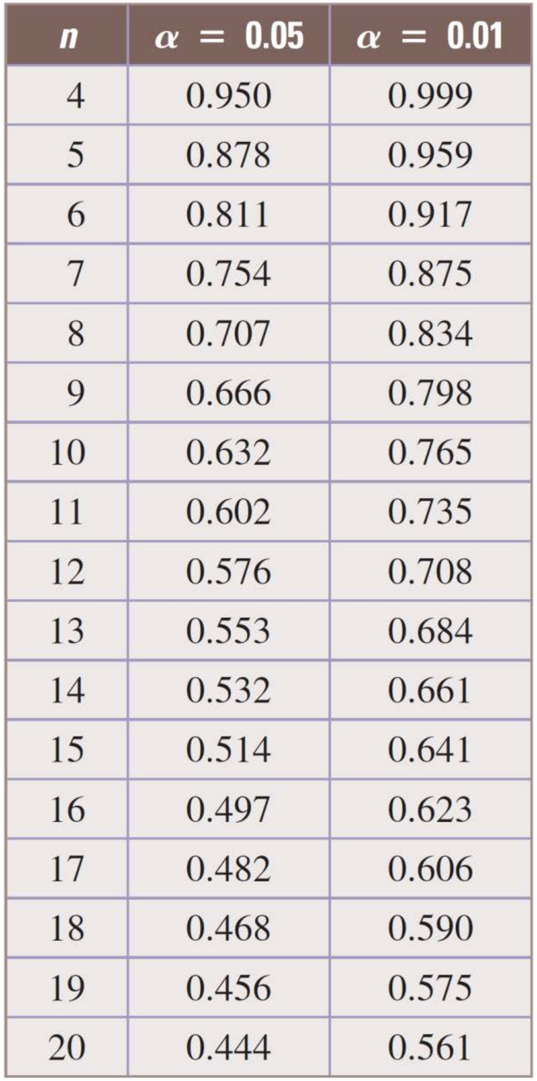
r is very close to 1.
(a)

r is close to 0.
(b)

r is negative.
(c)

There is a *positive correlation* between the variables x and y if whenever x increases or decreases, then y changes in the same way. We will say that there is a *negative correlation* between the variables x and y if whenever x increases or decreases, then y changes in the opposite way.

Linear Correlation

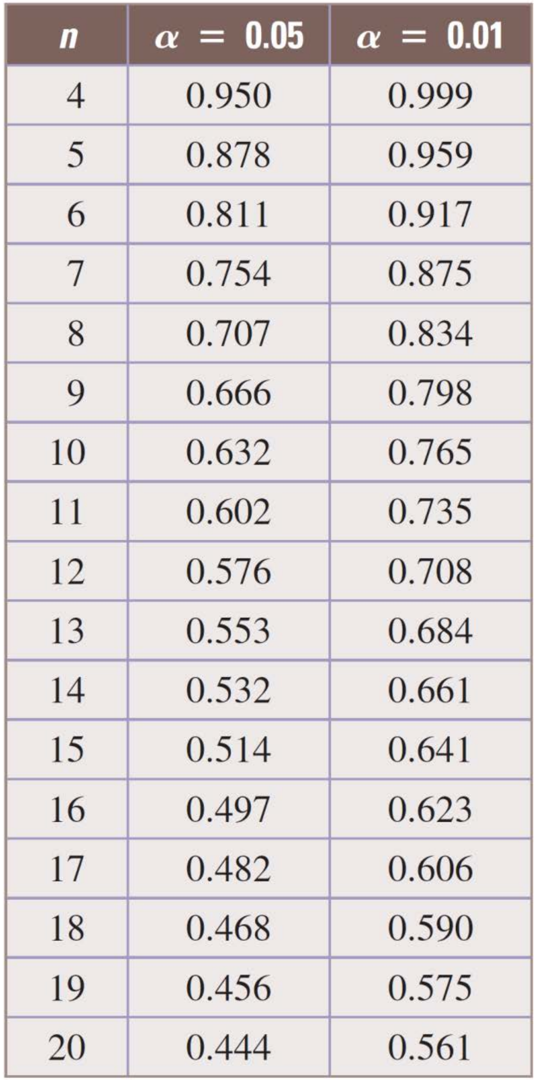


n	$\alpha = 0.05$	$\alpha = 0.01$
4	0.950	0.999
5	0.878	0.959
6	0.811	0.917
7	0.754	0.875
8	0.707	0.834
9	0.666	0.798
10	0.632	0.765
11	0.602	0.735
12	0.576	0.708
13	0.553	0.684
14	0.532	0.661
15	0.514	0.641
16	0.497	0.623
17	0.482	0.606
18	0.468	0.590
19	0.456	0.575
20	0.444	0.561

1. Compute the linear correlation coefficient *r* for *n* pairs of data.

2. Go to line *n* in the table.

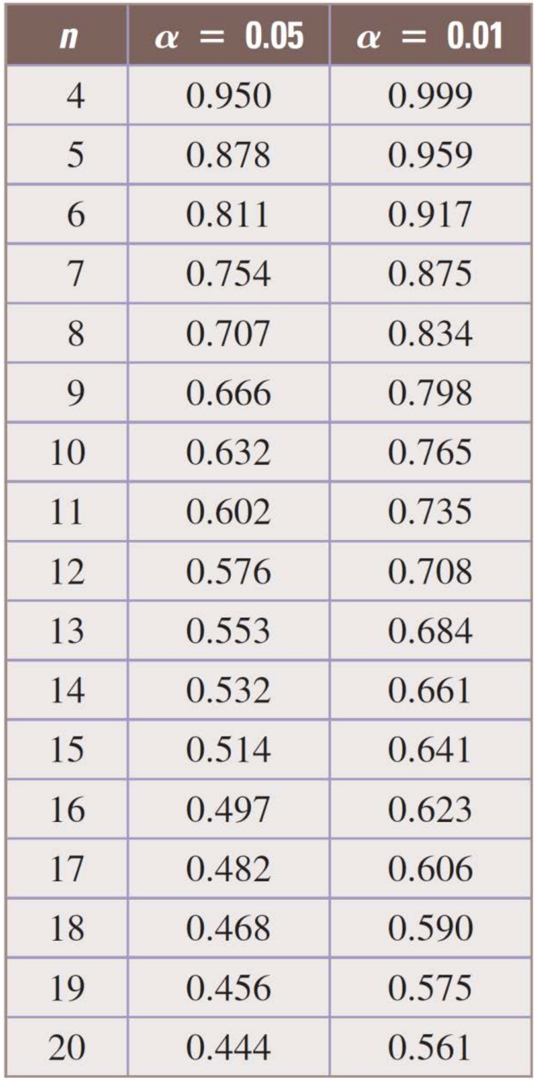
Linear Correlation (cont)



n	$\alpha = 0.05$	$\alpha = 0.01$
4	0.950	0.999
5	0.878	0.959
6	0.811	0.917
7	0.754	0.875
8	0.707	0.834
9	0.666	0.798
10	0.632	0.765
11	0.602	0.735
12	0.576	0.708
13	0.553	0.684
14	0.532	0.661
15	0.514	0.641
16	0.497	0.623
17	0.482	0.606
18	0.468	0.590
19	0.456	0.575
20	0.444	0.561

3. If the absolute value of *r* exceeds the number in the column labeled α = 0.05, we can be 95% confident that there is significant linear correlation between the variables. That is, there is only a 5% chance we will be incorrect if we conclude that there is a linear correlation between the variables.

Linear Correlation (cont)



n	$\alpha = 0.05$	$\alpha = 0.01$
4	0.950	0.999
5	0.878	0.959
6	0.811	0.917
7	0.754	0.875
8	0.707	0.834
9	0.666	0.798
10	0.632	0.765
11	0.602	0.735
12	0.576	0.708
13	0.553	0.684
14	0.532	0.661
15	0.514	0.641
16	0.497	0.623
17	0.482	0.606
18	0.468	0.590
19	0.456	0.575
20	0.444	0.561

4. If the absolute value of r exceeds the number in the column labeled α = 0.01, we can be 99% confident that there is significant linear correlation between the variables. That is, there is only a 1% chance we will be incorrect if we conclude that there is a linear correlation between the variables.

Example: Determining Correlation between Car Weight and Mileage

Let us suppose that you are considering buying a used car. After talking to several car dealers, you are convinced that there is a relationship between the weight of the car and gas mileage. To check this, you gather data regarding the gas mileage of cars with various weights, to see whether there is a linear correlation between weight and gas mileage for these data.

Weight (hundreds of pounds)	29	28	31	24	25	30	24	28	32	26
City MPG	21	22	22	23	23	21	24	21	20	22

Example: Determining Correlation between Car Weight and Mileage (cont)

Find the correlation coefficient for these data and determine whether there is significant linear correlation at the 5% or 1% level.

Solution

We will represent the weight by *x* and the gas mileage by *y*. We find that

$$\Sigma x = 277,\ \Sigma y = 219,\ \Sigma x^2 = 7{,}747,\ \Sigma y^2 = 4{,}809$$

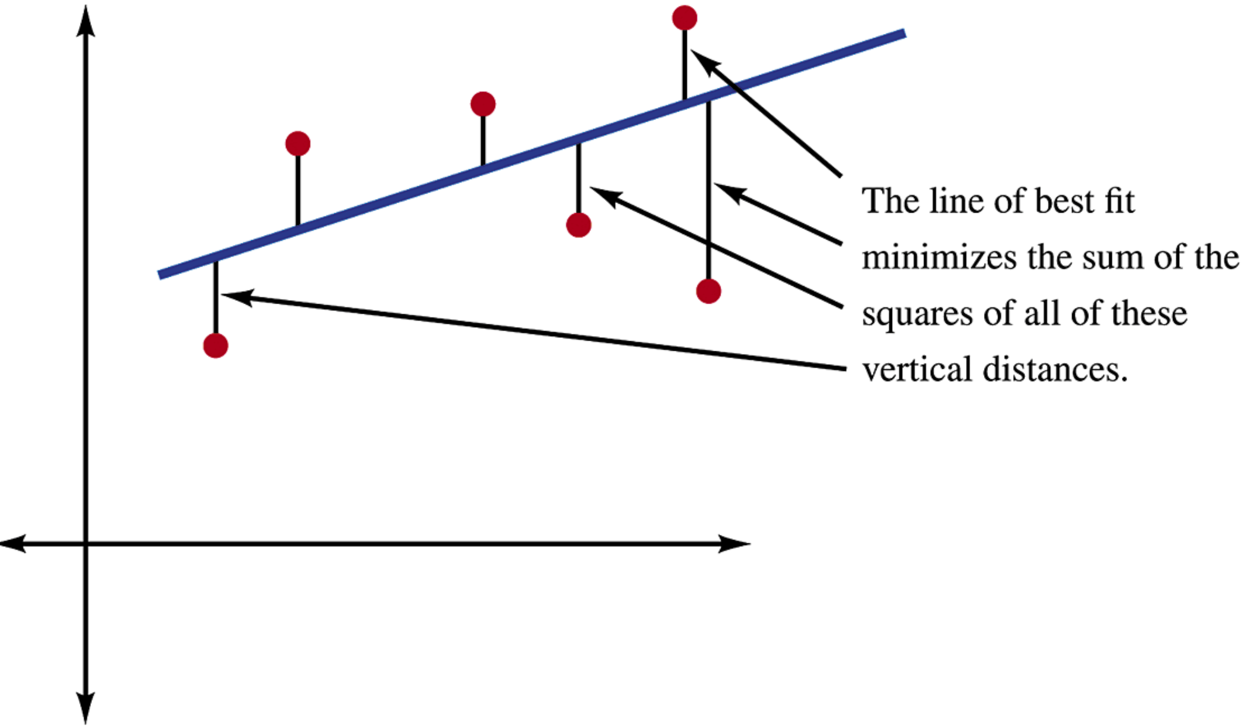
$$\text{and } \Sigma xy = 6{,}040.$$

We now compute *r* = –0.8507.

Example: Determining Correlation between Car Weight and Mileage (cont)

Because there are 10 pairs of data, we will use line 10 of the table to determine how confident we can be that there is a significant linear correlation. The absolute value of *r* is 0.85, which exceeds 0.765 in line 10 of the table. Therefore, we can be 99% confident that there is significant negative linear correlation between the variables car weight and gas mileage.

The Line of Best Fit

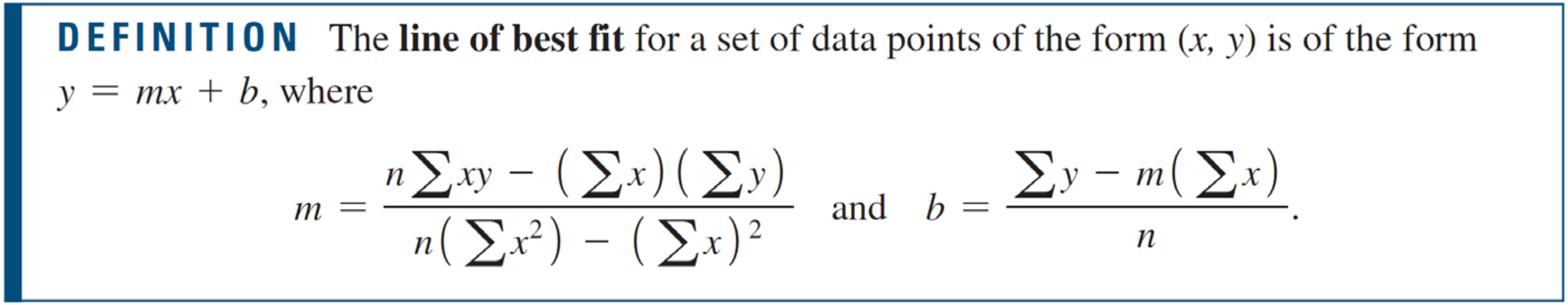


The line of best fit minimizes the sum of the squares of all of these vertical distances.

We wish to find the line that best models our data.

This line is called *the line of best fit*. It is the line that minimizes the sum of all the vertical distances from the data points to the line.

The Line of Best Fit



DEFINITION The **line of best fit** for a set of data points of the form (x, y) is of the form $y = mx + b$, where

$$m = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2} \quad \text{and} \quad b = \frac{\sum y - m\left(\sum x\right)}{n}.$$

Example: Finding the Line of Best Fit for a Set of Data Points

Find the line of best fit.

Number of Tutoring Session, *x*	18	6	16	14	0	4	0	10	12	18
Exam Grade, *y*	42	31	46	41	25	38	28	39	42	44

Solution

We compute

$$n = 10,\ \Sigma xy = 4{,}090,\ \Sigma x = 98,\ \Sigma y = 376,\ \Sigma x^2 = 1{,}396.$$

Example: Finding the Line of Best Fit for a Set of Data Points (cont)

Slope:

$$m = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2} = \frac{10 \cdot 4,090 - 98 \cdot 376}{10 \cdot 1,396 - (98)^2}$$

$$= \frac{40,900 - 36,848}{13,960 - 9,604}$$

$$= \frac{4,052}{4,356}$$

$$\approx 0.9302.$$

Example: Finding the Line of Best Fit for a Set of Data Points (cont)

y-intercept:

$$b = \frac{\sum y - m\left(\sum x\right)}{n} = \frac{376 - (0.9302)98}{10}$$

$$= \frac{376 - 91.1596}{10}$$

$$\approx \frac{284.84}{10} = 28.48$$

Example: Finding the Line of Best Fit for a Set of Data Points (cont)

We can interpret the slope and intercept of this equation. A student with 0 tutoring sessions would expect a score of approximately 28.5. For each additional tutoring session, we expect the exam score to increase by just under 1 point (0.93 point). Note that our slope here is positive, as was the linear correlation coefficient (0.907) we found in an earlier example.