Linear Regression

Objectives

1. Determine and interpret the linear correlation coefficient
2. Determine the linear regression equation

Linear Correlation Coefficient

In the previous section, we examined correlation types (positive, negative, or none) with the help of the means of the explanatory (x) and response variables (y).

Linear Correlation Coefficient

In the previous section, we examined correlation types (positive, negative, or none) with the help of the means of the explanatory (x) and response variables (y).

In this section, we will examine the correlation type the way it is done in the real world: calculating the linear correlation coefficient (r).

Linear Correlation Coefficient

Correlation Coefficient

The **correlation coefficient**, r, is a numerical value with $-1 \le r \le 1$ that measures the type of linear correlation of a bivariate dataset.

Linear Correlation Coefficient

Correlation Coefficient

The **correlation coefficient**, r, is a numerical value with $-1 \leq r \leq 1$ that measures the type of linear correlation of a bivariate dataset.

- $r > 0$: positive linear correlation

Linear Correlation Coefficient

Correlation Coefficient

The **correlation coefficient**, r, is a numerical value with $-1 \leq r \leq 1$ that measures the type of linear correlation of a bivariate dataset.

- $r > 0$: positive linear correlation
- $r = 0$: no linear correlation

Linear Correlation Coefficient

Correlation Coefficient

The **correlation coefficient**, r, is a numerical value with $-1 \leq r \leq 1$ that measures the type of linear correlation of a bivariate dataset.

- $r > 0$: positive linear correlation
- $r = 0$: no linear correlation
- $r < 0$: negative linear correlation

Linear Correlation Coefficient

$$r = \frac{\sum(x-\overline{x})(y-\overline{y})}{\sqrt{\sum(x-\overline{x})^2 \cdot \sum(y-\overline{y})^2}}$$

Linear Correlation Coefficient

$$r = \frac{\sum(x - \overline{x})(y - \overline{y})}{\sqrt{\sum(x - \overline{x})^2 \cdot \sum(y - \overline{y})^2}}$$

We will use technology to calculate r

Linear Correlation Coefficient

$$r = \frac{\sum(x-\overline{x})(y-\overline{y})}{\sqrt{\sum(x-\overline{x})^2 \cdot \sum(y-\overline{y})^2}}$$

We will use technology to calculate r

The closer r is to 1 (or -1), the more the data points "fall in line"

Linear Correlation Coefficient

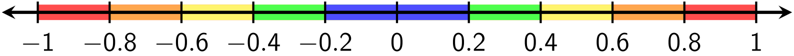
$$r = \frac{\sum(x-\overline{x})(y-\overline{y})}{\sqrt{\sum(x-\overline{x})^2 \cdot \sum(y-\overline{y})^2}}$$

We will use technology to calculate r

The closer r is to 1 (or -1), the more the data points "fall in line"

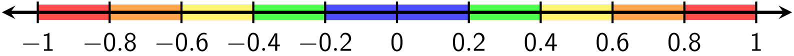
The closer r is to 0, the more the data points resemble a "cloud"

Interpreting r



None
Weak
Moderate
Strong
Very Strong

Interpreting r



None
Weak
Moderate
Strong
Very Strong

Note: These interpretations are not universal.

Example 1

Find and interpret the linear correlation coefficient, r, for each.

(a)

x	y
7.6	19.1
9.2	22.9
3.3	10.3
1.1	6.6
3.7	10.6
3.9	11.3
4.6	12.9
2.3	8.6
5.1	15.2
5.3	15.1
2.5	13
3.4	11.2
3.1	10.6
1.7	6.8
3.7	13.7

Example 1

Find and interpret the linear correlation coefficient, r, for each.

(a)

x	y
7.6	19.1
9.2	22.9
3.3	10.3
1.1	6.6
3.7	10.6
3.9	11.3
4.6	12.9
2.3	8.6
5.1	15.2
5.3	15.1
2.5	13
3.4	11.2
3.1	10.6
1.7	6.8
3.7	13.7

$r \approx 0.9588$

Example 1

Find and interpret the linear correlation coefficient, r, for each.

(a)

x	y
7.6	19.1
9.2	22.9
3.3	10.3
1.1	6.6
3.7	10.6
3.9	11.3
4.6	12.9
2.3	8.6
5.1	15.2
5.3	15.1
2.5	13
3.4	11.2
3.1	10.6
1.7	6.8
3.7	13.7

$r \approx 0.9588$

Very strong positive linear correlation

Example 1

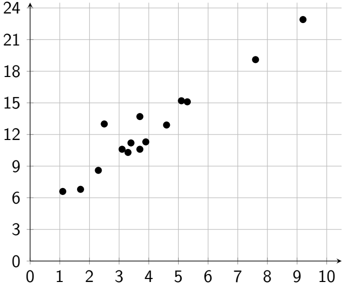
Find and interpret the linear correlation coefficient, r, for each.

(a)

x	y
7.6	19.1
9.2	22.9
3.3	10.3
1.1	6.6
3.7	10.6
3.9	11.3
4.6	12.9
2.3	8.6
5.1	15.2
5.3	15.1
2.5	13
3.4	11.2
3.1	10.6
1.7	6.8
3.7	13.7

$r \approx 0.9588$

Very strong positive linear correlation



Example 1

(b)

x	y
7.6	11.0
9.2	3.6
6.3	8.9
1.1	14.9
6.7	8.1
3.9	12.0
4.6	9.4
2.3	10.3
5.1	11.4
5.3	12.4
2.5	9.0
3.4	8.9
3.1	14.2
1.7	10.9
3.7	13.3

(b)

x	y
7.6	11.0
9.2	3.6
6.3	8.9
1.1	14.9
6.7	8.1
3.9	12.0
4.6	9.4
2.3	10.3
5.1	11.4
5.3	12.4
2.5	9.0
3.4	8.9
3.1	14.2
1.7	10.9
3.7	13.3

$r \approx -0.6273$

Example 1

(b)

x	y
7.6	11.0
9.2	3.6
6.3	8.9
1.1	14.9
6.7	8.1
3.9	12.0
4.6	9.4
2.3	10.3
5.1	11.4
5.3	12.4
2.5	9.0
3.4	8.9
3.1	14.2
1.7	10.9
3.7	13.3

$r \approx -0.6273$

Strong negative linear correlation

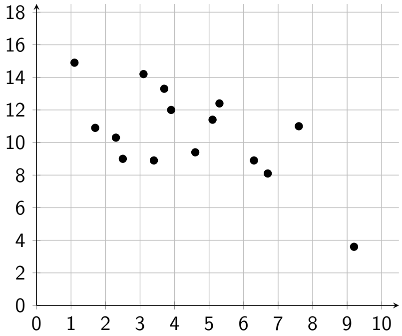
Example 1

(b)

x	y
7.6	11.0
9.2	3.6
6.3	8.9
1.1	14.9
6.7	8.1
3.9	12.0
4.6	9.4
2.3	10.3
5.1	11.4
5.3	12.4
2.5	9.0
3.4	8.9
3.1	14.2
1.7	10.9
3.7	13.3

$r \approx -0.6273$

Strong negative linear correlation



Example 1

(c)

x	y
6.9	3.4
7.7	4.5
0.9	9.8
3.4	1.5
8.9	3.3
5.7	8.9
3.1	8.4
2.2	8.1
4.5	6.8
4.1	0.5
5.0	0.4
7.8	8.4
2.5	3.1
6.1	9.0
1.1	8.5

Example 1

(c)

x	y
6.9	3.4
7.7	4.5
0.9	9.8
3.4	1.5
8.9	3.3
5.7	8.9
3.1	8.4
2.2	8.1
4.5	6.8
4.1	0.5
5.0	0.4
7.8	8.4
2.5	3.1
6.1	9.0
1.1	8.5

$r \approx -0.2218$

Example 1

(c)

x	y
6.9	3.4
7.7	4.5
0.9	9.8
3.4	1.5
8.9	3.3
5.7	8.9
3.1	8.4
2.2	8.1
4.5	6.8
4.1	0.5
5.0	0.4
7.8	8.4
2.5	3.1
6.1	9.0
1.1	8.5

$r \approx -0.2218$

Weak negative correlation

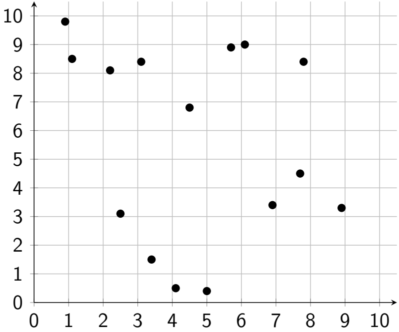
Example 1

(c)

x	y
6.9	3.4
7.7	4.5
0.9	9.8
3.4	1.5
8.9	3.3
5.7	8.9
3.1	8.4
2.2	8.1
4.5	6.8
4.1	0.5
5.0	0.4
7.8	8.4
2.5	3.1
6.1	9.0
1.1	8.5

$r \approx -0.2218$

Weak negative correlation



Objectives

1. Determine and interpret the linear correlation coefficient
2. Determine the linear regression equation

Linear Regression Equation

While determining the linear correlation coefficient is valuable, it is also helpful to be able to predict data values not contained in the data set.

Linear Regression Equation

While determining the linear correlation coefficient is valuable, it is also helpful to be able to predict data values not contained in the data set.

To do this, we can create the **least squares regression equation**, (also called the *line of best fit*) which will **minimize** the total squared distance each data point is from the line:

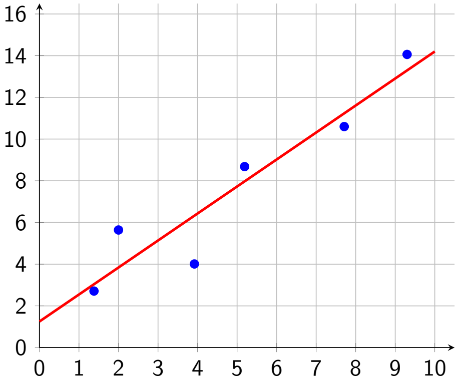
Linear Regression Equation

While determining the linear correlation coefficient is valuable, it is also helpful to be able to predict data values not contained in the data set.

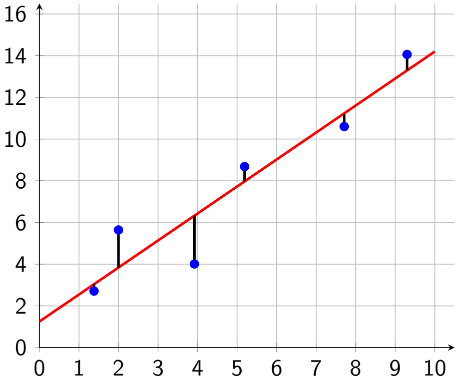
To do this, we can create the **least squares regression equation**, (also called the *line of best fit*) which will **minimize** the total squared distance each data point is from the line:

$$\hat{y} = mx + b$$

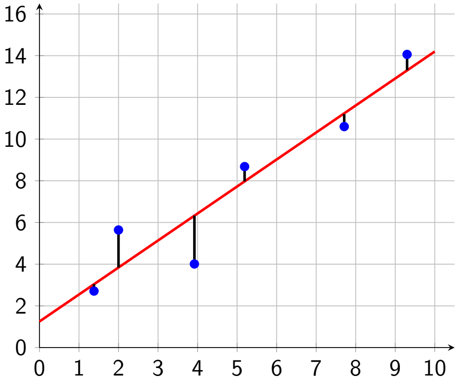
Line of Best Fit



Line of Best Fit

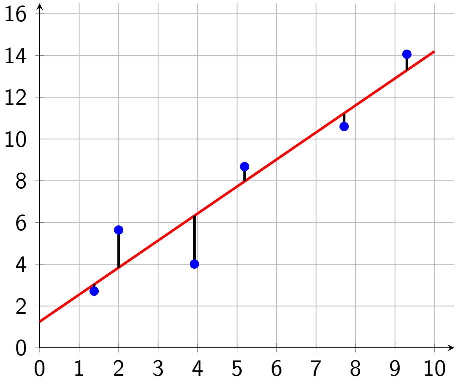


Line of Best Fit



The black lines are **residuals**

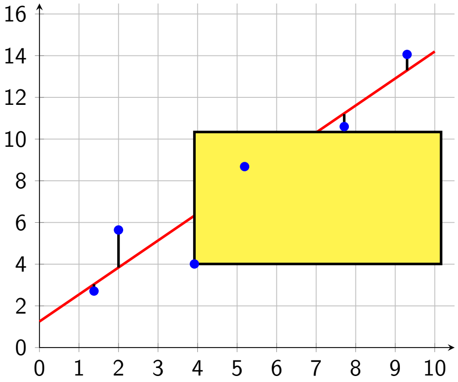
Line of Best Fit



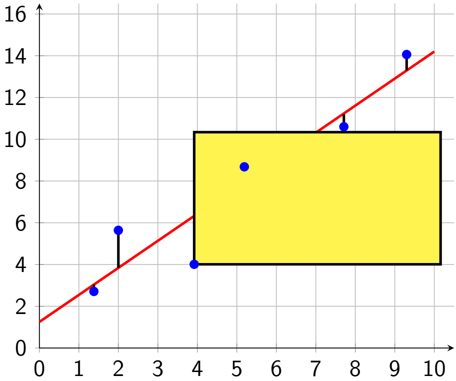
The black lines are **residuals**

Like deviations from the mean, the sum of the residuals is 0

Line of Best Fit



Line of Best Fit



The line of best fit minimizes the sum of the areas of the squares