

Correlation and Regression

Correlation

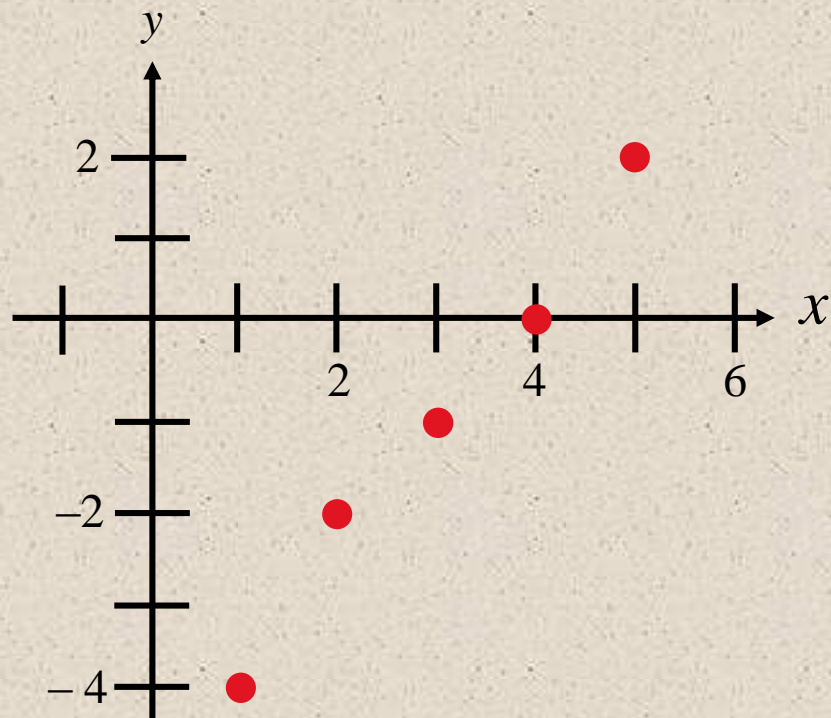
Correlation

A **correlation** is a relationship between two variables. The data can be represented by the ordered pairs (x, y) where x is the **independent** (or **explanatory**) **variable**, and y is the **dependent** (or **response**) **variable**.

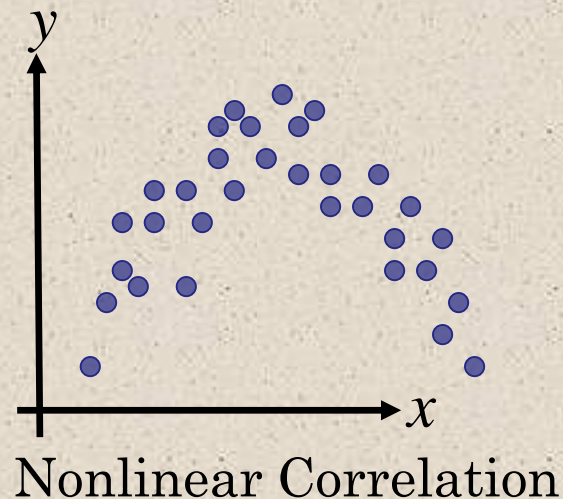
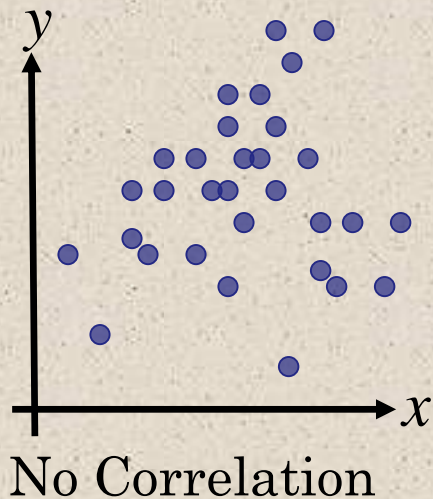
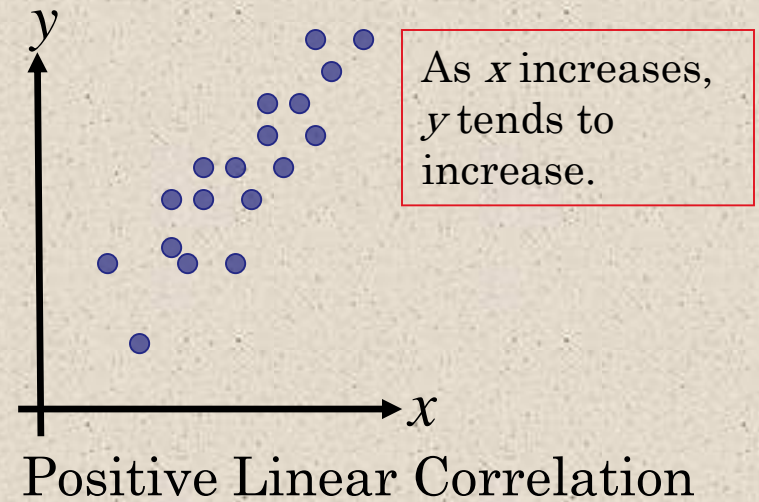
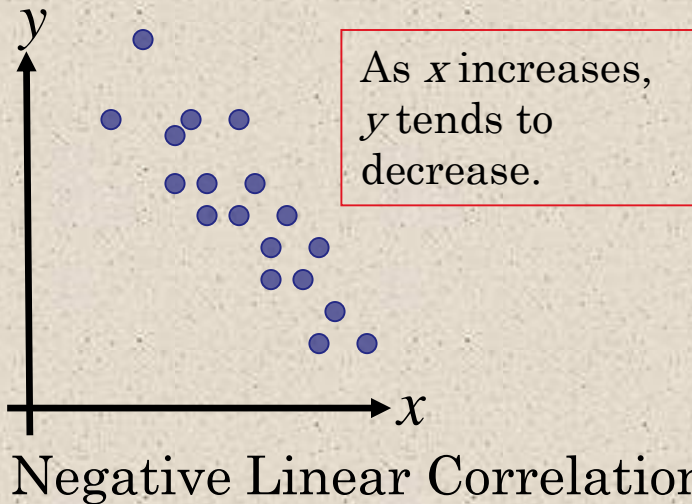
A **scatter plot** can be used to determine whether a linear (straight line) correlation exists between two variables.

Example:

x	1	2	3	4	5
y	-4	-2	-1	0	2



Linear Correlation



Correlation Coefficient

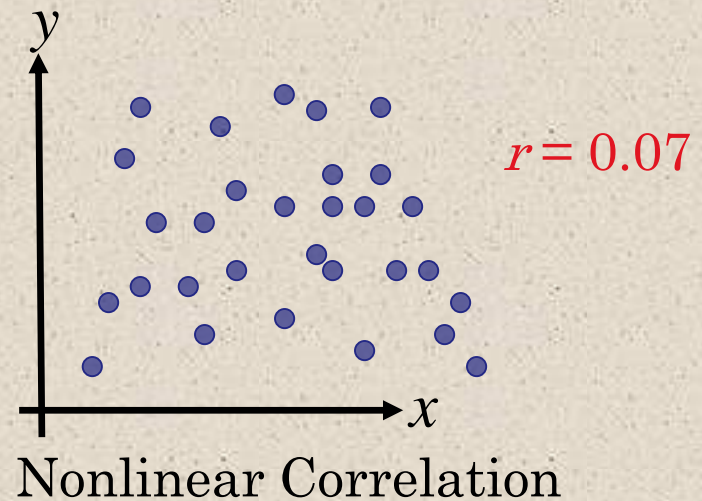
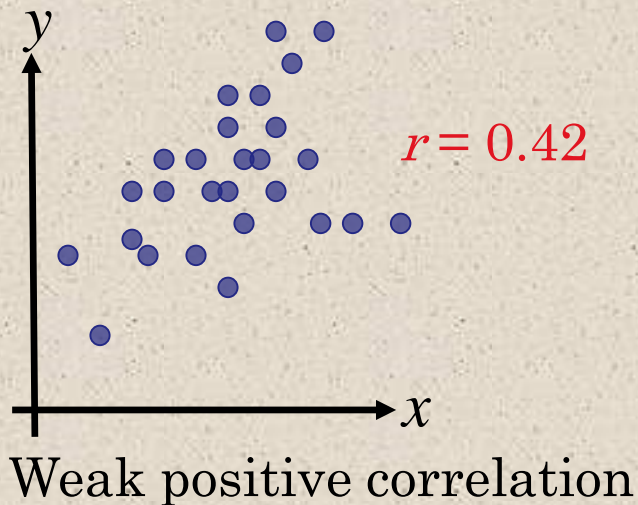
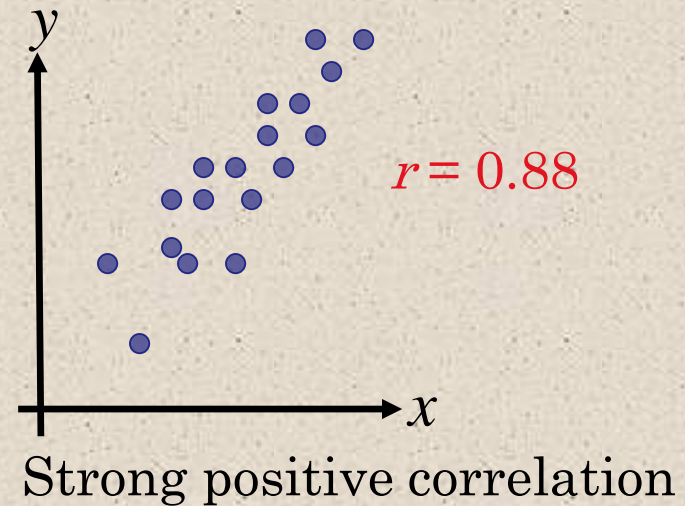
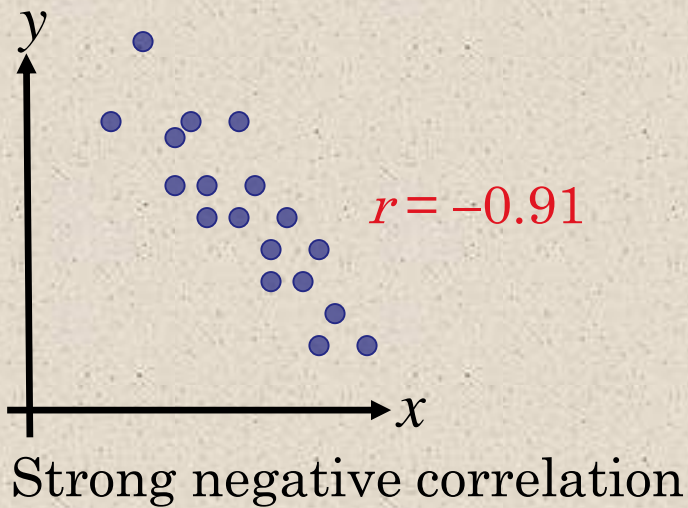
The **correlation coefficient** is a measure of the strength and the direction of a linear relationship between two variables. The symbol r represents the sample correlation coefficient. The formula for r is

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

Looks
sp ewa

The range of the correlation coefficient is -1 to 1 . If x and y have a strong positive linear correlation, r is close to 1 . If x and y have a strong negative linear correlation, r is close to -1 . If there is no linear correlation or a weak linear correlation, r is close to 0 .

Linear Correlation



Calculating a Correlation Coefficient

Calculating a Correlation Coefficient

In Words

1. Find the sum of the x -values.
2. Find the sum of the y -values.
3. Multiply each x -value by its corresponding y -value and find the sum.
4. Square each x -value and find the sum.
5. Square each y -value and find the sum.
6. Use these five sums to calculate the correlation coefficient.

In Symbols

$$\sum x$$

$$\sum y$$

$$\sum xy$$

$$\sum x^2$$

$$\sum y^2$$

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

Continued.

Correlation Coefficient

Example:

Calculate the correlation coefficient r for the following data.

x	y	xy	x^2	y^2
1	-3	-3	1	9
2	-1	-2	4	1
3	0	0	9	0
4	1	4	16	1
5	2	10	25	4
$\Sigma x = 15$	$\Sigma y = -1$	$\Sigma xy = 9$	$\Sigma x^2 = 55$	$\Sigma y^2 = 15$

$$r = \frac{n \Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n \Sigma x^2 - (\Sigma x)^2} \sqrt{n \Sigma y^2 - (\Sigma y)^2}} = \frac{5(9) - (15)(-1)}{\sqrt{5(55) - 15^2} \sqrt{5(15) - (-1)^2}}$$
$$= \frac{60}{\sqrt{50} \sqrt{74}} \approx 0.986$$

There is a strong positive linear correlation between x and y .

Correlation Coefficient

Example:

The following data represents the number of hours 12 different students watched television during the weekend and the scores of each student who took a test the following Monday.

- Display the scatter plot.
- Calculate the correlation coefficient r .

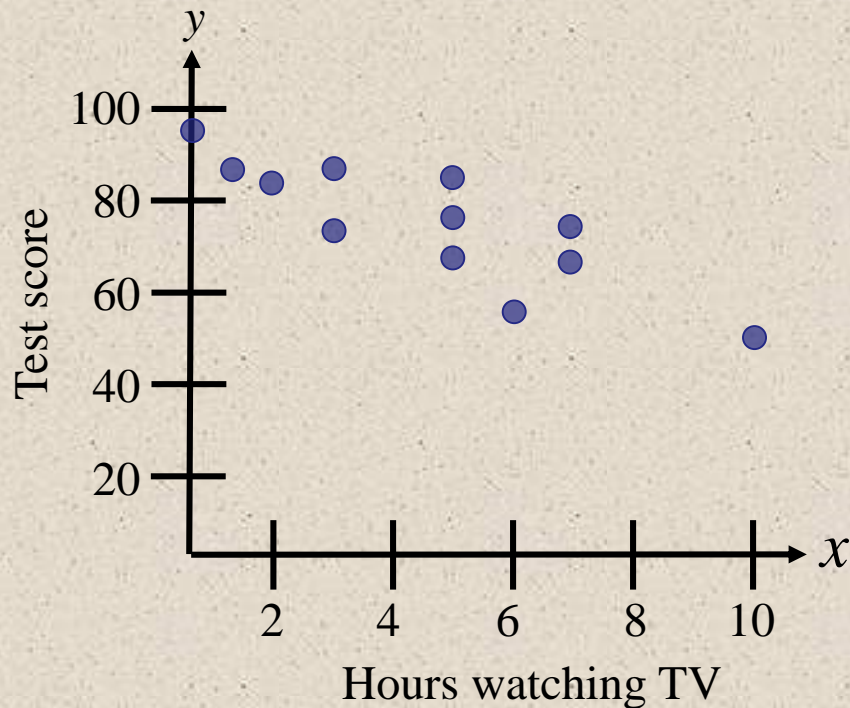
Hours, x	0	1	2	3	3	5	5	5	6	7	7	10
Test score, y	96	85	82	74	95	68	76	84	58	65	75	50

Continued.

Correlation Coefficient

Example continued:

Hours, x	0	1	2	3	3	5	5	5	6	7	7	10
Test score, y	96	85	82	74	95	68	76	84	58	65	75	50



Continued.

Correlation Coefficient

Example continued:

Hours, x	0	1	2	3	3	5	5	5	6	7	7	10
Test score, y	96	85	82	74	95	68	76	84	58	65	75	50
xy	0	85	164	222	285	340	380	420	348	455	525	500
x^2	0	1	4	9	9	25	25	25	36	49	49	100
y^2	9216	7225	6724	5476	9025	4624	5776	7056	3364	4225	5625	2500

$$\sum x = 54 \quad \sum y = 908 \quad \sum xy = 3724 \quad \sum x^2 = 332 \quad \sum y^2 = 70836$$

$$r = \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{n \sum x^2 - (\sum x)^2} \sqrt{n \sum y^2 - (\sum y)^2}} = \frac{12(3724) - (54)(908)}{\sqrt{12(332) - 54^2} \sqrt{12(70836) - (908)^2}} \approx -0.831$$

There is a strong negative linear correlation.

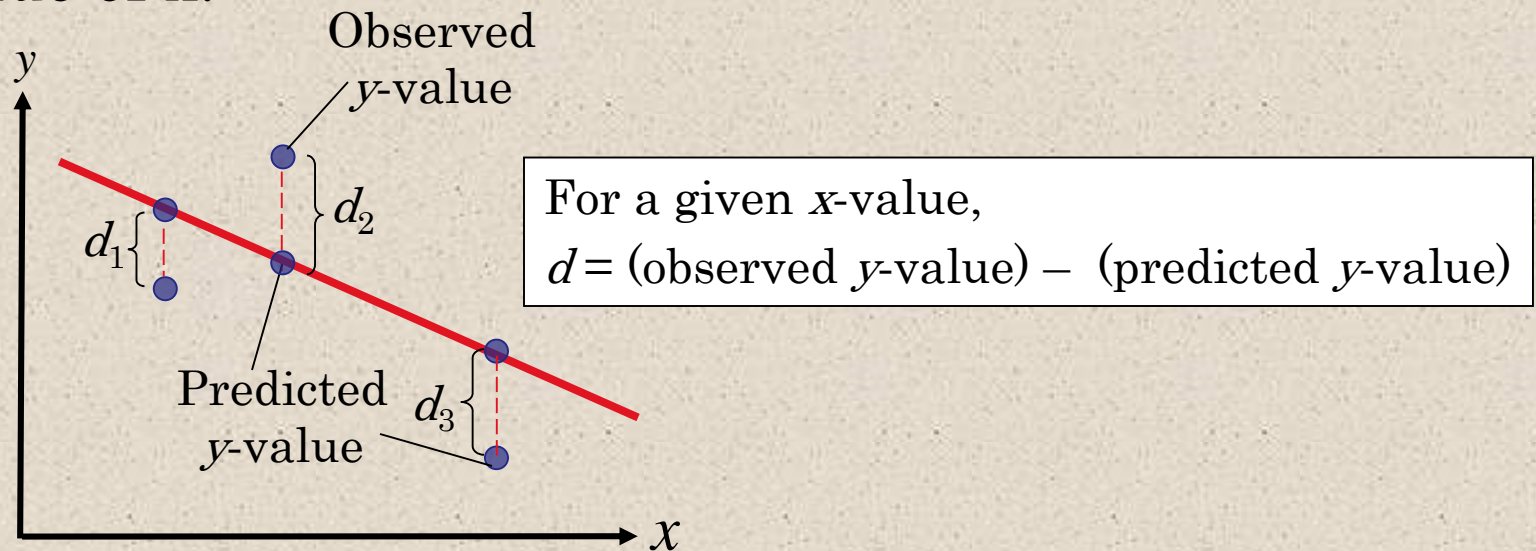
As the number of hours spent watching TV increases, the test scores tend to decrease.

§ 9.2

Linear Regression

Residuals

After verifying that the linear correlation between two variables is significant, next we determine the equation of the line that can be used to predict the value of y for a given value of x .



Each data point d_i represents the difference between the observed y -value and the predicted y -value for a given x -value on the line. These differences are called **residuals**.

Regression Line

A **regression line**, also called a **line of best fit**, is the line for which the sum of the squares of the residuals is a minimum.

The Equation of a Regression Line

The equation of a regression line for an independent variable x and a dependent variable y is

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

where \hat{y} is the predicted y -value for a given x -value. The slope m and y -intercept b are given by

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

where \bar{y} is the mean of the y -values and \bar{x} is the mean of the x -values. The regression line always passes through (\bar{x}, \bar{y}) .

Regression Line

Example:

Find the equation of the regression line.

x	y	xy	x^2	y^2
1	-3	-3	1	9
2	-1	-2	4	1
3	0	0	9	0
4	1	4	16	1
5	2	10	25	4
$\Sigma x = 15$	$\Sigma y = -1$	$\Sigma xy = 9$	$\Sigma x^2 = 55$	$\Sigma y^2 = 15$

$$m = \frac{n \Sigma xy - (\Sigma x)(\Sigma y)}{n \Sigma x^2 - (\Sigma x)^2} = \frac{5(9) - (15)(-1)}{5(55) - (15)^2} = \frac{60}{50} = 1.2$$

Continued.

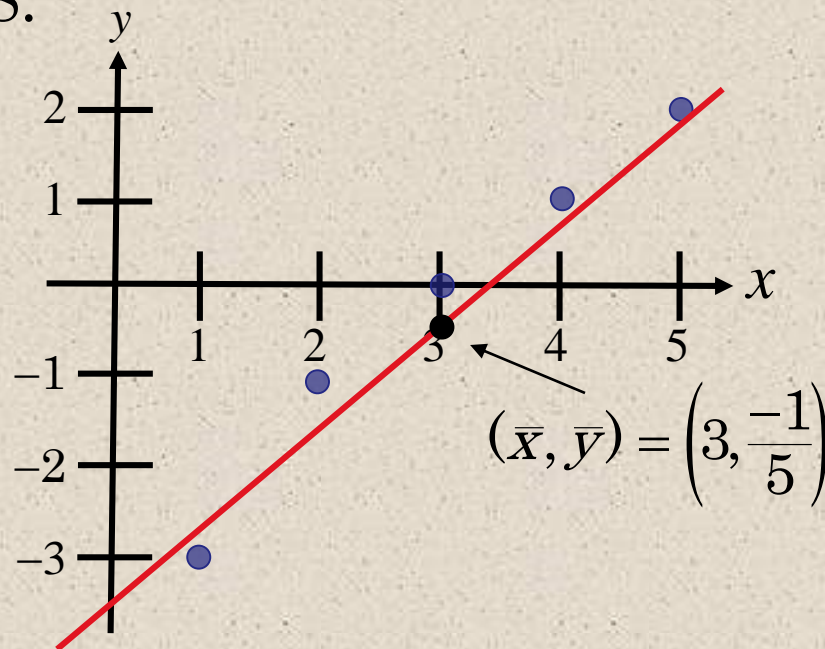
Regression Line

Example continued:

$$b = \bar{y} - m\bar{x} = \frac{-1}{5} - (1.2)\frac{15}{5} = -3.8$$

The equation of the regression line is

$$\hat{y} = 1.2x - 3.8.$$



Regression Line

Example:

The following data represents the number of hours 12 different students watched television during the weekend and the scores of each student who took a test the following Monday.

- Find the equation of the regression line.
- Use the equation to find the expected test score for a student who watches 9 hours of TV.

Hours, x	0	1	2	3	3	5	5	5	6	7	7	10
Test score, y	96	85	82	74	95	68	76	84	58	65	75	50
xy	0	85	164	222	285	340	380	420	348	455	525	500
x^2	0	1	4	9	9	25	25	25	36	49	49	100
y^2	9216	7225	6724	5476	9025	4624	5776	7056	3364	4225	5625	2500

$$\Sigma x = 54$$

$$\Sigma y = 908$$

$$\Sigma xy = 3724$$

$$\Sigma x^2 = 332$$

$$\Sigma y^2 = 70836$$

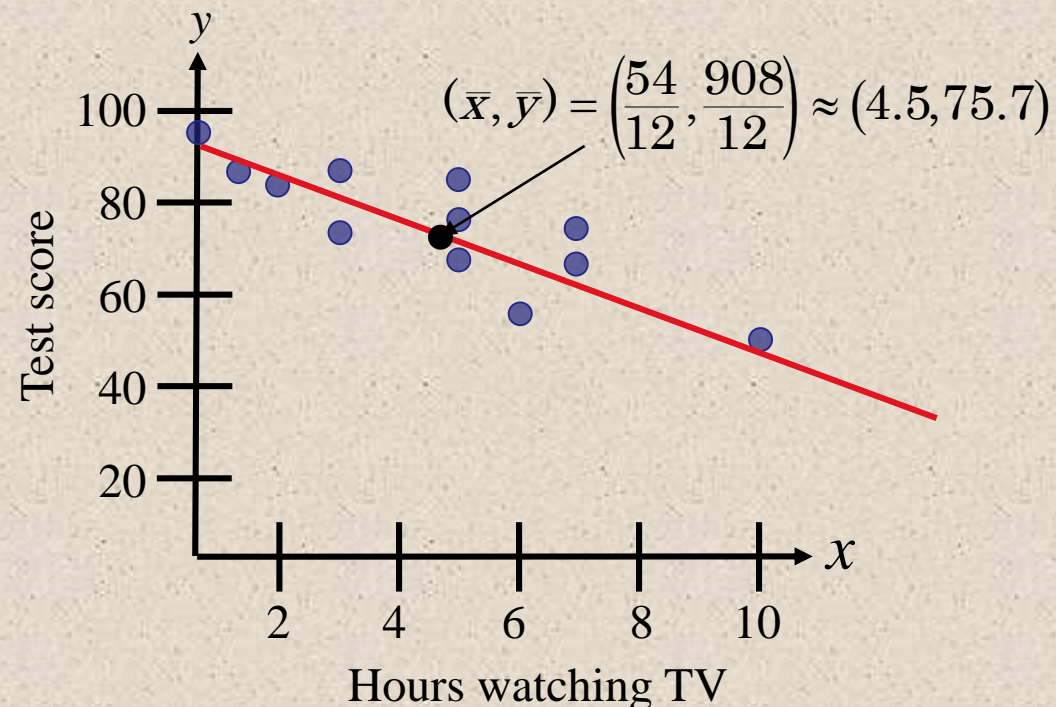
Regression Line

Example continued:

$$m = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2} = \frac{12(3724) - (54)(908)}{12(332) - (54)^2} \approx -4.067$$

$$\begin{aligned} b &= \bar{y} - m\bar{x} \\ &= \frac{908}{12} - (-4.067)\frac{54}{12} \\ &\approx 93.97 \end{aligned}$$

$$\hat{y} = -4.07x + 93.97$$



Continued.

Regression Line

Example continued:

Using the equation $\hat{y} = -4.07x + 93.97$, we can predict the test score for a student who watches 9 hours of TV.

$$\begin{aligned}\hat{y} &= -4.07x + 93.97 \\ &= -4.07(9) + 93.97 \\ &= 57.34\end{aligned}$$

A student who watches 9 hours of TV over the weekend can expect to receive about a 57.34 on Monday's test.