Correlation and Regression

fifth lecture

We will learn in this lecture:

1- Linear Correlation Coefficient of Pearson

2- Simple Linear Regression

Correlation and Regression

Definition of 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**.

Example:

A. The relation exits between the number of hours for group of students spent studying for a test and their scores on that test.

B. The relation exits between the high outdoor temperature (in degrees Fahrenheit) and coffee sales (in hundreds of dollars) for a coffee shop for eight randomly selected days.

C. The relation exists between an individual's weight (in pounds) and daily water consumption (in ounces).

D. The relation exists between income per year (in thousand of dollars) and a mount spent on milk per year (in dollars).

Example:

x = hours spent studying , y= scores on that test

x = temperature (in degrees Fahrenheit) , y= coffee sales

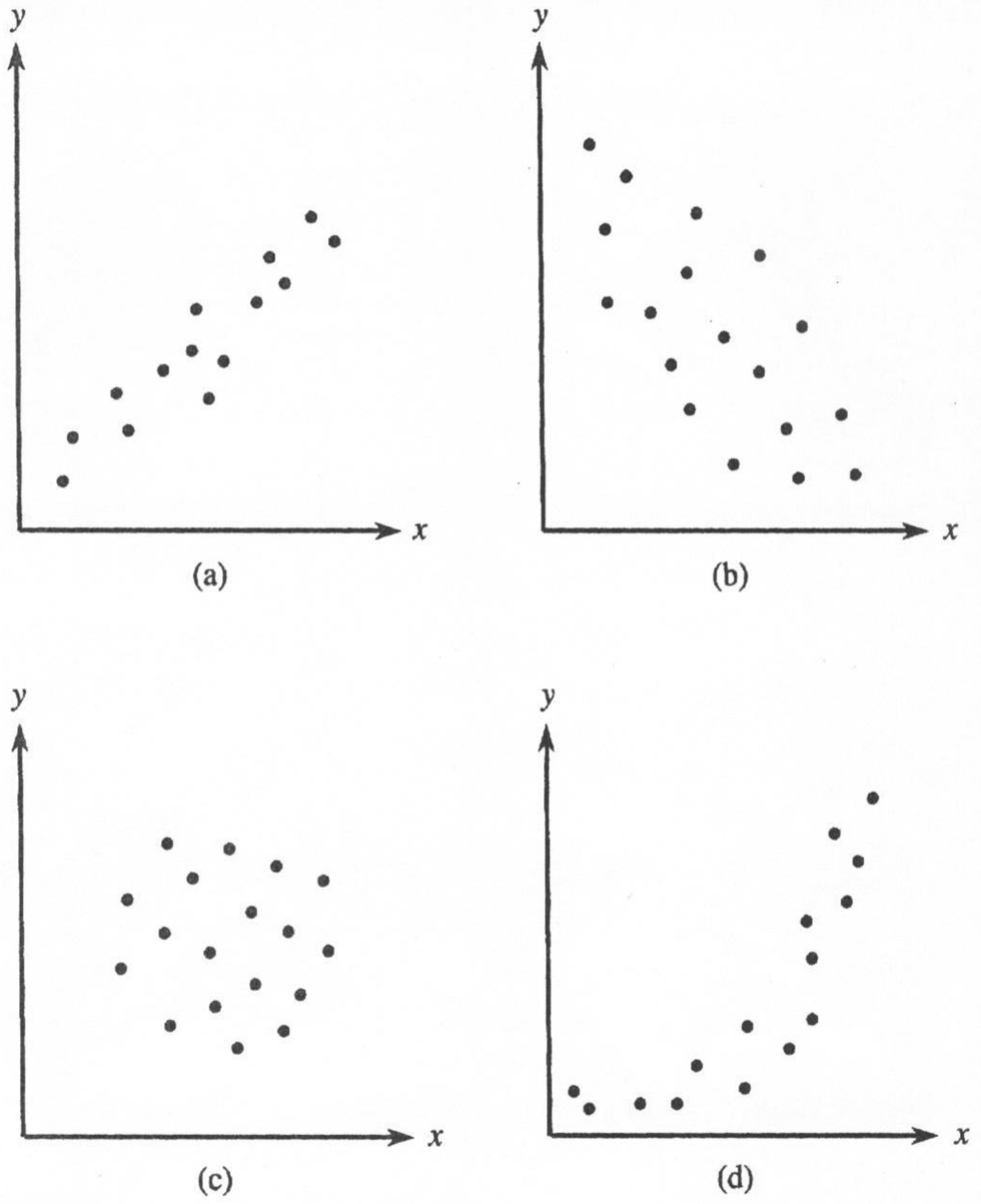
x = an individual’s weight (in pounds) , y= water consumption

x = money spent on advertising , y= company sales

Scatter plot

x = hours spent studying
y= scores on that test

x=temperat-ure in degrees Fahrenheit
y= coffee sales



x= an individual’s weight (in pounds)
y= water consumption

x= Income per year
y=a mount spent on milk

Linear Correlation Coefficient of Pearson

Definition of Correlation :

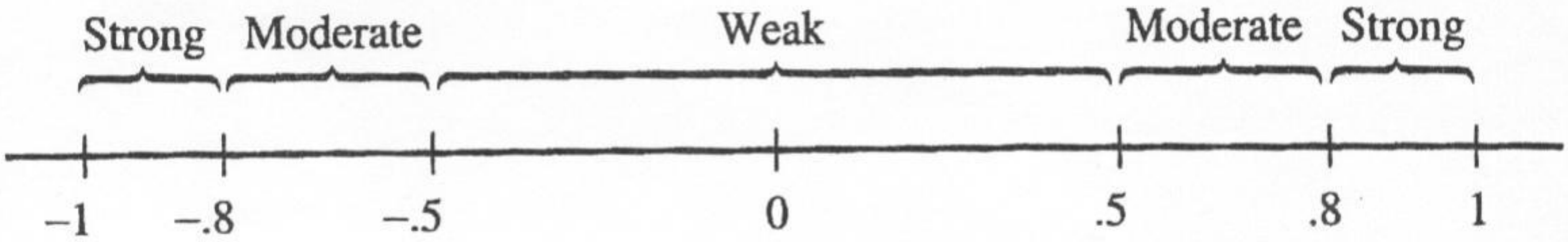
The **correlation coefficient** is a measure of the strength and the direction of a liner relationship between two variables. The symbol *r* represents the sample correlation coefficient.

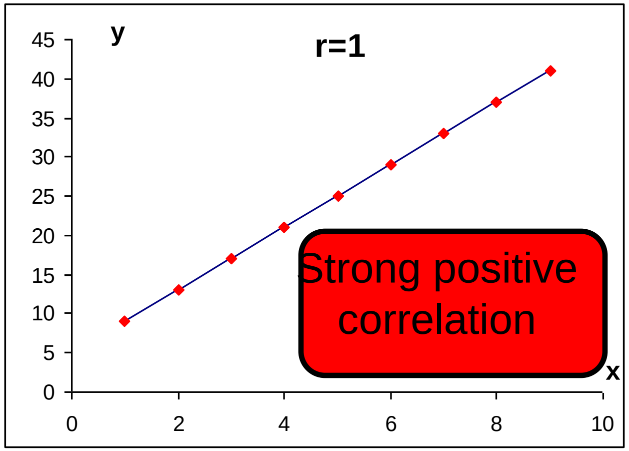
$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}}$$

Where *n* is the number of pairs of data.

Remark

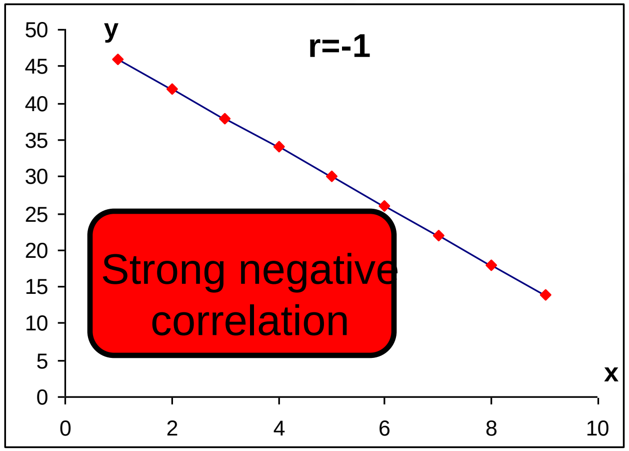
The range of correlation coefficient is -1 to 1.



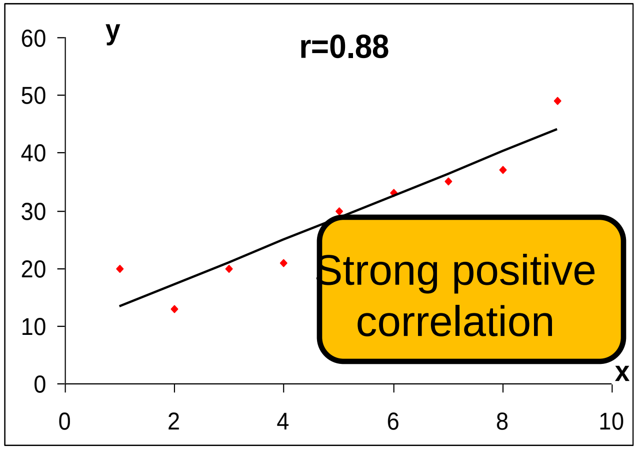


Strong positive correlation

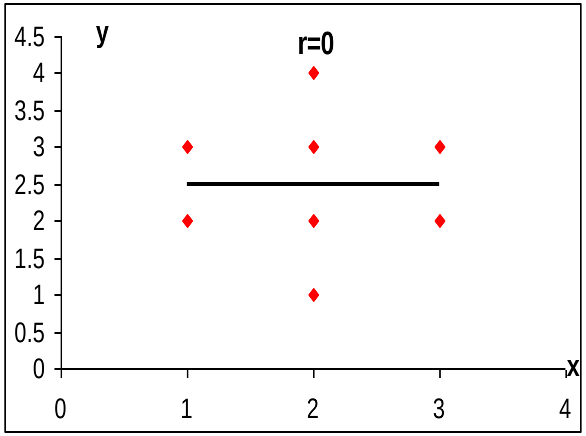
No correlation

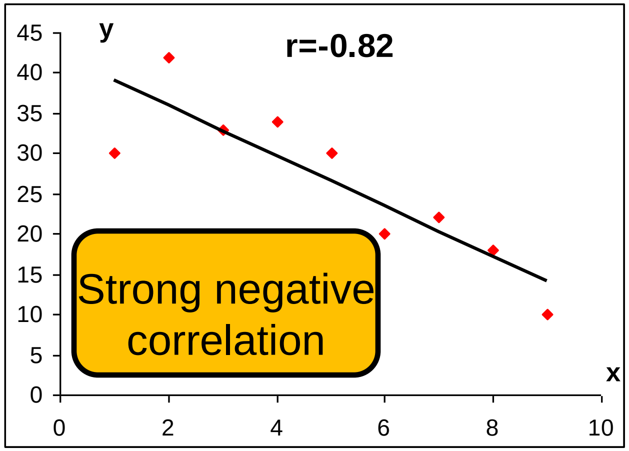


Strong negative correlation

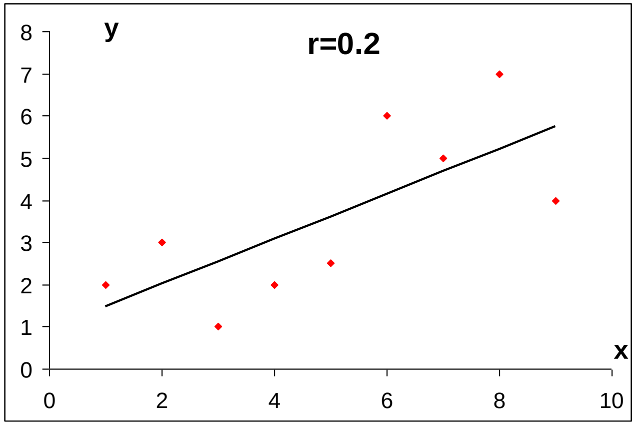


Strong positive correlation



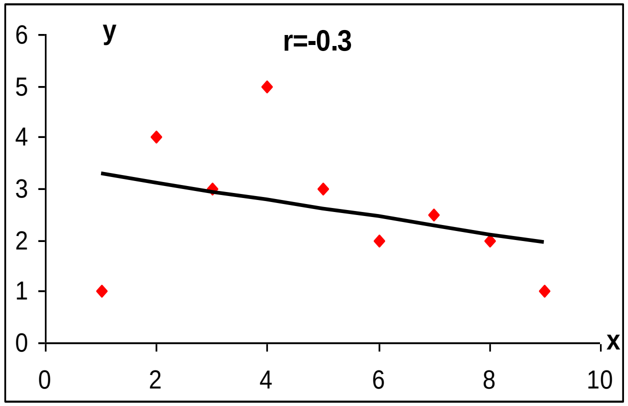


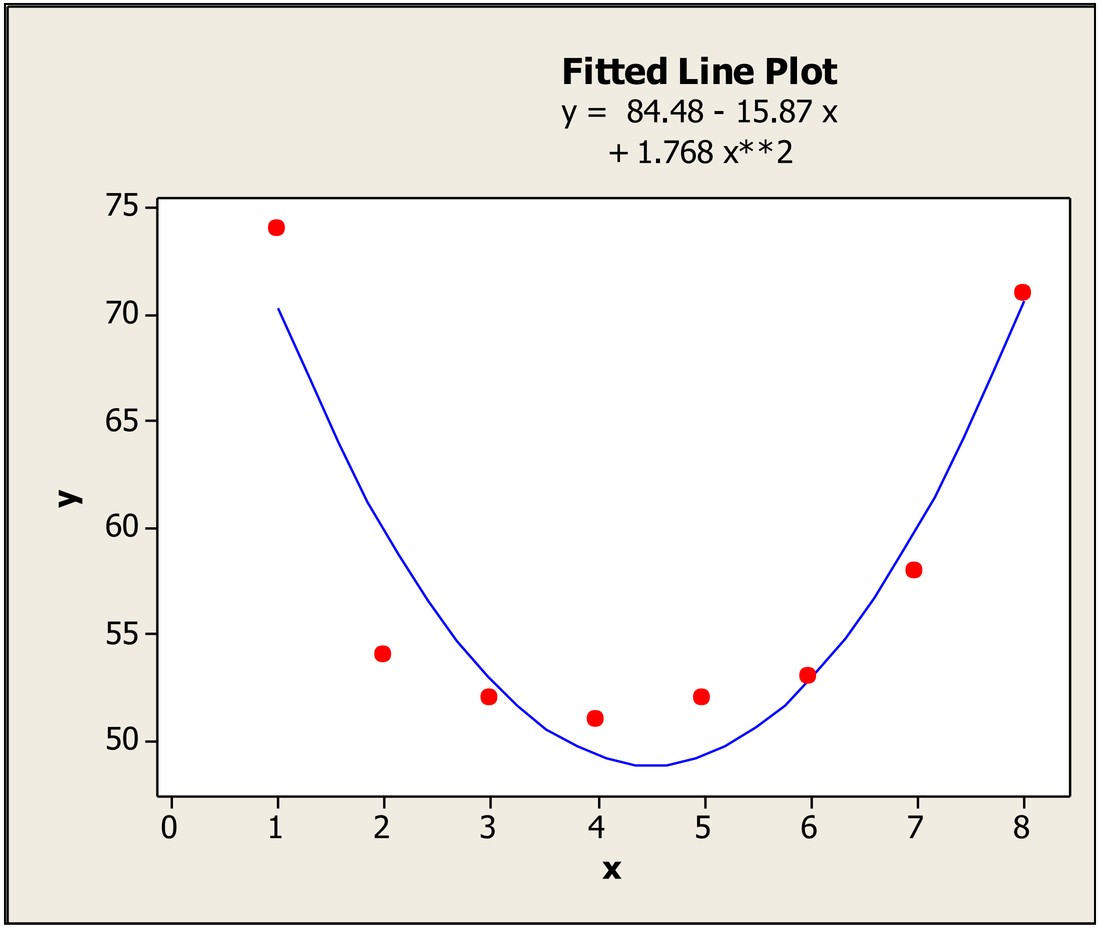
Strong negative correlation



Weak positive correlation

Weak negative correlation





Note

Weak linear correlation coefficient does not mean no any relationship

Example:

A marketing manager conducted a study to determine whether there is a linear relationship between money spent on advertising and company sales. The data are shown in the table below.

A. Calculate the correlation coefficient for the advertising expenditures and company sales data.

B. Display the data in a scatter plot then determine the types of correlation .

C. What can you conclude

Advertising expenses	2.4	1.6	2	2.6	1.4	1.6	2	2.2
Company sales	225	184	220	240	180	184	186	215

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}}$$

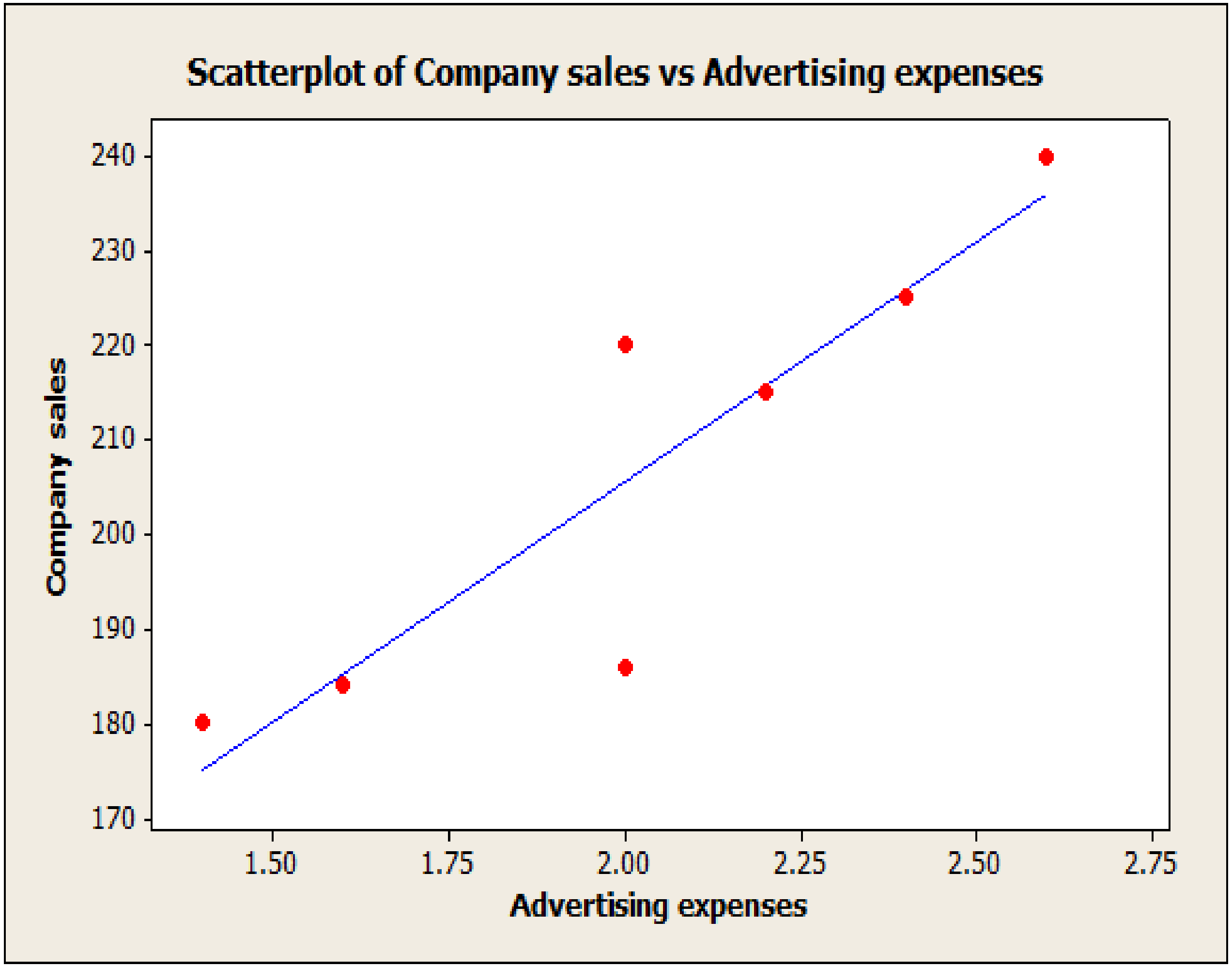
XY	Y^2	X^2	Y	X	
540	50.625	5.76	225	2.4	
294.4	33.856	2.56	184	1.6	
440	48.400	4	220	2.0	
			240	2.6	
			180	1.4	
			184	1.6	
			186	2.0	
			215	2.2	
					Total

XY	Y^2	X^2	Y	X	
540	50.625	5.76	225	2.4	
294.4	33.856	2.56	184	1.6	
440	48.400	4	220	2.0	
624	57.600	6.76	240	2.6	
252	32.400	1.96	180	1.4	
294.4	33.856	2.56	184	1.6	
372	34.596	4	186	2.0	
473	46.225	4.84	215	2.2	
3289.8	337.558	32.44	1634	15.8	**Total**

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}}$$

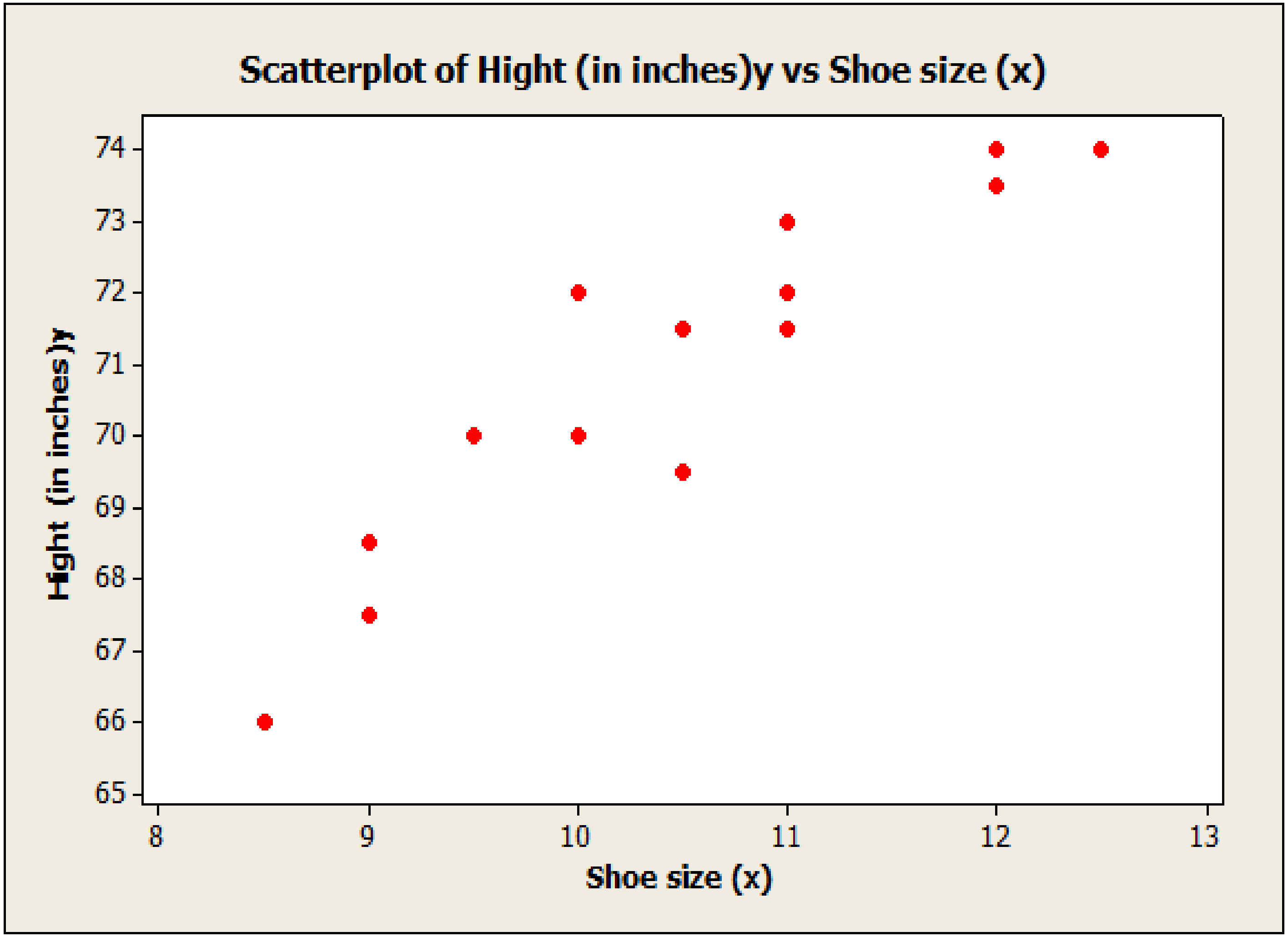
$$\boldsymbol{r} = \frac{8(3289.8) - 15.8(1634)}{\sqrt{[8(32.44) - (15.8)^2][8(337.558) - (1634)^2]}}$$

=0.913



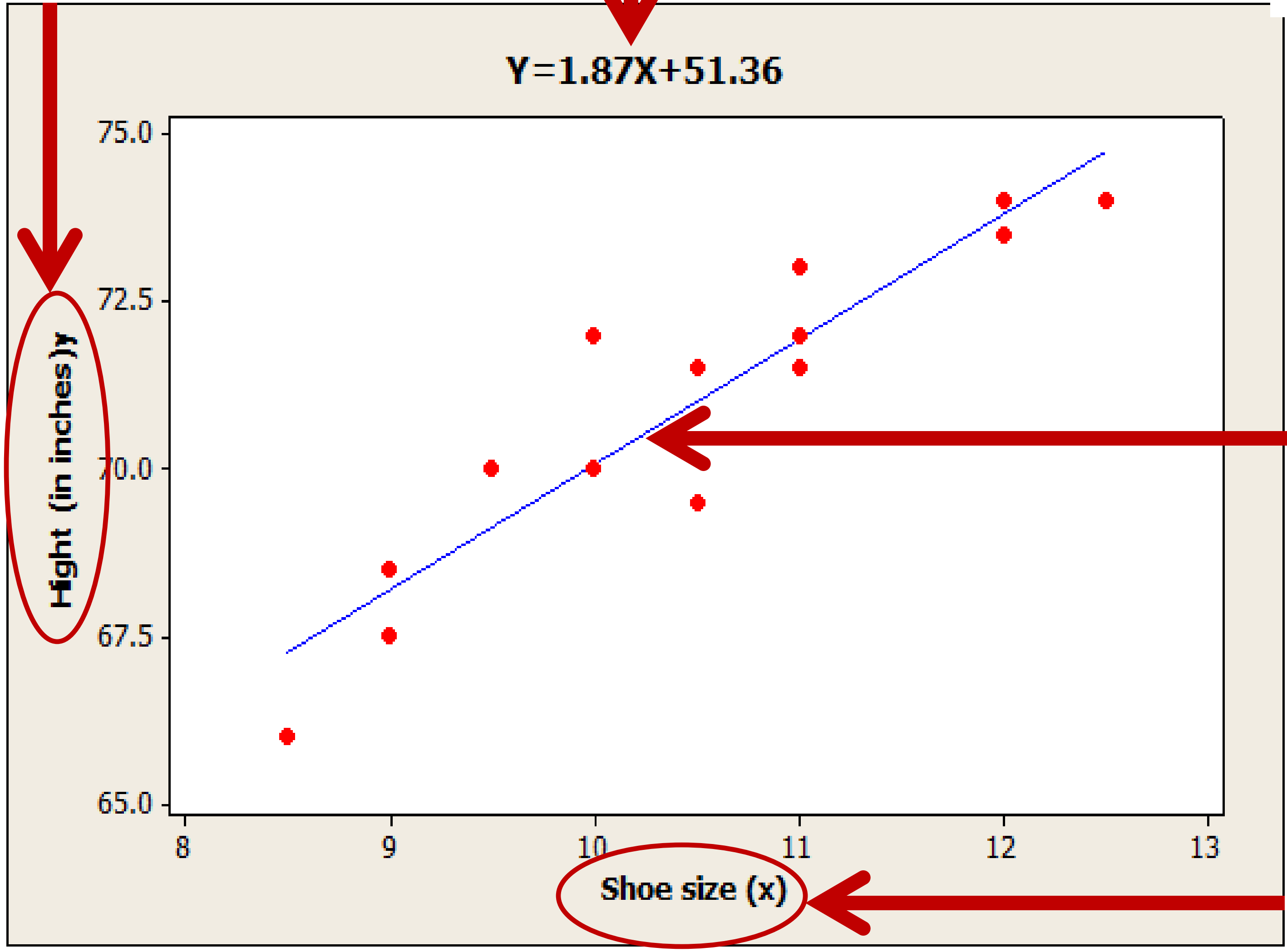
Strong positive correlation

Simple Linear Regression


Scatterplot of Hight (in inches)y vs Shoe size (x)
Hight (in inches)y
74
73
72
71
70
69
68
67
66
65
8
9
10
11
12
13
Shoe size (x)

The equation of a regression line

dependent variable(response)



Regression line

Independent variable

The equation of a regression line for an independent variable X and a dependent variable Y is:

$$Y = mX + b$$

where Y is the predicted Y-value for a given X-value. The slope m and Y-intercept b are given by:

$$m = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2}$$ and $$b = \bar{Y} - m\bar{X}$$

where $\bar{Y}$ is the mean of the Y-value in the data set and $\bar{X}$ is the mean of the X-value.

Example:

A marketing manager conducted a study to determine whether there is a linear relationship between money spent on advertising and company sales. The data are shown in the table below.

Find the equation of the regression line for the advertising expenditures and company sales data

Advertising expenses	2.4	1.6	2	2.6	1.4	1.6	2	2.2
Company sales	225	184	220	240	180	184	186	215

$$m = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2}$$

$$b = \bar{Y} - m\bar{X}$$

XY	Y^2	X^2	Y	X	
540	50.625	5.76	225	2.4	
294.4	33.856	2.56	184	1.6	
440	48.400	4	220	2.0	
624	57.600	6.76	240	2.6	
252	32.400	1.96	180	1.4	
294.4	33.856	2.56	184	1.6	
372	34.596	4	186	2.0	
473	46.225	4.84	215	2.2	
3289.8	337.558	32.44	1634	15.8	**Total**

XY	Y^2	X^2	Y	X	
540	50.625	5.76	225	2.4	
294.4	33.856	2.56	184	1.6	
440	48.400	4	220	2.0	
624	57.600	6.76	240	2.6	
252	32.400	1.96	180	1.4	
294.4	33.856	2.56	184	1.6	
372	34.596	4	186	2.0	
473	46.225	4.84	215	2.2	
3289.8	337.558	32.44	1634	15.8	**Total**

$$m = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2}$$

$$m = \frac{8(3289.9) - 15.8(1634)}{[8(32.44) - (15.8)^2]}$$

=50.7287

$$b = \bar{Y} - m\bar{X}$$

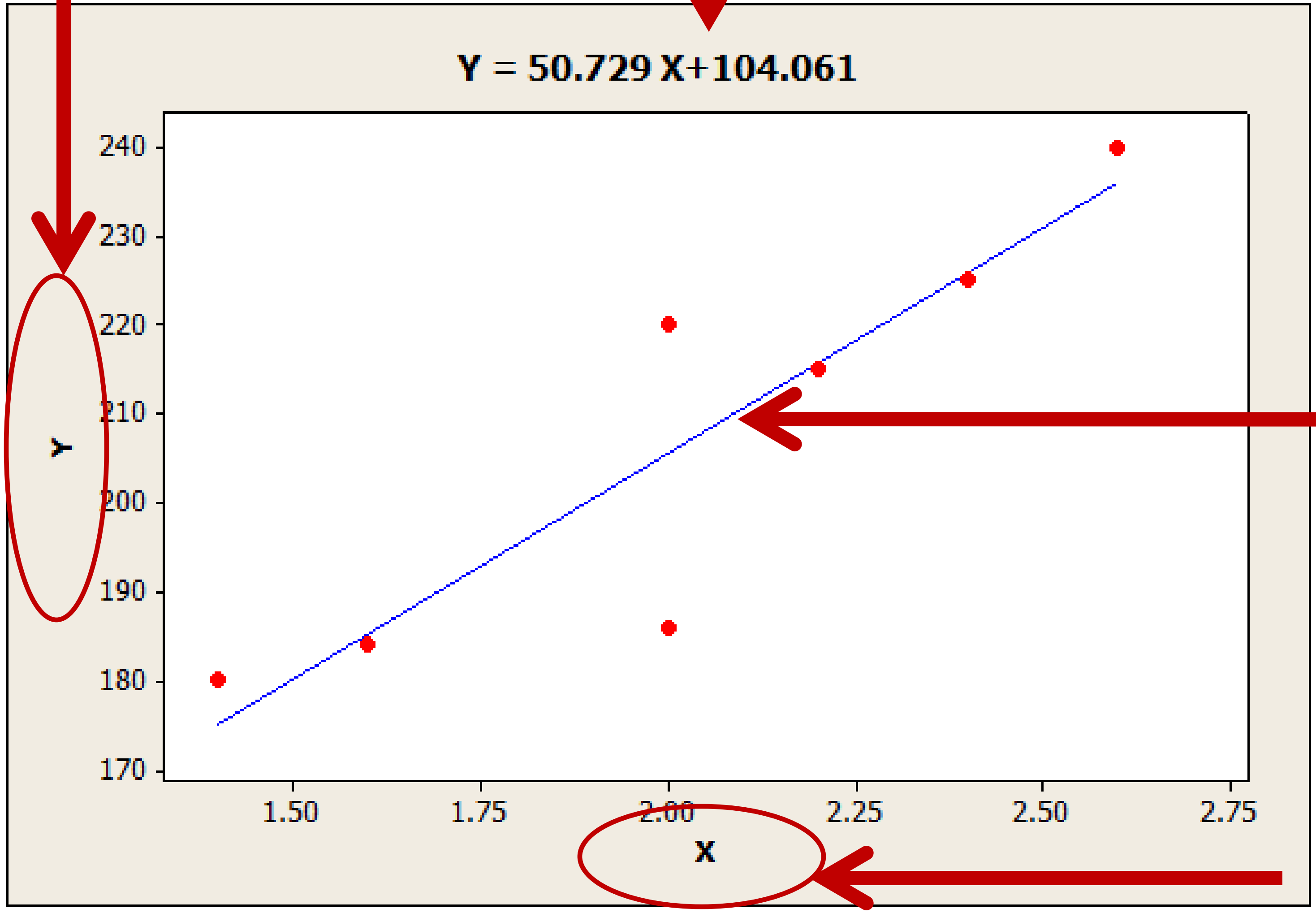
$$b = \frac{1634}{8} - 50.7287(\frac{15.8}{8})$$

=104.0608

Y=50.729X+104.061

The equation of a regression line

Company sales



Y = 50.729 X+104.061

Regression line

Advertising expenses

To calculate the Pearson correlation coefficient , there are several **hypotheses**

- The data must be in the form of pairs
- Data for each variable must be normal distributed.
- The sample must be random, **that means any individual value does not depend on the values of another individual.**
- The relationship between the two variables must be linear. Because this coefficient measures the strength of the linear relationship.

Notation:

- A strong linear relationship does not mean that a causal relationship between two variables.
- If the variables are not normal distributed, there are another correlation coefficients , such as Sperman, Kindal