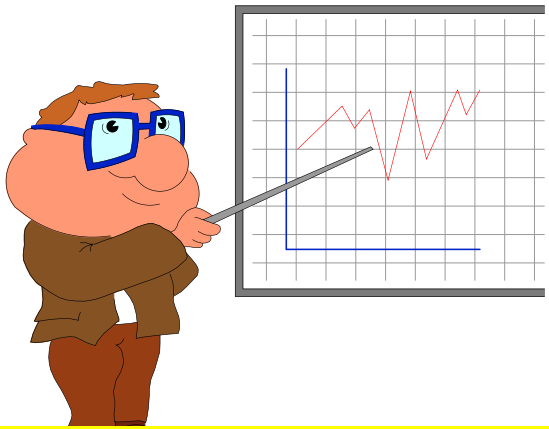
TAIBAH UNIVERSITY
Faculty of Science
Department of Math.

جامعة طيبة
كلية العلوم
قسم الرياضيات

STAT 301



Teacher :Dr

Lesson

8

Regression

Regression Analysis

The idea behind linear regression is to build a model that describes the dependence of one variable (the response or dependent variable) on another variable(s) (the explanatory or independent variable) .

Regression Analysis

In regression analysis we analyze the relationship between two or more variables.

The relationship between two or more variables could be linear or non linear.

If the relationship between only two variables called Simple regression.

Simple Linear Regression is a linear Regression Between Two Variables

Simple Linear Regression Model

A statistical technique that uses a straight-line relationship to predict a numerical dependent variable Y from a single numerical independent variable X.

The general form of a linear equation with one independent variable can be written as

$$y = b_0 + b_1 x$$

Where b_0 and b_1 are constants (fixed numbers), X is the independent variable, and Y is the dependent variable. The graph of a linear equation with one independent variable is a straight line.

Simple Linear Regression Model

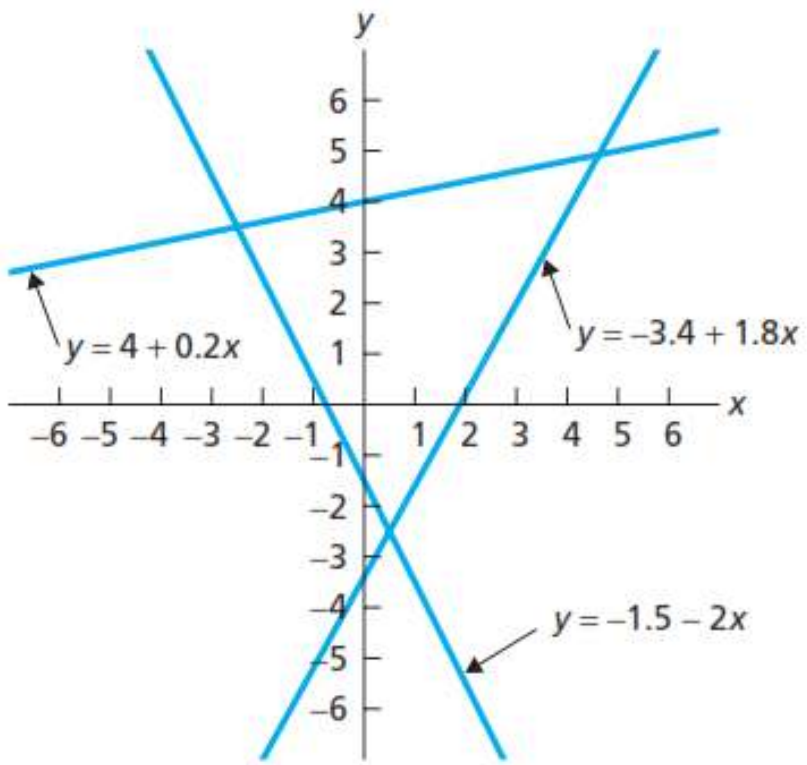
Examples of linear equations with one independent variable are

$y = 4 + 0.2x$

$y = -1.5 - 2x$, and

$y = -3.4 + 1.8x$.

The graphs of these three linear equations are shown in this Fig.



Graphs of three linear equations

Intercept and slope

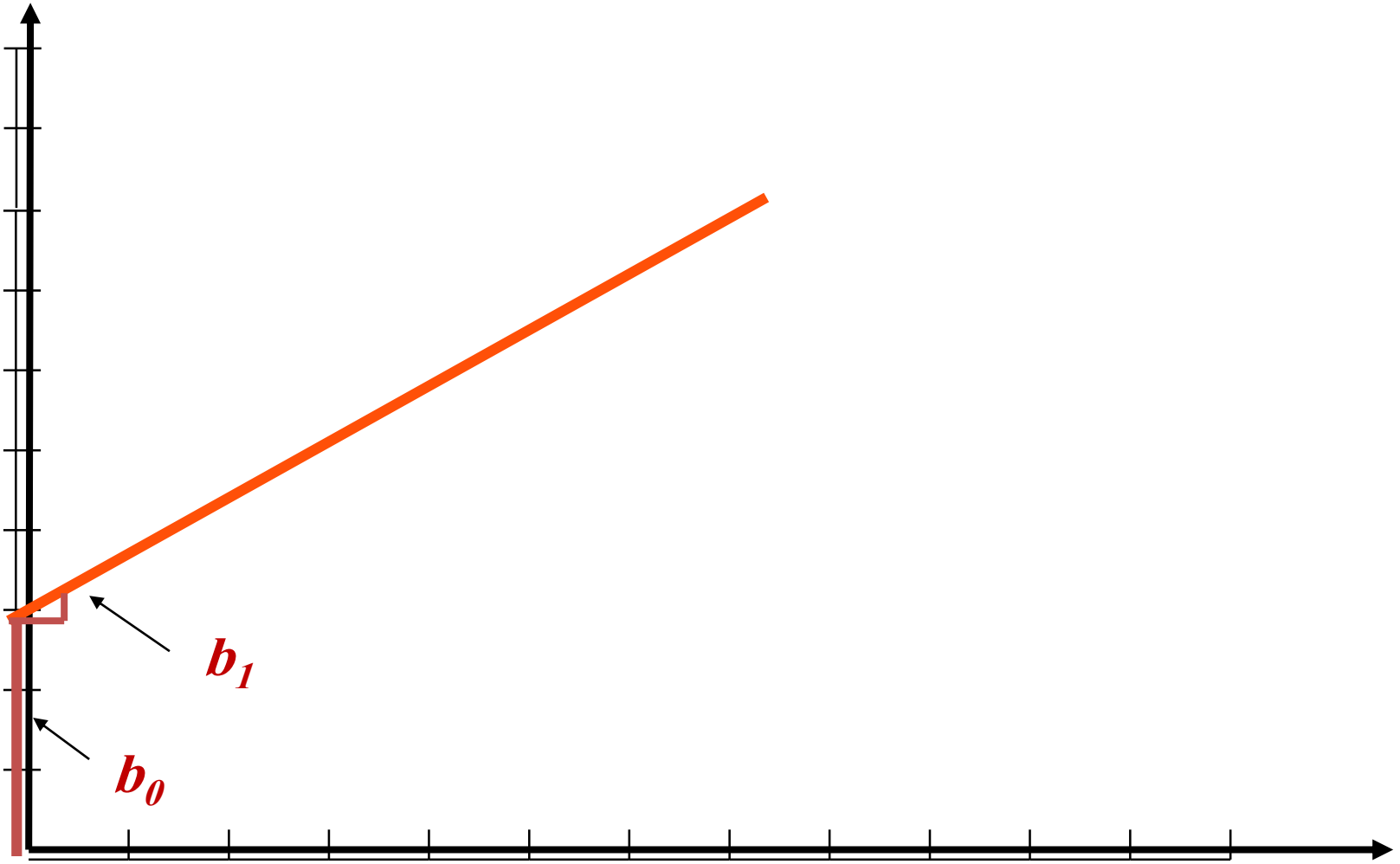
For a linear equation $y = b_0 + b_1x$ the number b_0 is called the ***y-intercept*** and the number b_1 is called the **slope.**

b_0 is the value of *Y when X = 0,*

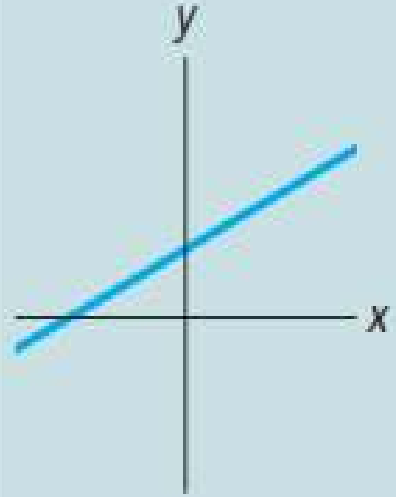
b_1 is the change in *Y per unit change in X .*
Positive slope means *Y increases as X increases.*
Negative slope means Y* decreases as *X increases.

The y intercept and the slope are known as the regression coefficients.

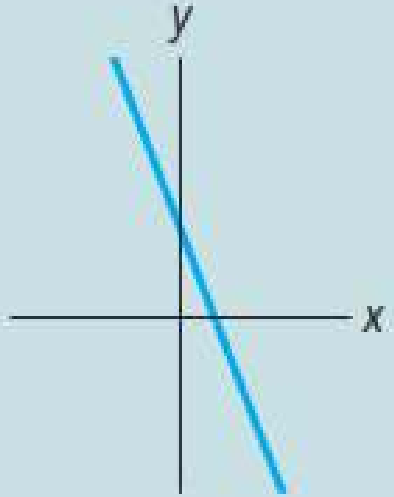
Intercept and slope



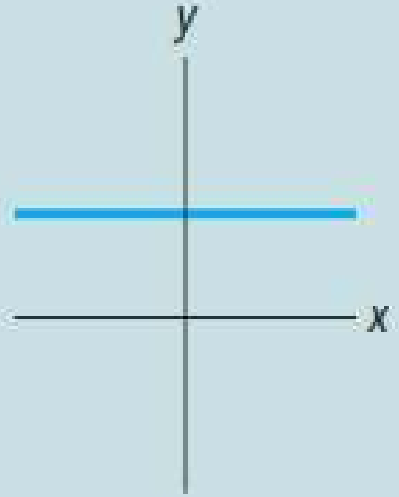
Graphical interpretation of slope



$b_1 > 0$



$b_1 < 0$



$b_1 = 0$

Least squares Method

For plotted sets of X and Y values, there are many possible straight lines, each with its own values of b_0 and b_1 , that might seem to fit the data.

The least-squares method finds the values for the y intercept and the slope that makes the sum of the squared differences between the actual values of the dependent variable Y and the predicted values of Y as small as possible.

Least squares Method

Consider the problem of fitting a line to the four data points, whose scatterplot is shown in Fig. below. Many lines can "fit" those four data points. Two possibilities are shown in Figs.(a) and (b) on the next slide .

x	y
1	1
1	2
2	2
4	6

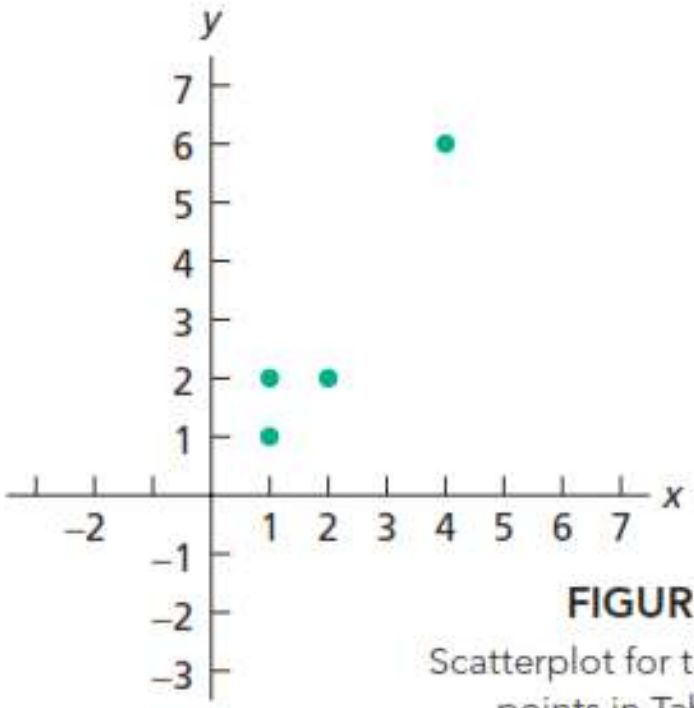
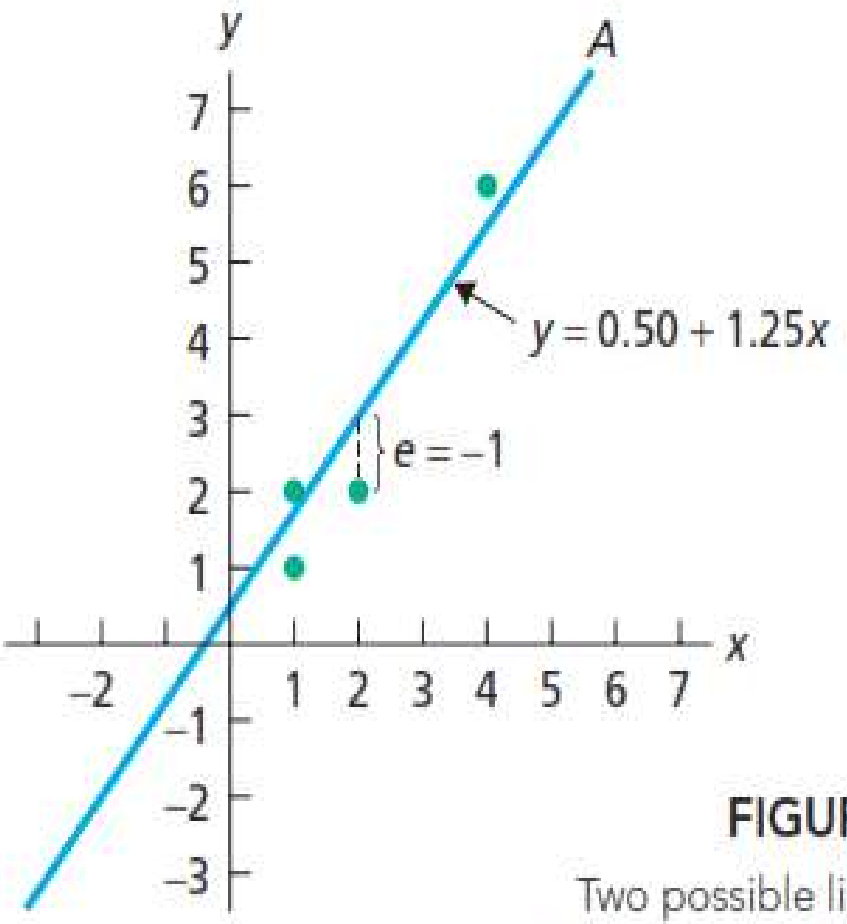


FIGURE
Scatterplot for the data points in Table 14.3

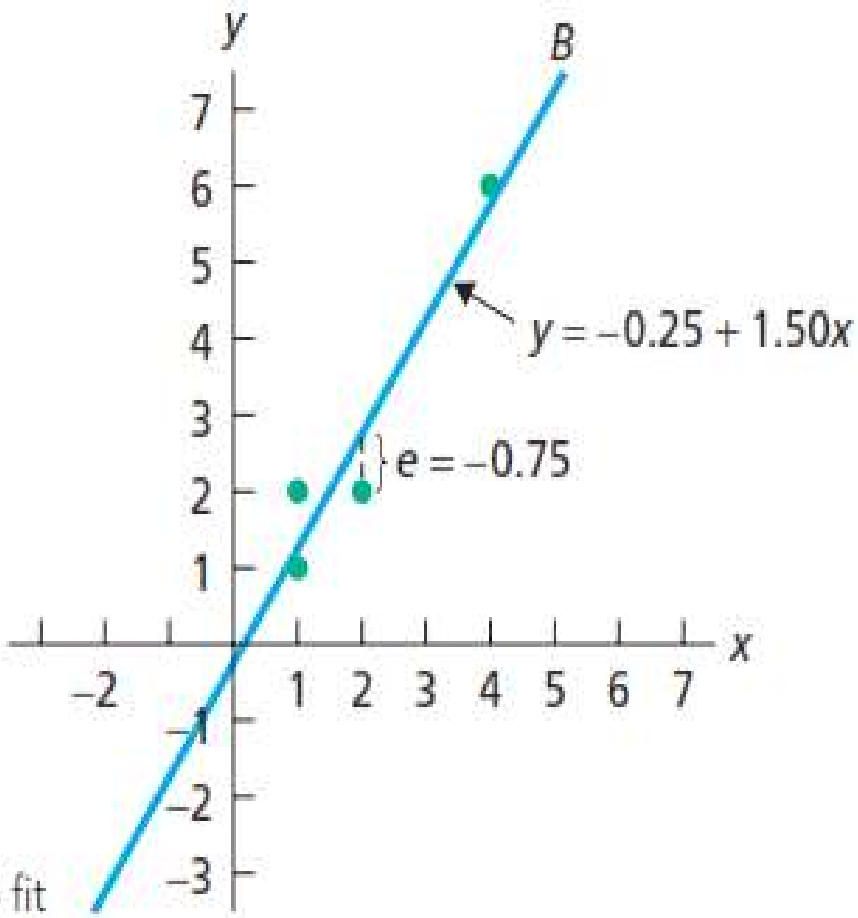
Least squares Method

Line A: $y = 0.50 + 1.25x$



(a)

Line B: $y = -0.25 + 1.50x$



(b)

FIGURE

Two possible lines to fit the data points in Table 14.3

Least squares Method

For example , as we have just demonstrated, Line A predicts a Y-value of $\widehat{Y}$= 3 when x = 2. The actual y-value for x = 2 is Y = 2 . So, the error made in using Line A to predict the Y-value of the data point (2, 2) is

$$e = Y - \hat{Y} = 2 - 3 = -1$$

An Error e:

In general, an error, e, is the signed vertical distance from the line to a data point.

Least squares Method

The fourth column of Table (a) below shows the errors made by Line A for all four data points; the fourth column of Table (b) shows the same for Line B.

Line A: $y = 0.50 + 1.25x$

x	y	$\hat{y}$	e	e^2
1	1	1.75	−0.75	0.5625
1	2	1.75	0.25	0.0625
2	2	3.00	−1.00	1.0000
4	6	5.50	0.50	0.2500
				1.8750

(a)

Line B: $y = -0.25 + 1.50x$

x	y	$\hat{y}$	e	e^2
1	1	1.25	−0.25	0.0625
1	2	1.25	0.75	0.5625
2	2	2.75	−0.75	0.5625
4	6	5.75	0.25	0.0625
				1.2500

(b)

Least squares Method

To decide which line, Line A or Line B, fits the data better, we first compute the sum of the squared errors, e_i^2, in the final column of Table (a) and Table (b). The line having the smaller sum of squared errors, in this case Line B, is the one that fits the data better. Among all lines, the least-squares criterion is that the line having the smallest sum of squared errors is the one that fits the data best.

Least squares Method

The least-squares method is that the line that best fits a set of data points is the one having the smallest possible sum of squared errors.

Least squares Method

Regression line:

- The line that best fits a set of data points according to the least-squares method.

Regression equation:

- The equation of the regression line.

Least squares Method

Notation use in Regression

For a set of n data points, the defining and computing formulas for S_{xx} , S_{xy}, and S_{yy} are as follows:

Quantity	Defining formula	Computing formula
S_{xx}	$\Sigma(x_i - \bar{x})^2$	$\Sigma x_i^2 - (\Sigma x_i)^2/n$
S_{xy}	$\Sigma(x_i - \bar{x})(y_i - \bar{y})$	$\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)/n$
S_{yy}	$\Sigma(y_i - \bar{y})^2$	$\Sigma y_i^2 - (\Sigma y_i)^2/n$

Least squares Method

Regression equation

The regression equation for a set of n data points is

$$\hat{y} = b_0 + b_1 x$$

Where:

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum x_i y_i - (\sum x_i)(\sum y_i)/n}{\sum {x_i}^2 - (\sum x_i)^2/n}$$

$$b_0 = \frac{1}{n}\left(\sum y_i - b_1 \sum x_i\right) = \bar{y} - b_1 \bar{x}$$

1. $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$, where $S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$ and $S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$.
2. $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$.

Regression equation (Example)

Age and Price of Orion's in the first two columns of Table below, we repeat our data on age and price for a sample of 11 Orions.

Age (yr) x	Price ($100) y
5	85
4	103
6	70
5	82
5	89
5	98
6	66
6	95
2	169
7	70
7	48
58	975

Regression equation (Example)

a) Determine the regression equation for the data.

b) Graph the regression equation and the data points.

c) Describe the apparent relationship between age and price of Orions.

d) Interpret the slope of the regression line in terms of prices for Orions.

e) Use the regression equation to predict the price of a 3-year-old Orion and a 4-year-old Orion.

Solution

We first need to compute b_1 and b_0 by using Formula. We did so by constructing a table of values for x (age), y (price), xy, x 02 , and their sums in

Table 14.5. The slope of the regression line therefore is

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum x_i y_i - (\sum x_i)(\sum y_i)/n}{\sum {x_i}^2 - (\sum x_i)^2/n}$$

$$= \frac{4732 - (58)(975)/11}{326 - (58)^2/11} = -20.26$$

$$b_0 = \frac{1}{n}(\sum y_i - b_1 \sum x_i) = \frac{1}{11}\big(975 - (-20.26)(58)\big) = 195.47$$

Solution

So the regression line is

$$\hat{y} = 195.47 - 20.26x$$

(b) To graph the regression equation, we need to substitute two different *x-values* in the regression equation to obtain two distinct points. Let's use the *x-values 2* and 8. The corresponding *y-values are*

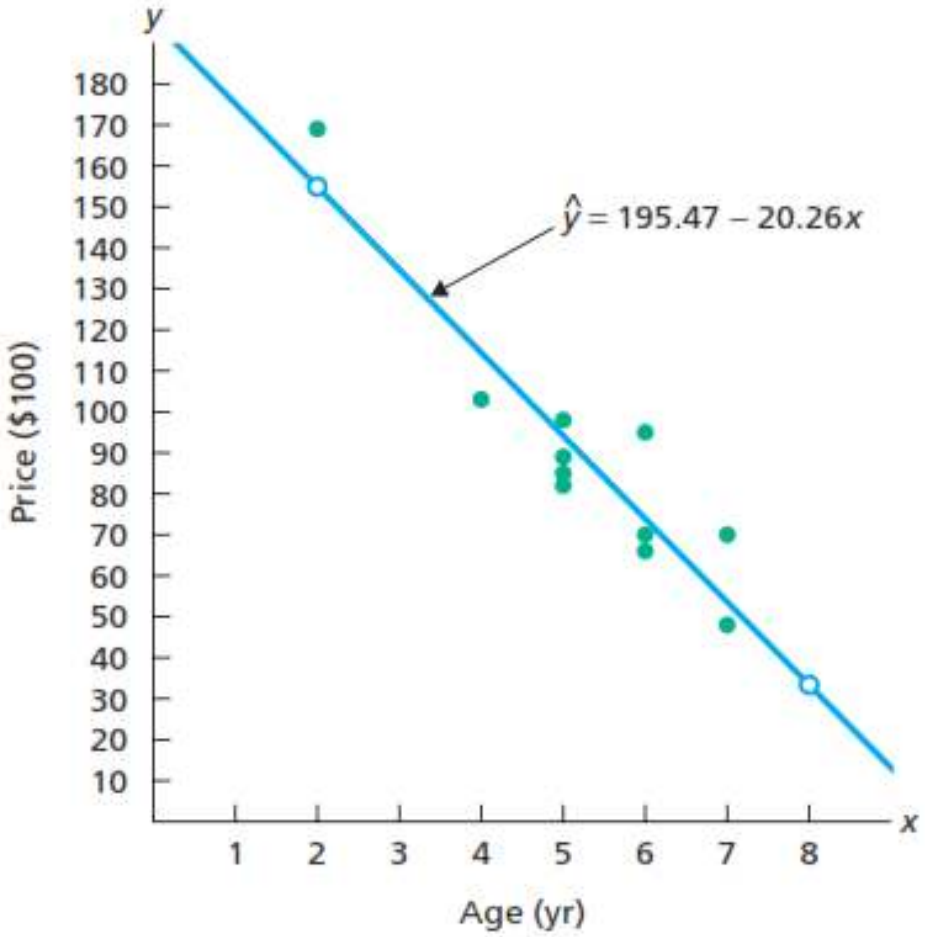
$$\hat{y} = 195.47 - 20.26(2) = 154.95$$

$$\hat{y} = 195.47 - 20.26(8) = 33.39$$

Therefore, the regression line goes through the two points (2, 154.95) and (8, 33.39).

In Fig. below , we plotted these two points with open dots. Drawing a line through the two open dots yields the regression line, the graph of the regression equation. This Figure also shows the data points from the first two columns of Table.

Age (yr) x	Price ($100) y	xy	x^2
5	85	425	25
4	103	412	16
6	70	420	36
5	82	410	25
5	89	445	25
5	98	490	25
6	66	396	36
6	95	570	36
2	169	338	4
7	70	490	49
7	48	336	49
58	975	4732	326



(c) Because the slope of the regression line is negative, price tends to decrease as age increases, which is no particular surprise.

(d) Because x represents age in years and y represents price in hundreds of dollars, the slope of -20.26 indicates that Orions depreciate an estimated $2026 per year, at least in the 2- to 7-year-old range.

(e) For a 3-year-old Orion, x = 3, and the regression equation yields the predicted price of

$$\hat{y} = 195.47 - 20.26(3) = 134.69$$

Similarly, the predicted price for a 4-year-old Orion is

$$\hat{y} = 195.47 - 20.26(4) = 114.43$$

Interpretation The estimated price of a 3-year-old Orion is $13,469, and the estimated price of a 4-year-old Orion is $11,443.

quiz

Let's have the following data points.

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

- **Find the regression equation for the data point.**
- **Graph the regression equation and the data points.**

Linear Correlation

The linear correlation Coefficient

- **The linear correlation coefficient is a descriptive measure of the strength and direction of the linear (straight-line) relationship between two variables.**
- **Represented by the symbol *r*.**

The linear correlation Coefficient

- The values of this coefficient vary from –1, which indicates perfect negative correlation, to +1, which indicates perfect positive correlation.
- The sign of the correlation coefficient *r* is the same as the sign of the slope.
  - If the slope is positive, *r* is positive.
  - If the slope is negative, *r* is negative.

The linear correlation Coefficient

For a set of *n* data points, the linear correlation coefficient is defined by

$$r = \frac{\frac{1}{n-1}\Sigma(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$

Where S_x , S_y denote the sample standard deviation of the x-values and y-values respectively

Using algebra, we can show that the linear correlation coefficient can be expressed as

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \hat{\beta}_1\sqrt{\frac{S_{xx}}{S_{yy}}}.$$

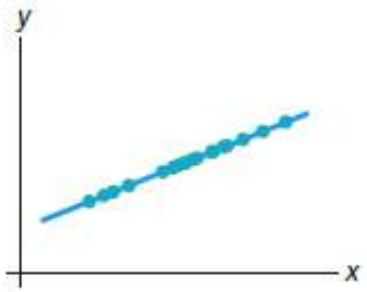
The linear correlation Coefficient

The computing formula for a linear correlation coefficient is

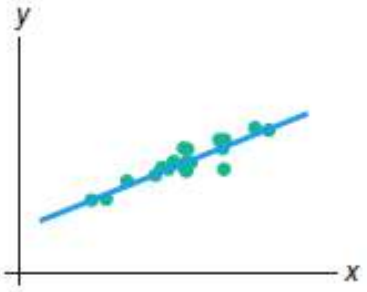
$$r = \frac{\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)/n}{\sqrt{[\Sigma x_i^2 - (\Sigma x_i)^2/n][\Sigma y_i^2 - (\Sigma y_i)^2/n]}}.$$

The computing formula is almost always preferred for hand calculations, but the defining formula reveals the meaning and basic properties of the linear correlation coefficient.

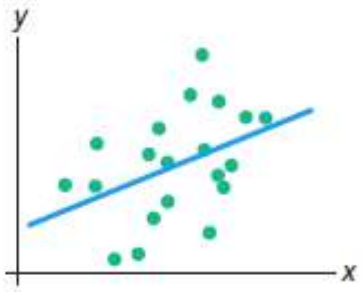
Various degrees of linear correlation



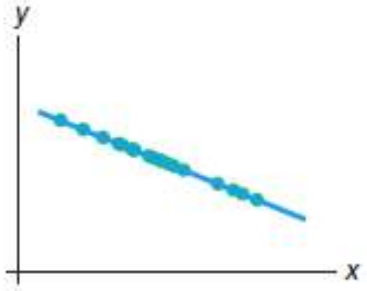
(a) Perfect positive linear correlation
$r = 1$



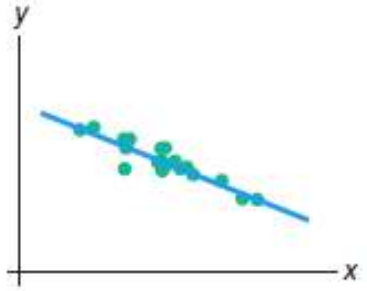
(b) Strong positive linear correlation
$r = 0.9$



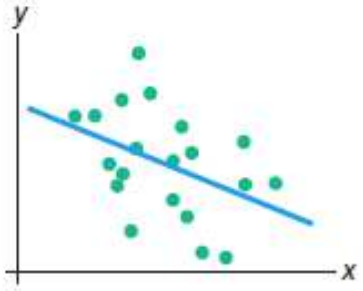
(c) Weak positive linear correlation
$r = 0.4$



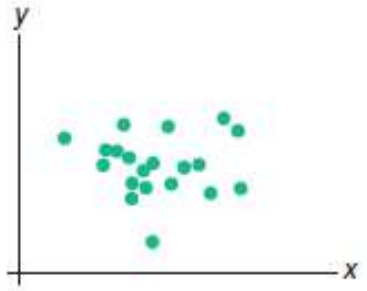
(d) Perfect negative linear correlation
$r = -1$



(e) Strong negative linear correlation
$r = -0.9$



(f) Weak negative linear correlation
$r = -0.4$



Example

Age (yr) x	Price ($100) y	xy	x^2	y^2
5	85	425	25	7,225
4	103	412	16	10,609
6	70	420	36	4,900
5	82	410	25	6,724
5	89	445	25	7,921
5	98	490	25	9,604
6	66	396	36	4,356
6	95	570	36	9,025
2	169	338	4	28,561
7	70	490	49	4,900
7	48	336	49	2,304
58	975	4732	326	96,129

1) Compute the linear correlation coefficient for the data.
2) Interpret the result in terms of the relationship between the variables age and price of Orions .
3) Discuss the graphical implications of the value of r.

Example

Age (yr) x	Price ($100) y	xy	x^2	y^2
5	85	425	25	7,225
4	103	412	16	10,609
6	70	420	36	4,900
5	82	410	25	6,724
5	89	445	25	7,921
5	98	490	25	9,604
6	66	396	36	4,356
6	95	570	36	9,025
2	169	338	4	28,561
7	70	490	49	4,900
7	48	336	49	2,304
58	975	4732	326	96,129

$$r = \frac{\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)/n}{\sqrt{\left[\Sigma x_i^2 - (\Sigma x_i)^2/n\right]\left[\Sigma y_i^2 - (\Sigma y_i)^2/n\right]}}$$

$$= \frac{4732 - (58)(975)/11}{\sqrt{\left[326 - (58)^2/11\right]\left[96,129 - (975)^2/11\right]}} = -0.924.$$

Example

Interpretation:

The linear correlation coefficient, $r = -0.924$, suggests a strong negative linear correlation between age and price of Orions. In particular, it indicates that as age increases, there is a strong tendency for price to decrease, which is not surprising.

Because the correlation coefficient, $r = -0.924$, is quite close to −1, the data points should be clustered closely about the regression line.

Correlation and causation

- Two variables may have a high correlation without being causally related
- A correlation coefficient close to zero does not necessarily mean that X and Y are not related .
- Two variables may be strongly correlated because they are both associated with other variables, called lurking variables, that cause changes in the two variables under consideration.