CHAPTER 4: DESCRIBING THE RELATION BETWEEN TWO VARIABLES

[This chapter is based on Chapter 4 from the textbook]

In Chapters 1, 2 and 3 we learned how to collect and summarize data in which a single variable is measured. That is, we ask one question to each individual of the sample or the population. For example, in the last example of Chapter 3 the variable was the number of chocolate chips in Keebler's and store brand cookies.

In this chapter we will learn how to describe bivariate data, that is, data in which we study two variables. In the chocolate chip cookies example, two variables could be the number of chocolate chips and the weight of each cookie. We are interested in learning whether the two variables are related, how they are related, and if one can explain the variability of the other variable. For example, can the weight of a cookie be predicted using the number of chocolate chips it has?

4.1 Scatter Diagrams and Correlation

We start with a key definition.

Definition 4.1. *The response variable is the variable whose value can be explained by the value of the explanatory or predictor variable.*

Let's see some examples.

Example 4.1. *In the following examples, determine which variable is the likely explanatory variable and which one is the likely response variable.*

(a) Speed at which a golf club is swung, and distance the golf ball travels.

(b) Gestation period when a baby is born, and weight of a baby when born

Solution.

(a) The speed of the golf club is the explanatory variable and the distance the golf ball travels is the response variable.

(b) The gestaion period is the explanatory variable and the weight when born is the response variable.

□

The first step to analyze the relationship between the explanatory and the reponse variables is to draw their values. We use a scatter diagram, defined below.

Definition 4.2. *A scatter diagram is a graph that shows the relationship between two quantitative variables measured on the same individual. Each individual in the dataset is represented by a point in the scatter diagram. The explanatory variable is plotted on the horizontal axis, and the response variable on the vertical axis.*

Let's see an example.

Example 4.2. *An engineer wants to determine how the weight of a car affects gas mileage. The following data represent the weights of various domestic cars and their gas mileages in the city for the 2015 model year.*

Car	*Weight (lb)*	*Miles per Gallon*
Buick LaCrosse	*4724*	*17*
Cadillac CTS	*4006*	*18*
Chevrolet Cruze	*3097*	*22*
Chevrolet Impala	*3555*	*19*
Chrysler 300	*4029*	*19*
Dodge Charger	*3934*	*19*
Dodge Dart	*3242*	*24*
Ford Focus	*2960*	*26*
Ford Mustang	*3530*	*19*
Lincoln MKZ	*3823*	*18*

Source: Manufacturers' website

(a) Determine which variable is the likely explanatory variable and which is the likely response variable

(b) Draw a scatter diagram of the data. What do you observe?

The following algorithm shows how to construct a scatter diagram in the TI83/84 calculator.

Algorithm 4.1 (Scatter diagram in TI83/84)**.**

1 Enter the data:

i. Press `stat` *and* `1:Edit...`

ii. Enter the explanatory variable in `L1` *and the response variable in* `L2`

iii. Press the keys `2nd` *and* `mode` *to quit*

2 Press `2nd` *and* `Y=` *to open the* `stat plot` *menu. Select* `1:Plot1`

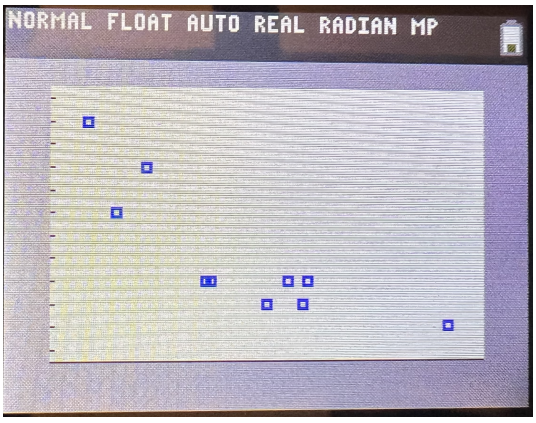
3 Turn `Plot 1` *by highlighting the* `On` *button and pressing* `Enter`

4 Highlight the scatter diagram icon (first icon from left to right) and press `Enter`.

5 Make sure that `Xlist` *is* `L1` *and* `Ylist` *is* `L2`

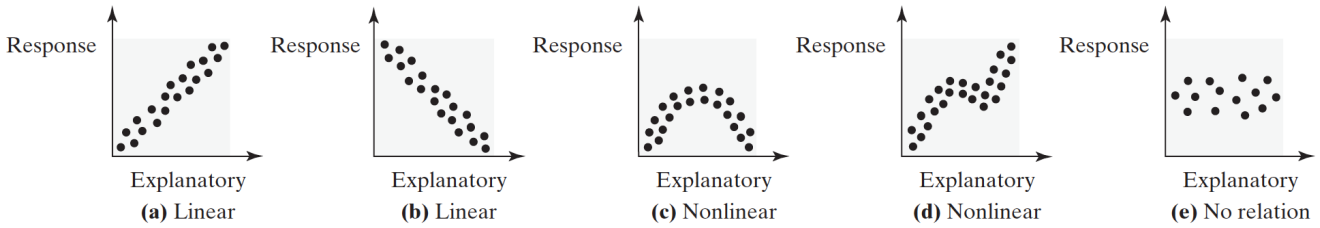
6 Press `ZOOM` *and select* `9:ZoomStat`

Solution. It is likely that the weight of the cars is the explanatory variable and the mileage is the response variable. Plotting the weight of each car in the horizontal axis and the mileage in the vertical axis, we obtain the following scatter diagram:



We see that the larger the weight, the smaller the mileage. □

Different pairs of variables will show different scatter diagrams. In the figure below we summarize some possible outcomes.



We are interested in determining if the relation is linear or nonlinear. When the scatter diagram shows the data around a straight line as in figures (a) and (b), we say that the relation is linear. If the diagram shows any other function, such as (c) and (d), we say that the relation between the variables in nonlinar. Finally, if the scatter diagram shows a cloud of points without clear shape as in figure (e), we say that there is no relation between the variables.

In the following definition we introduce another relation that variables can have.

Definition 4.3. *Two variables that are linearly related are:*

(i) *positively associated* *if, whenever the value of one variable increases, the value of the other variable also increases.*

(ii) *negatively associated* *if, whenever the value of one variable increases, the value of the other variable decreases.*

For example, panel (a) shows a pair of variables positively associated, and pannel (b) shows a pair that is negatively associated. Similarly, our scatter diagram in Example 4.2 shows that weight and mileage are negatively associated. That is, as the weight of a car increases, its mileage decreases.

Now, relying on a graph can be tricky. Remember that we've been discussing how the scale of the axes can confuse our eyes. Below we formally define a coefficient that can tell us if two variables are positively or negatively associated.

Definition 4.4. *The linear correlation coefficient or Pearson product moment correlation coefficient is a measure of the strength and direction of the linear relation between two quantitative variables.*

The sample correlation coefficient is represented by the letter r, and is defined as follows:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right)$$

where:

x_i : i^{th} *observation of explanatory variable*
$\overline{x}$: *sample mean of explanatory variable*
s_x : *standard deviation of explanatory variable*
y_i : i^{th} *observation of response variable*
$\overline{y}$: *sample mean of response variable*
s_y : *standard deviation of response variable*
n : *number of individuals in the sample*

Let's revise some properties of the correlation coefficient, and then we will analyze how the scatter diagram of various values of r look like.

Properties:

(1) The linear correlation coefficient r always satisfies $-1 \leq r \leq 1$

(2) If $r = +1$, then a perfect positive linear correlation exists. Plot (a) in the figure below shows this situation.

(3) The closer r is to $+1$, the stronger is the evidence of positive association of the two variables. Plots (a), (b) and (c) show different values of positive r.

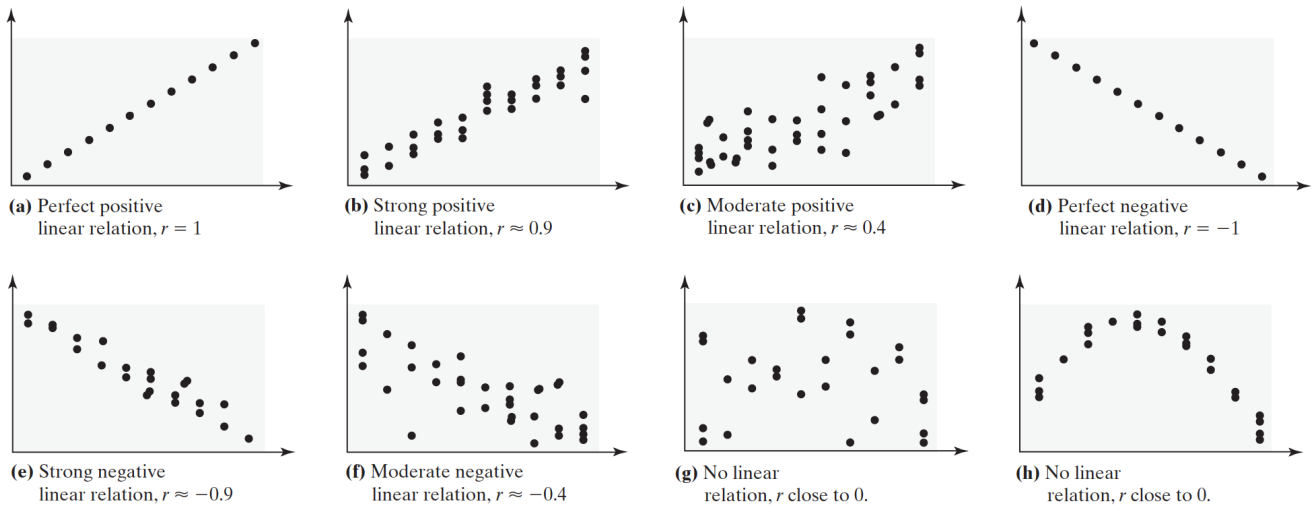
(4) If $r = -1$, then a perfect negative correlation exists. Plot (d) shows this situation.

(5) The closer r is to -1, the stronger is the evidence of negative association. Plots (d), (e) and (f) show different values of r when negative.

(6) If $r = 0$ or r is close to zero, there is no evidence of linear negative association. This does **not** mean that the variables are not related; it only means that they are not linearly related. Plots (g) and (h) show scatter diagrams of variables with r close to zero.

(7) r is a uniteless measure, so the unit of measure of x and y does not matter.

(8) An observation that does not follow the overall pattern of the data could affect the value of the linear correlation coefficient.



(a) Perfect positive linear relation, $r = 1$

(b) Strong positive linear relation, $r \approx 0.9$

(c) Moderate positive linear relation, $r \approx 0.4$

(d) Perfect negative linear relation, $r = -1$

(e) Strong negative linear relation, $r \approx -0.9$

(f) Moderate negative linear relation, $r \approx -0.4$

(g) No linear relation, r close to 0.

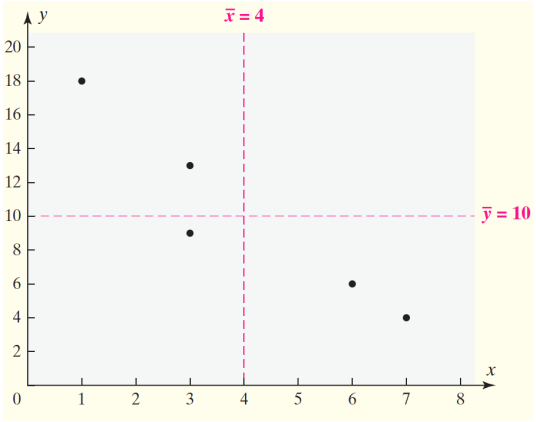
(h) No linear relation, r close to 0.

Observe that the terms of the sum in the definition of r are the product between the z-score of each observation of x and y. Then, positive r means that whenever the observations of x are above its mean, the observations of y are also above the mean. Similarly when the observations are below the mean. If r is negative, instead, it means that whenever the z-score of one variable is positive, the other variable is negative. That is, whenever an observation of x is above the mean, the corresponding value of y is below the mean, and vice-versa.

Let's solve an example to understand how r works.

Example 4.3. *For the data shown in the following table, compute the linear correlation coefficient. A scatter diagram of the data is presented too. Use the diagram to interpret the value of* r.

x	y
1	18
3	13
3	9
6	6
7	4



Solution. Before computing the coefficient r we compute the mean and sample standard deviation of x and y. We obtain the following means:

$$\overline{x} = \frac{1+3+3+6+7}{5} = 4$$
$$\overline{y} = \frac{18+13+9+6+4}{5} = 10$$

and the following standard deviation:

$$s_x = \sqrt{\frac{(1-4)^2+(3-4)^2+(3-4)^2+(6-4)^2+(7-4)^2}{4}} = 2.449$$
$$s_y = \sqrt{\frac{(18-10)^2+(13-10)^2+(9-10)^2+(6-10)^2+(4-10)^2}{4}} = 5.612$$

We use the following table to compute r:

x	y	z-score of x (z_x)	z-score of y (z_y)	$z_x z_y$
1	18	$\frac{1-4}{2.449} = -1.2247$	$\frac{18-10}{5.612} = 1.4254$	-1.7457
3	13	$\frac{3-4}{2.449} = -0.4083$	$\frac{13-10}{5.612} = 0.5345$	-0.2182
3	9	$\frac{3-4}{2.449} = -0.4083$	$\frac{9-10}{5.612} = -0.1782$	0.0727
6	6	$\frac{6-4}{2.449} = 0.8165$	$\frac{6-10}{5.612} = -0.7217$	-0.5819
7	4	$\frac{7-4}{2.449} = 1.2247$	$\frac{4-10}{5.612} = -1.0690$	-1.3093

Summing the last column we obtain:

$$r = \frac{-3.7824}{5-1} = -0.946$$

That is, there is a strong negative correlation. As we see in the diagram, the variables x and y form an almost perfect decreasing straight line. □

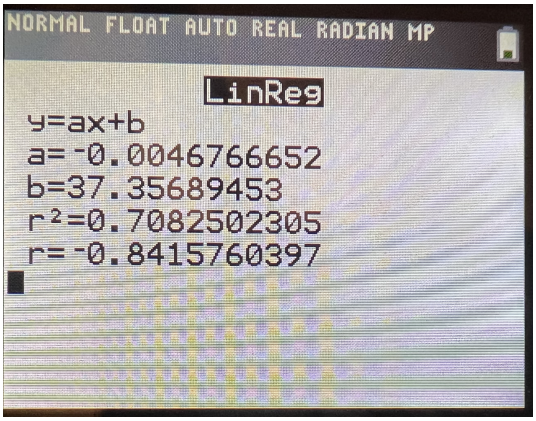
In general, we will not compute the coefficient r by hand. Instead, we will use our calculator as indicated in the following algorithm.

Algorithm 4.2 (Computing r with a TI83/84)**.**

1. *Enter the explanatory variable in* `L1` *and the response variable in* `L2` *as indicated in step 1 of Algorithm 4.1*
2. *Turn on the diagnostics as follows (you only need to do this once):*
   i. *Select the catalog pressing* `2nd` *and* `0`
   ii. *Scroll down and select* `DiagnosticsOn`
   iii. *Press* `Enter` *twice*
3. *From the home screen, press* `stat`
4. *Highlight* `CALC` *and select* `4:LinReg(ax+b)`
5. *Select* `L1` *for* `Xlist` *and* `L2` *for* `Ylist`.
6. *Highlight* `Calculate` *and press* `Enter`

Example 4.4. *Use your calculator to compute the linear correlation coefficient of the data in Example 4.2*

Solution. We obtain the following on the calculator.



Hence, $r = -0.842$. That is, there is negative linear correlation. □

So far, we only said that if r is close to zero, then there is no evidence of linear correlation. But how much is "close" to zero? In the Table from the figure[1] below we present the critical value of the absolute value of r to determine that there is no correlation.

[1] Table II from Appendix A of the textbook

Table II

Critical Values for Correlation Coefficient

n	
3	0.997
4	0.950
5	0.878
6	0.811
7	0.754
8	0.707
9	0.666
10	0.632
11	0.602
12	0.576
13	0.553
14	0.532
15	0.514
16	0.497
17	0.482
18	0.468
19	0.456
20	0.444
21	0.433
22	0.423
23	0.413
24	0.404
25	0.396
26	0.388
27	0.381
28	0.374
29	0.367
30	0.361

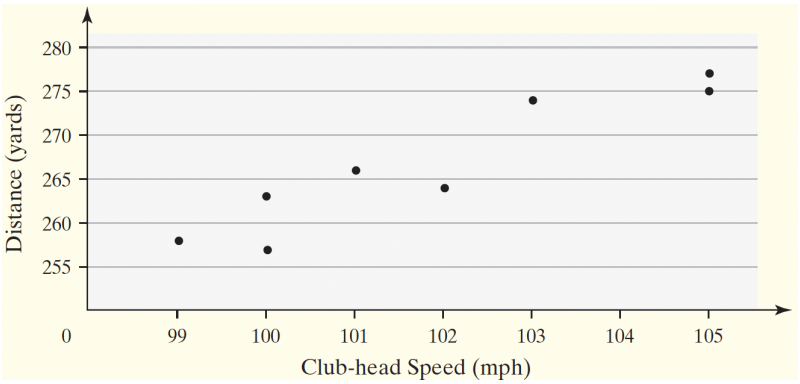
For example, if a sample of size 10 has $r = 0.5$, then there is no evidence of correlation. However, if a sample of size 30 has $r = 0.6$, we conclude that there is positive correlation. Similarly, if a sample of size 15 has $r = -0.7$, we say that there is negative correlation between the variables because $|r| = 0.7 > 0.514$.

To close this section, let's discuss the difference between correlation and causation. Correlation means that the index r has big absolute value (with respect to the table above). However, causation means that changes in the variable x provoke changes in the variable y. When we design an experiment and measure correlation, we usually observe causation too. However, many variables have positive correlation by coincidence. To prevent concluding causation wrongly, we need to be careful with lurking variables. For example, we may study the relation between the speed at which an ice cream melts with the number of people getting sunburn in differet days. We will probably observe that the faster an ice cream melts, the more people get sunburn. But, is the ice cream melt causing the sun burn? No! There is a lurking variable that we should consider: the weather.

4.2 Least-Squares Regression

We've been classifying the relation between two variables as linear or nonlinear depending on what we see in a scatter plot. Further, we introduced the linear correlation coefficient r to confirm what we see. In this chapter we focus on characterizing the line that best described the relation between the explanatory and the response variable.

We develop the main ideas using the following scatter plot, that relates the speed of a club head when it hits a golf ball, and the distance at which the ball lands.



Observe that there is a positive correlation between the variables, and it seems the relation is linear. Indeed, these data points have linear correlation coefficient $r = 0.939$ so they are linearly correlated. But what is the best line to describe the relation between the variables?

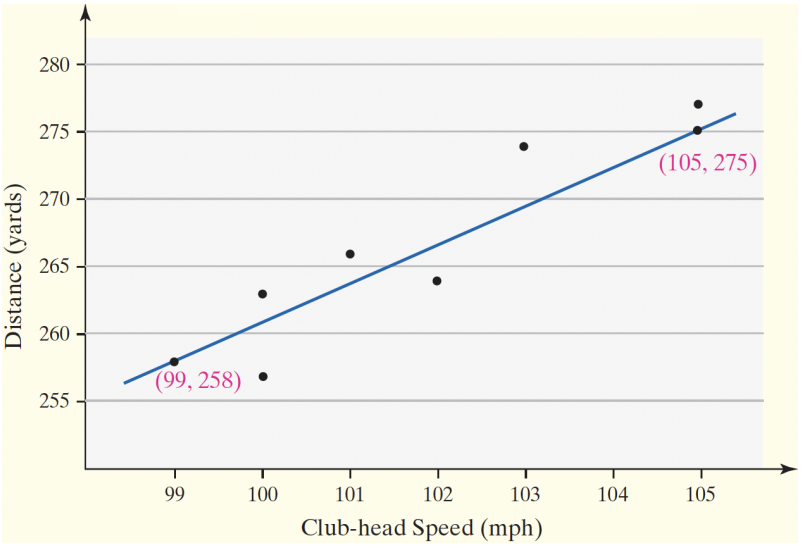
We are interested in computing this line because the data only gives us the value of the response variable y for some values of the explanatory x. However, we would like to predict the value of y for values of x that are not necessary part of the sample. For example, in the example above, we would like to predict the distance at which a ball lands if we hit it at 104 mph with the club head, or even at 98 mph.

Recall that a line can be described with two parameters: the slope and the intercept with the vertical axis. Using a to denote the slope and b to denote the intercept, the formula of the corresponding line is $ax + b$, where x is the explanatory variable. Continuing with the examples above, if we knew the values of a and b that relate the speed of the golf club head and the distance at which the ball lands, we could predict the landing distance if we hit the ball at 98 mph or 104 mph. We would have:

$$\hat{y}_1 = a \cdot 98 + b$$
$$\hat{y}_2 = a \cdot 104 + b$$

We add a hat on top of y because it is not an *observed* value from the sample; instead, it is a *predicted* value.

The easiest way to choose a line that predicts the relation between two variables is choosing two points in the scatter plot and draw a line between them, as shown in the figure below:



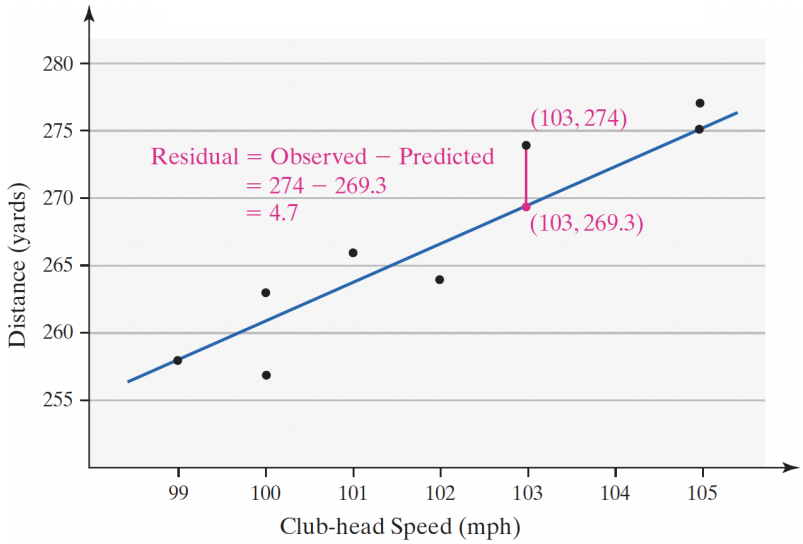
In the case of the line drawn in the figure, we have $a = 2.8333$ and $b = -22.4967$. Then, the prediction for the landing distance if the speed of the club head is 98 mph and 104 mph are:

$$\hat{y}_1 = a \cdot 98 + b = 2.8333 \cdot 98 - 22.4967 = 255.1667$$
$$\hat{y}_2 = a \cdot 104 + b = 2.8333 \cdot 104 - 22.4967 = 272.1665$$

Now, what happens with the points in the scatter plot that are not on top of the line? There will be a difference between the predicted value $\hat{y}$ and the observed value y, as we define below.

Definition 4.5. *The error or residual is the difference between the observed value of the response variable and the predicted value.*

The figure below shows an example:



where the residual of the observation $(103, 274)$ is computed as follows. The observed value is $y = 274$ and the predicted value is

$$\hat{y} = a \cdot 103 + b = 2.8333 \cdot 103 - 22.4967 = 269.3332$$

Then, the residual is:

$$\text{Residual} = y - \hat{y} = 274 - 269.3332 = 4.6668$$

When we computed the line in this example we simply took two representative points in the scatter plot and drew the line between them. In the rest of this chapter we will use a systematic approach to compute such line, where the goal is to minimize the total square of the residuals. We consider the square because the residuals can be positive or negative, and we don't want the sum cancelling out terms.

Definition 4.6. *The least-squares regression line is the line that minimizes the sum of the squared errors (or residuals). The line minimizes the sum of the squared distance between the observed values y and those predicted by the line $\hat{y}$.*

As a result of the least-squares regression line definition, we obtain a line of the form

$$\hat{y} = ax + b,$$

where

$$a = r\frac{s_y}{s_x} \qquad \text{and} \qquad b = \overline{y} - a\overline{x},$$

with

r	:	Linear correlation coefficient
s_x	:	Standard deviation of explanatory variable x
s_y	:	Standard deviation of response variable y
$\overline{x}$	:	Sample mean of explanatory variable x
$\overline{y}$	:	Sample mean of response variable y

Let's see an example.

Example 4.5. *Consider the data from Example 4.3 and compute the least-squares regression line.*

Solution. Recall that we already computed:

$$\overline{x} = 4, \quad \overline{y} = 10, \quad s_x = 2.499, \quad s_y = 5.612, \quad r = -0.946$$

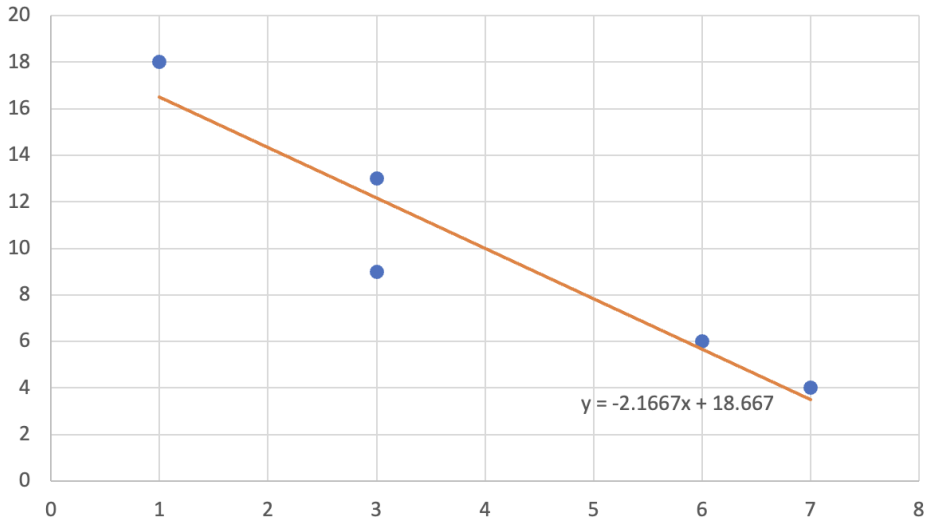
Then, we obtain:

$$a = -0.946 \cdot \frac{5.612}{2.499} = -2.1244$$
$$b = 10 - (-2.1244) \cdot 4 = 18.4977$$

That is, the least-squares regression line is

$$\hat{y} = -2.1244x + 18.4977$$

The figure below shows the scatter graph with the regression line, generated in Excel.

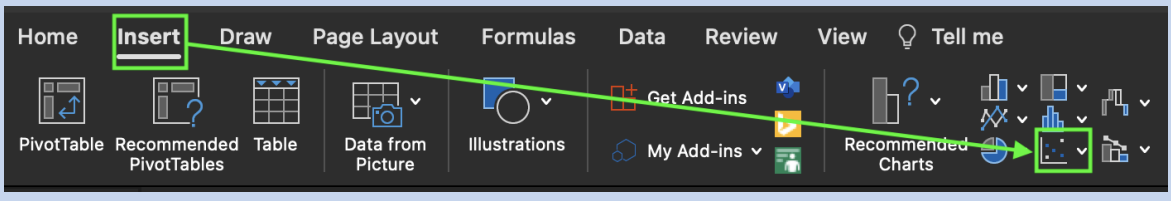


□

To obtain a scatter plot and add the least-squares regression line in Excel, we use the following algorithm.

Algorithm 4.3 (Scatter plot and least-squares regression line in Excel)**.**

1. *Enter the data using one column for the explanatory variable, and another column for the response column*
2. *Scatter plot:*
   - i. *Select the data*
   - ii. *Click on the* `Insert` *tab*
   - iii. *Click on the scatter plot icon as shown below*

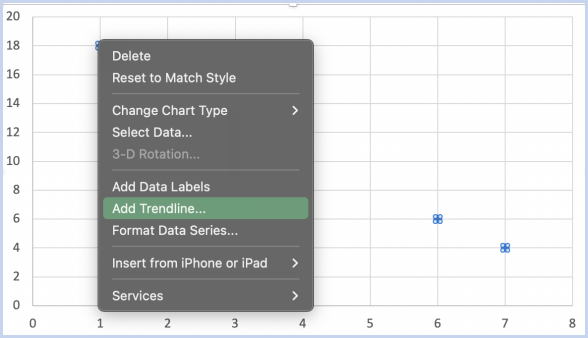


3. *Least-squares regression line:*

   i. *Click on one point of the scatter diagram*

   ii. *Click the right button of your mouse to open the menu*

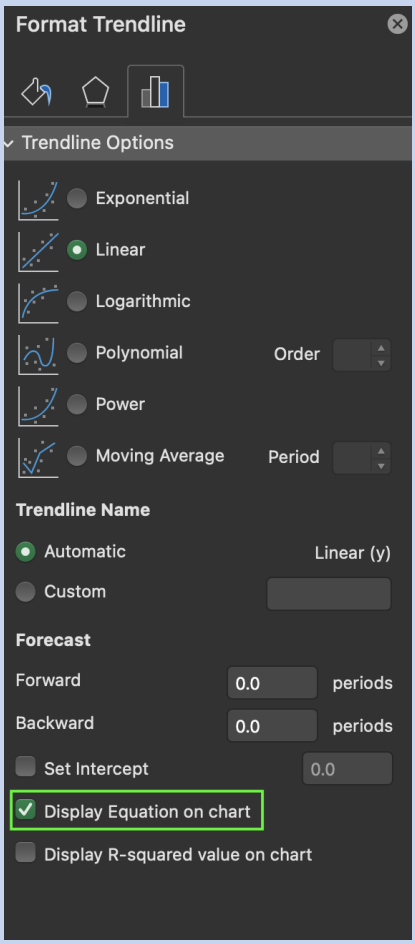
   iii. *Click on* `Add trend line`*, as shown in the figure below:*



4. *To show the parameters a and b of the line, follow these steps:*

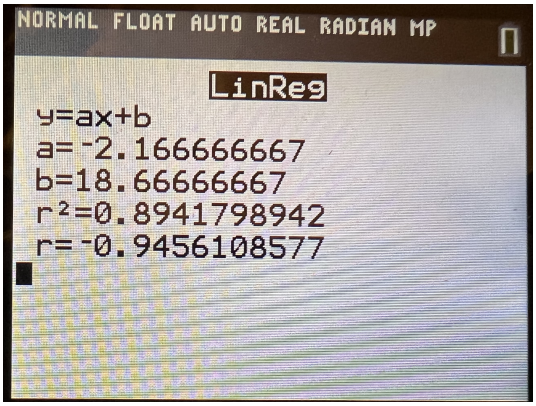
   i. *Double-click on the line obtained from the previous step*

   ii. *On the menu bar on the right-hand side of the screen, click the box that says* `Display Equation on chart`*, as shown in the figure below:*



We can easily compute the parameters a and b from the least-squares regression line using the TI calculator too. To do that, we use the same steps as in Algorithm 4.1 and extract the value of a and b.

Using the data from Example Example 4.3, we obtain the following result:



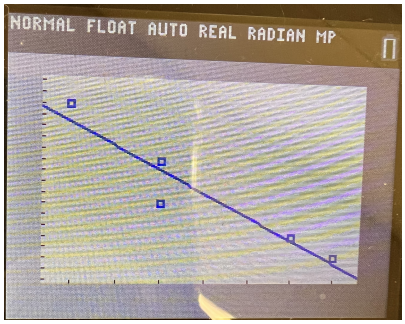
There is a small difference between our result above and the calculator, due to the amount of decimals we used in our manual computation.

We can also plot the regression line together with the scatter plot in a TI83/84 calculator, using the following algorithm.

Algorithm 4.4 (Adding the regression line to a scatter plot in a TI83/84)**.**

0. *Make sure you already plotted your scatter diagram (see Algorithm 4.1) and you computed the regression equation (see Algorithm 4.1)*
1. *Press* `y=` *and then* `vars`
2. *Use the arrows to navigate to* `5: Statistics` *or, alternatively, press* `5`
3. *Use the arrows to navigate until* `EQ` *on the menu*
4. *Navigate until* `1: RegEq` *and press* `enter` *or, alternatively, press* `1`
5. *Press* `graph` *to see the line and the scatter plot together.*

Using the data from Example 4.3, we obtain the following graph on the calculator:



The least-squares regression line is tremendously useful to understand the trend of the data points and characterize the average (or expected) change in the response variable for any unit change in the explanatory variable. However, the computation of this line is completely based on the observations we have. Hence, we need to be careful with the use of these predictions. We can only assure that the predictions are accurate within the scope of the observations (and nearby). In the golf example, predicting the distance of a ball hit at 98 mph is okay because the smallest speed in the data is 99 mph. However, we should not use the regression line to predict the distance if we hit the ball at 50 mph.

Let's do one more example before finishing the chapter. The following example includes everything we learned in Chapter 4.

Example 4.6. *A pediatrician wants to determine the relation that exists between a child's height (x) and head circumference (y). She randomly selects 11 children from her practice, and obtains the following data:*

Height (inches)	*Head circumference (inches)*
27.75	*17.5*
24.50	*17.1*
25.50	*17.1*
26.00	*17.3*
25.00	*16.9*
27.75	*17.6*
26.50	*17.3*
27.00	*17.5*
26.75	*17.3*
26.75	*17.5*
27.50	*17.5*

(a) Draw a scatter diagram of the data

(b) Determine the linear correlation coefficient

(c) Does a linear relation exist? If yes, is the association positive or negative?

(d) Determine the slope and intercept of the least-squares regression line, and draw the line on the scatter diagram.

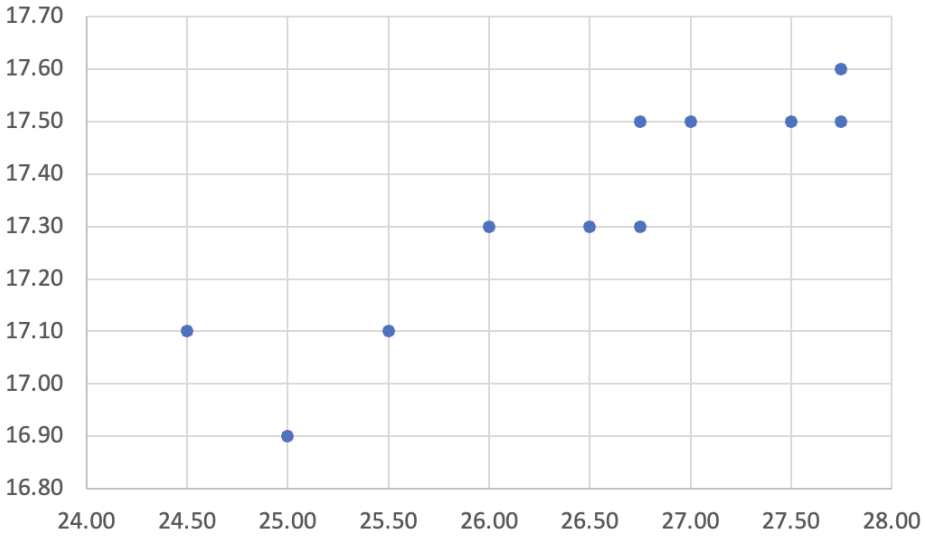
(e) Use the regression equation to predict the head circumference of a child who is 25 inches tall

(f) Compute the residual based on the observed head of the 25-inch-tall child from the data. Is the head circumference of this child above average or below average?

(g) Would it be reasonable to use the least-squares regression line to predict the head circumference of a child who is 32 inches tall? Why?

Solution.

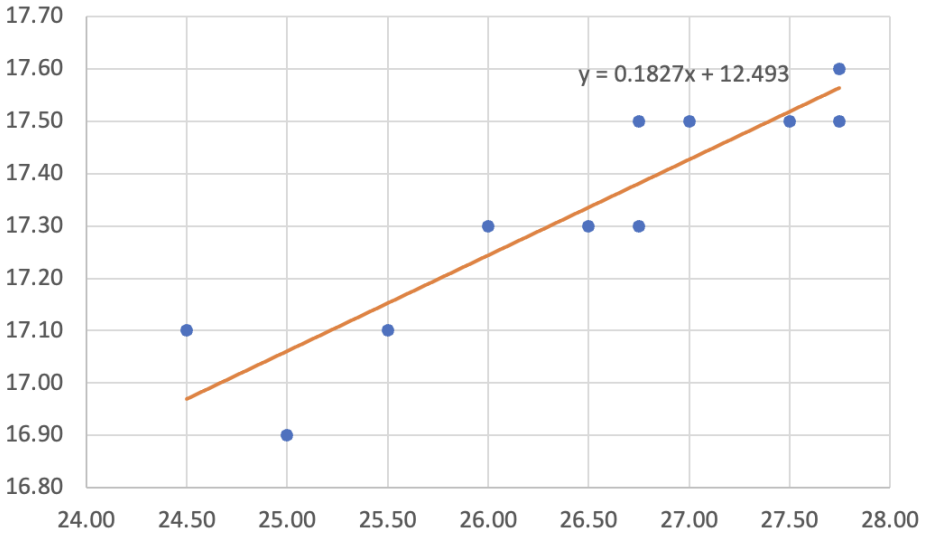
(a) We use Excel using Algorithm 4.3, and obtain the following scatter diagram



(b) We compute the value of r and obtain: $r = 0.9110$

(c) Yes, there is a linear relation with positive association. In the scatter diagram, we observe that the data points are around an increasing line and the value of r is close to $+1$.

(d) We obtain the following line:



Hence, the slope is $a = 0.1827$ and the intercept is $b = 12.493$

(e) Using the regression, we obtain that a child that is 25-inch-tall has the following predicted head circumference:

$$\hat{y} = 0.1827 \cdot 25 + 12.493 = 17.0605$$

That is, we predict a head circumference of 17.06 inches for a child that is 25-inch-tall.

(f) From the data points, we see that a child that is 25-inch-tall has a head circumference of 16.9 inches. Hence, the residual is:

$$\text{Residual} = y - \hat{y} = 16.9 - 17.06 = -0.16$$

Since the residual is negative, the child from the sample is below average.

(g) No because the maximum height of a child in the sample is 27.75 inches, which is considerably smaller than 32 inches.

□