

Figure 12.8

In this chapter, we are interested in scatter plots that show a linear pattern. Linear patterns are quite common. The linear relationship is strong if the points are close to a straight line, except in the case of a horizontal line where there is no relationship. If we think that the points show a linear relationship, we would like to draw a line on the scatter plot. This line can be calculated through a process called **linear regression**. However, we only calculate a regression line if one of the variables helps to explain or predict the other variable. If  $x$  is the independent variable and  $y$  the dependent variable, then we can use a regression line to predict  $y$  for a given value of  $x$ .

## 12.3 The Regression Equation

Data rarely fit a straight line exactly. Usually, you must be satisfied with rough predictions. Typically, you have a set of data whose scatter plot appears to "**fit**" a straight line. This is called a **Line of Best Fit or Least-Squares Line**.



### COLLABORATIVE EXERCISE

If you know a person's pinky (smallest) finger length, do you think you could predict that person's height? Collect data from your class (pinky finger length, in inches). The independent variable,  $x$ , is pinky finger length and the dependent variable,  $y$ , is height. For each set of data, plot the points on graph paper. Make your graph big enough and **use a ruler**. Then "by eye" draw a line that appears to "fit" the data. For your line, pick two convenient points and use them to find the slope of the line. Find the  $y$ -intercept of the line by extending your line so it crosses the  $y$ -axis. Using the slopes and the  $y$ -intercepts, write your equation of "best fit." Do you think everyone will have the same equation? Why or why not? According to your equation, what is the predicted height for a pinky length of 2.5 inches?

### EXAMPLE 12.6

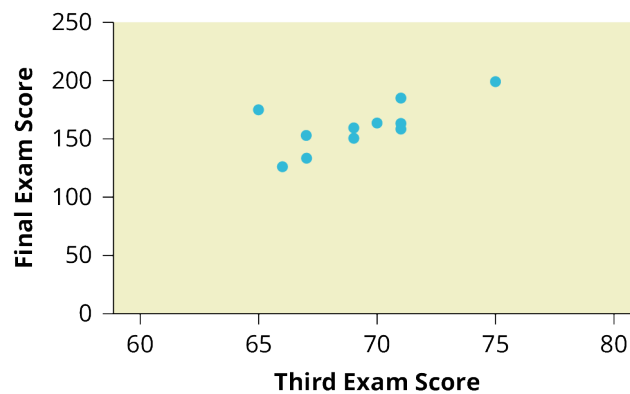
A random sample of 11 statistics students produced the following data, where  $x$  is the third exam score out of 80, and  $y$  is the final exam score out of 200. Can you predict the final exam score of a random student if you know the third exam score?

$x$ (third exam score)	$y$ (final exam score)
65	175
67	133
71	185
71	163
66	126

**Table 12.3** Table showing the scores on the final exam based on scores from the third exam.

x (third exam score)	y (final exam score)
75	198
67	153
70	163
71	159
69	151
69	159

**Table 12.3** Table showing the scores on the final exam based on scores from the third exam.



**Figure 12.9** Scatter plot showing the scores on the final exam based on scores from the third exam.

### > TRY IT 12.6

SCUBA divers have maximum dive times they cannot exceed when going to different depths. The data in [Table 12.4](#) show different depths with the maximum dive times in minutes. Use your calculator to find the least squares regression line and predict the maximum dive time for 110 feet.

X (depth in feet)	Y (maximum dive time)
50	80
60	55
70	45
80	35

**Table 12.4**

X (depth in feet)	Y (maximum dive time)
90	25
100	22

Table 12.4

The third exam score,  $x$ , is the independent variable and the final exam score,  $y$ , is the dependent variable. We will plot a regression line that best "fits" the data. If each of you were to fit a line "by eye," you would draw different lines. We can use what is called a **least-squares regression line** to obtain the best fit line.

Consider the following diagram. Each point of data is of the form  $(x, y)$  and each point of the line of best fit using least-squares linear regression has the form  $(x, \hat{y})$ .

The  $\hat{y}$  is read "**y hat**" and is the **estimated value of y**. It is the value of  $y$  obtained using the regression line. It is not generally equal to  $y$  from data.

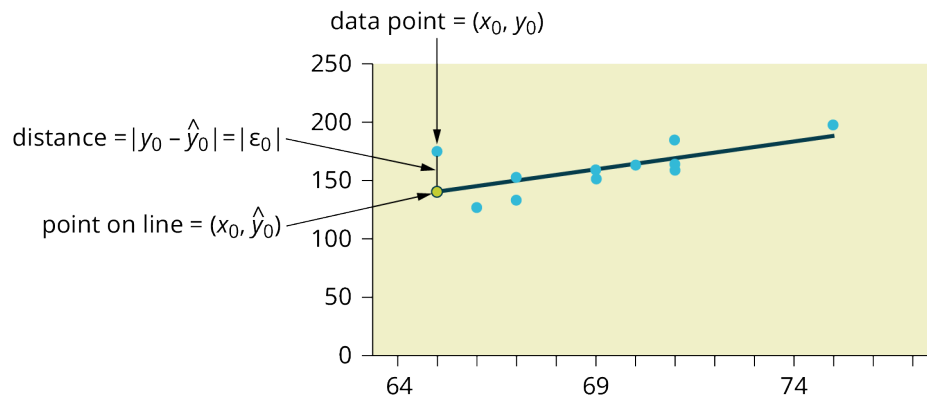


Figure 12.10

The term  $y_0 - \hat{y}_0 = \epsilon_0$  is called the **"error" or residual**. It is not an error in the sense of a mistake. The **absolute value of a residual** measures the vertical distance between the actual value of  $y$  and the estimated value of  $y$ . In other words, it measures the vertical distance between the actual data point and the predicted point on the line.

If the observed data point lies above the line, the residual is positive, and the line underestimates the actual data value for  $y$ . If the observed data point lies below the line, the residual is negative, and the line overestimates that actual data value for  $y$ .

In the diagram in [Figure 12.10](#),  $y_0 - \hat{y}_0 = \epsilon_0$  is the residual for the point shown. Here the point lies above the line and the residual is positive.

$\epsilon$  = the Greek letter **epsilon**

For each data point, you can calculate the residuals or errors,  $y_i - \hat{y}_i = \epsilon_i$  for  $i = 1, 2, 3, \dots, 11$ .

Each  $|\epsilon|$  is a vertical distance.

For the example about the third exam scores and the final exam scores for the 11 statistics students, there are 11 data points. Therefore, there are 11  $\epsilon$  values. If you square each  $\epsilon$  and add, you get

$$(\epsilon_1)^2 + (\epsilon_2)^2 + \dots + (\epsilon_{11})^2 = \sum_{i=1}^{11} \epsilon^2$$

This is called the **Sum of Squared Errors (SSE)**.

Using calculus, you can determine the values of  $a$  and  $b$  that make the **SSE** a minimum. When you make the **SSE** a minimum, you have determined the points that are on the line of best fit. It turns out that the line of best fit has the equation:

$$\hat{y} = a + bx$$

where  $a = \bar{y} - b\bar{x}$  and  $b = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}$ .

The sample means of the  $x$  values and the  $y$  values are  $\bar{x}$  and  $\bar{y}$ , respectively. The best fit line always passes through the point  $(\bar{x}, \bar{y})$ .

The slope  $b$  can be written as  $b = r \left( \frac{s_y}{s_x} \right)$  where  $s_y$  = the standard deviation of the  $y$  values and  $s_x$  = the standard deviation of the  $x$  values.  $r$  is the correlation coefficient, which is discussed in the next section.

## Residuals Plots

A residuals plot can be used to help determine if a set of  $(x, y)$  data is linearly correlated. For each data point used to create the correlation line, a residual  $y - \hat{y}$  can be calculated, where  $y$  is the observed value of the response variable and  $\hat{y}$  is the value predicted by the correlation line. The difference between these values is called the residual. A residuals plot shows the explanatory variable  $x$  on the horizontal axis and the residual for that value on the vertical axis. The residuals plot is often shown together with a scatter plot of the data. While a scatter plot of the data should resemble a straight line, a residuals plot should appear random, with no pattern and no outliers. It should also show constant error variance, meaning the residuals should not consistently increase (or decrease) as the explanatory variable  $x$  increases.

A residuals plot can be created using StatCrunch or a TI calculator. The plot should appear random. A box plot of the residuals is also helpful to verify that there are no outliers in the data. By observing the scatter plot of the data, the residuals plot, and the box plot of residuals, together with the linear correlation coefficient, we can usually determine if it is reasonable to conclude that the data are linearly correlated.

### EXAMPLE:

A shop owner uses a straight-line regression to estimate the number of ice cream cones that would be sold in a day based on the temperature at noon. The owner has data for a 2-year period and chose nine days at random. A scatter plot of the data is shown, together with a residuals plot.

Temperature ° F	Ice cream cones sold
70	105
85	240
65	49
72	147
80	231
61	38
75	193
78	196
68	89

**Table 12.5** Table showing the number of ice cream cones sold on nine random days and the temperature at noon on those days.

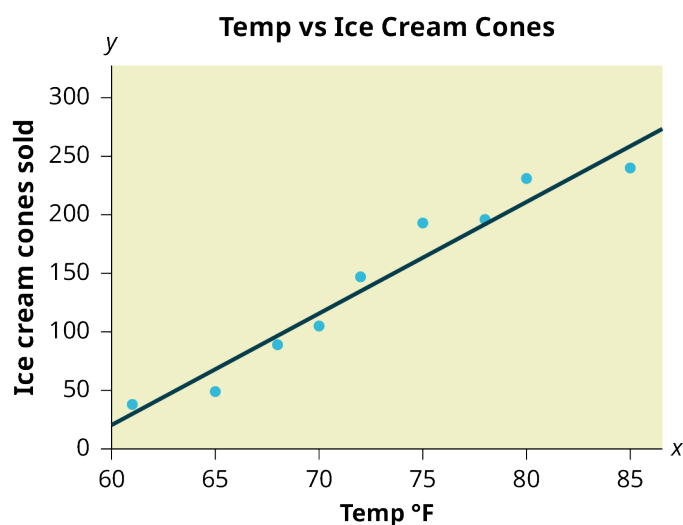


Figure 12.11 Scatter plot of the data looks like a straight line.

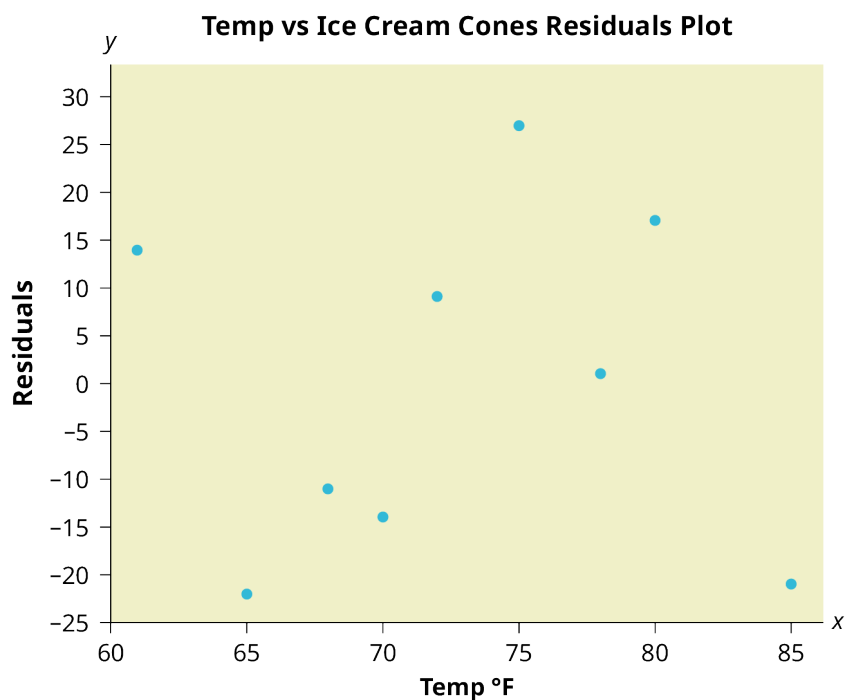


Figure 12.12 Residuals plot appears random.

### Least Squares Criteria for Best Fit

The process of fitting the best-fit line is called **linear regression**. The idea behind finding the best-fit line is based on the assumption that the data are scattered about a straight line. The criteria for the best fit line is that the sum of the squared errors (SSE) is minimized, that is, made as small as possible. Any other line you might choose would have a higher SSE than the best fit line. This best fit line is called the **least-squares regression line**.

#### NOTE

Computer spreadsheets, statistical software, and many calculators can quickly calculate the best-fit line and create the graphs. The calculations tend to be tedious if done by hand. Instructions to use the TI-83, TI-83+, and TI-84+ calculators to find the best-fit line and create a scatterplot are shown at the end of this section.

**THIRD EXAM vs FINAL EXAM EXAMPLE:**

The graph of the line of best fit for the third-exam/final-exam example is as follows:

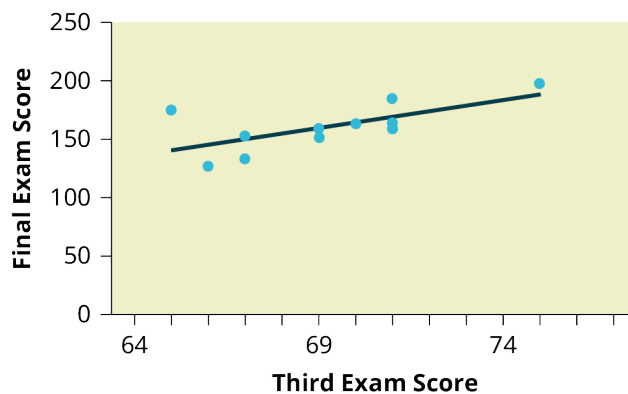


Figure 12.13

The least squares regression line (best-fit line) for the third-exam/final-exam example has the equation:

$$\hat{y} = -173.51 + 4.83x$$

**REMINDER**

Remember, it is always important to plot a scatter diagram first. If the scatter plot indicates that there is a linear relationship between the variables, then it is reasonable to use a best fit line to make predictions for  $y$  given  $x$  within the domain of  $x$ -values in the sample data, **but not necessarily for  $x$ -values outside that domain**. You could use the line to predict the final exam score for a student who earned a grade of 73 on the third exam. You should NOT use the line to predict the final exam score for a student who earned a grade of 50 on the third exam, because 50 is not within the domain of the  $x$ -values in the sample data, which are between 65 and 75.

**UNDERSTANDING SLOPE**

The slope of the line,  $b$ , describes how changes in the variables are related. It is important to interpret the slope of the line in the context of the situation represented by the data. You should be able to write a sentence interpreting the slope in plain English.

**INTERPRETATION OF THE SLOPE:** The slope of the best-fit line tells us how the dependent variable ( $y$ ) changes for every one unit increase in the independent ( $x$ ) variable, on average.

**THIRD EXAM vs FINAL EXAM EXAMPLE**

Slope: The slope of the line is  $b = 4.83$ .

Interpretation: For a one-point increase in the score on the third exam, the final exam score increases by 4.83 points, on average.

**USING THE TI-83, 83+, 84, 84+ CALCULATOR**

Using the Linear Regression T Test: LinRegTTest

1. In the STAT list editor, enter the X data in list L1 and the Y data in list L2, paired so that the corresponding  $(x,y)$  values are next to each other in the lists. (If a particular pair of values is repeated, enter it as many times as it appears in the data.)
2. On the STAT TESTS menu, scroll down with the cursor to select the LinRegTTest. (Be careful to select LinRegTTest, as some calculators may also have a different item called LinRegTInt.)
3. On the LinRegTTest input screen enter: Xlist: L1 ; Ylist: L2 ; Freq: 1
4. On the next line, at the prompt  $\beta$  or  $\rho$ , highlight " $\neq 0$ " and press ENTER
5. Leave the line for "RegEq:" blank

## 6. Highlight Calculate and press ENTER.

## LinRegTTest Input Screen and Output Screen

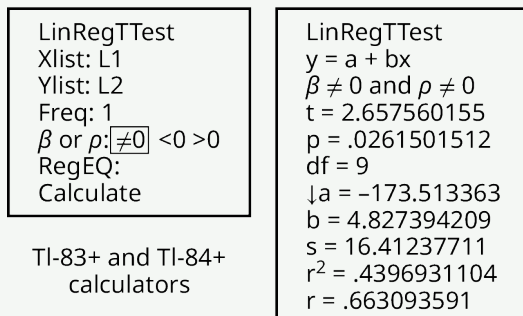
TI-83+ and TI-84+  
calculators

Figure 12.14

The output screen contains a lot of information. For now we will focus on a few items from the output, and will return later to the other items.

The second line says  $y = a + bx$ . Scroll down to find the values  $a = -173.513$ , and  $b = 4.8273$ ; the equation of the best fit line is  $\hat{y} = -173.51 + 4.83x$

The two items at the bottom are  $r^2 = 0.43969$  and  $r = 0.663$ . For now, just note where to find these values; we will discuss them in the next two sections.

## Graphing the Scatterplot and Regression Line

1. We are assuming your X data is already entered in list L1 and your Y data is in list L2
2. Press 2nd STATPLOT ENTER to use Plot 1
3. On the input screen for PLOT 1, highlight **On**, and press ENTER
4. For TYPE: highlight the very first icon which is the scatterplot and press ENTER
5. Indicate Xlist: L1 and Ylist: L2
6. For Mark: it does not matter which symbol you highlight.
7. Press the ZOOM key and then the number 9 (for menu item "ZoomStat"); the calculator will fit the window to the data
8. To graph the best-fit line, press the "Y=" key and type the equation  $-173.5 + 4.83X$  into equation Y1. (The X key is immediately left of the STAT key). Press ZOOM 9 again to graph it.
9. Optional: If you want to change the viewing window, press the WINDOW key. Enter your desired window using Xmin, Xmax, Ymin, Ymax

## NOTE

Another way to graph the line after you create a scatter plot is to use LinRegTTest.

1. Make sure you have done the scatter plot. Check it on your screen.
2. Go to LinRegTTest and enter the lists.
3. At RegEq: press VARS and arrow over to Y-VARS. Press 1 for 1:Function. Press 1 for 1:Y1. Then arrow down to Calculate and do the calculation for the line of best fit.
4. Press Y = (you will see the regression equation).
5. Press GRAPH. The line will be drawn."

The Correlation Coefficient  $r$ 

Besides looking at the scatter plot and seeing that a line seems reasonable, how can you tell if the line is a good predictor? Use the correlation coefficient as another indicator (besides the scatterplot) of the strength of the relationship between  $x$  and  $y$ .

The **correlation coefficient**,  $r$ , developed by Karl Pearson in the early 1900s, is numerical and provides a measure of strength and direction of the linear association between the independent variable  $x$  and the dependent variable  $y$ .

The correlation coefficient is calculated as

$$r = \frac{n\Sigma(xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

where  $n$  = the number of data points.

If you suspect a linear relationship between  $x$  and  $y$ , then  $r$  can measure how strong the linear relationship is.

#### What the VALUE of $r$ tells us:

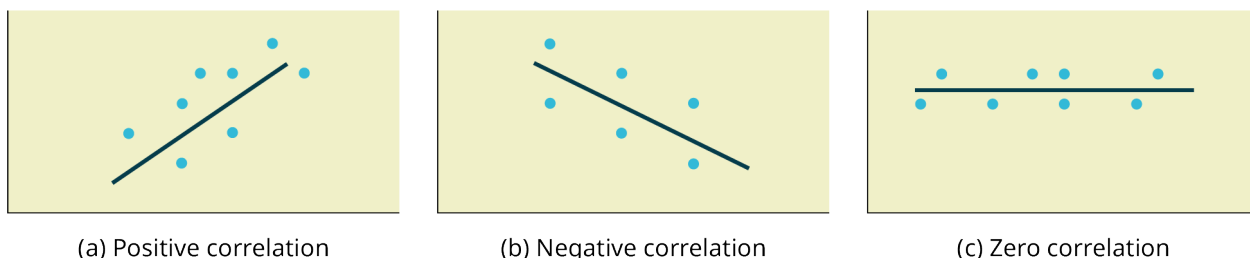
- The value of  $r$  is always between  $-1$  and  $+1$ :  $-1 \leq r \leq 1$ .
- The size of the correlation  $r$  indicates the strength of the linear relationship between  $x$  and  $y$ . Values of  $r$  close to  $-1$  or to  $+1$  indicate a stronger linear relationship between  $x$  and  $y$ .
- If  $r = 0$  there is likely no linear correlation. It is important to view the scatterplot, however, because data that exhibit a curved or horizontal pattern may have a correlation of 0.
- If  $r = 1$ , there is perfect positive correlation. If  $r = -1$ , there is perfect negative correlation. In both these cases, all of the original data points lie on a straight line. Of course, in the real world, this will not generally happen.

#### What the SIGN of $r$ tells us

- A positive value of  $r$  means that when  $x$  increases,  $y$  tends to increase and when  $x$  decreases,  $y$  tends to decrease (**positive correlation**).
- A negative value of  $r$  means that when  $x$  increases,  $y$  tends to decrease and when  $x$  decreases,  $y$  tends to increase (**negative correlation**).
- The sign of  $r$  is the same as the sign of the slope,  $b$ , of the best-fit line.

#### NOTE

Strong correlation does not suggest that  $x$  causes  $y$  or  $y$  causes  $x$ . We say "**correlation does not imply causation.**"



**Figure 12.15** (a) A scatter plot showing data with a positive correlation.  $0 < r < 1$  (b) A scatter plot showing data with a negative correlation.  $-1 < r < 0$  (c) A scatter plot showing data with zero correlation.  $r = 0$

The formula for  $r$  looks formidable. However, computer spreadsheets, statistical software, and many calculators can quickly calculate  $r$ . The correlation coefficient  $r$  is the bottom item in the output screens for the LinRegTTest on the TI-83, TI-83+, or TI-84+ calculator (see previous section for instructions).

## The Coefficient of Determination

The variable  $r^2$  is called the **coefficient of determination** and is the square of the correlation coefficient, but is usually stated as a percent, rather than in decimal form. It has an interpretation in the context of the data:

- $r^2$ , when expressed as a percent, represents the percent of variation in the dependent (predicted) variable  $y$  that can be explained by variation in the independent (explanatory) variable  $x$  using the regression (best-fit) line.
- $1 - r^2$ , when expressed as a percentage, represents the percent of variation in  $y$  that is NOT explained by variation in  $x$  using the regression line. This can be seen as the scattering of the observed data points about the regression line.

Consider the [third exam/final exam example](#) introduced in the previous section

- The line of best fit is:  $\hat{y} = -173.51 + 4.83x$
- The correlation coefficient is  $r = 0.6631$



- The coefficient of determination is  $r^2 = 0.6631^2 = 0.4397$
- **Interpretation of  $r^2$  in the context of this example:**
- Approximately 44% of the variation (0.4397 is approximately 0.44) in the final-exam grades can be explained by the variation in the grades on the third exam, using the best-fit regression line.
- Therefore, approximately 56% of the variation ( $1 - 0.44 = 0.56$ ) in the final exam grades can NOT be explained by the variation in the grades on the third exam, using the best-fit regression line. (This is seen as the scattering of the points about the line.)

## 12.4 Testing the Significance of the Correlation Coefficient

The correlation coefficient,  $r$ , tells us about the strength and direction of the linear relationship between  $x$  and  $y$ . However, the reliability of the linear model also depends on how many observed data points are in the sample. We need to look at both the value of the correlation coefficient  $r$  and the sample size  $n$ , together.

We perform a hypothesis test of the **"significance of the correlation coefficient"** to decide whether the linear relationship in the sample data is strong enough to use to model the relationship in the population.

The sample data are used to compute  $r$ , the correlation coefficient for the sample. If we had data for the entire population, we could find the population correlation coefficient. But because we have only sample data, we cannot calculate the population correlation coefficient. The sample correlation coefficient,  $r$ , is our estimate of the unknown population correlation coefficient.

The symbol for the population correlation coefficient is  $\rho$ , the Greek letter "rho."

$\rho$  = population correlation coefficient (unknown)

$r$  = sample correlation coefficient (known; calculated from sample data)

The hypothesis test lets us decide whether the value of the population correlation coefficient  $\rho$  is "close to zero" or "significantly different from zero". We decide this based on the sample correlation coefficient  $r$  and the sample size  $n$ .

**If the test concludes that the correlation coefficient is significantly different from zero, we say that the correlation coefficient is "significant."**

- Conclusion: There is sufficient evidence to conclude that there is a significant linear relationship between  $x$  and  $y$  because the correlation coefficient is significantly different from zero.
- What the conclusion means: There is a significant linear relationship between  $x$  and  $y$ . We can use the regression line to model the linear relationship between  $x$  and  $y$  in the population.

**If the test concludes that the correlation coefficient is not significantly different from zero (it is close to zero), we say that correlation coefficient is "not significant".**

- Conclusion: "There is insufficient evidence to conclude that there is a significant linear relationship between  $x$  and  $y$  because the correlation coefficient is not significantly different from zero."
- What the conclusion means: There is not a significant linear relationship between  $x$  and  $y$ . Therefore, we CANNOT use the regression line to model a linear relationship between  $x$  and  $y$  in the population.

### NOTE

- If  $r$  is significant and the scatter plot shows a linear trend, the line can be used to predict the value of  $y$  for values of  $x$  that are within the domain of observed  $x$  values.
- If  $r$  is not significant OR if the scatter plot does not show a linear trend, the line should not be used for prediction.
- If  $r$  is significant and if the scatter plot shows a linear trend, the line may NOT be appropriate or reliable for prediction OUTSIDE the domain of observed  $x$  values in the data.

## PERFORMING THE HYPOTHESIS TEST

- **Null Hypothesis:**  $H_0: \rho = 0$
- **Alternate Hypothesis:**  $H_a: \rho \neq 0$

### WHAT THE HYPOTHESES MEAN IN WORDS:

- **Null Hypothesis  $H_0$ :** The population correlation coefficient IS NOT significantly different from zero. There IS NOT a significant linear relationship (correlation) between  $x$  and  $y$  in the population.
- **Alternate Hypothesis  $H_a$ :** The population correlation coefficient IS significantly DIFFERENT FROM zero. There IS A