3 Descriptive Analysis and Presentation of Bivariate Data

3.2 Linear Correlation

Linear Correlation

The primary purpose of **linear correlation analysis** is to measure the strength of a linear relationship between two variables.

Let's examine some scatter diagrams that demonstrate different relationships between input, or independent variables, x, and output, or dependent variables, y.

If as x increases there is no definite shift in the values of y, we say there is **no correlation**, or no relationship between x and y.

Linear Correlation

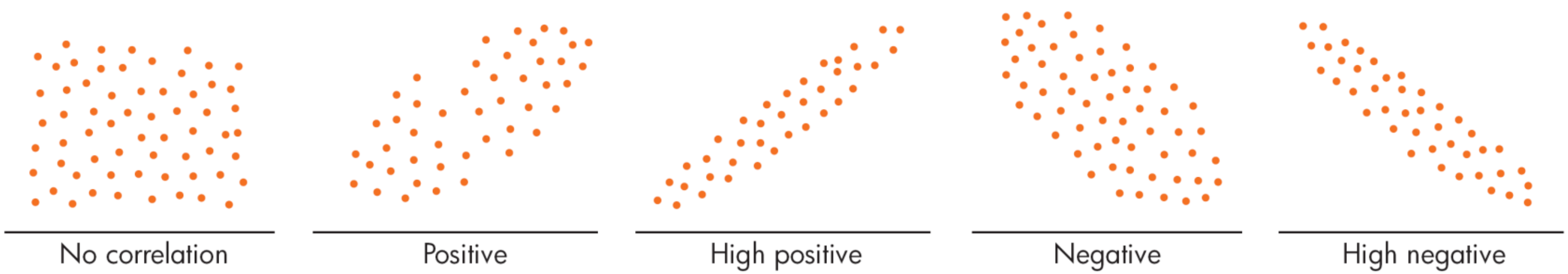
If as x increases there is a shift in the values of y, then there is a **correlation**. The correlation is **positive** when y tends to increase and **negative** when y tends to decrease.

If the ordered pairs (x, y) tend to follow a straight-line path, there is a linear correlation.

The preciseness of the shift in y as x increases determines the strength of the **linear correlation**.

Linear Correlation

The scatter diagrams in Figure 3.9 demonstrate these ideas.

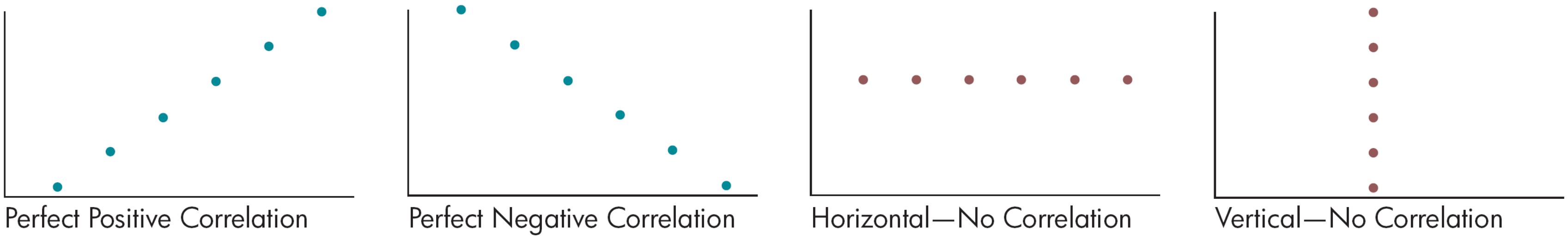


Scatter Diagrams and Correlation

Figure 3.9

Linear Correlation

Perfect linear correlation occurs when all the points fall exactly along a straight line, as shown in Figure 3.10.



Ordered Pairs Forming a Straight Line

Figure 3.10

The correlation can be either positive or negative, depending on whether *y* increases or decreases as *x* increases.

Linear Correlation

If the data form a straight horizontal or vertical line, there is no correlation, because one variable has no effect on the other, as also shown in Figure 3.7.

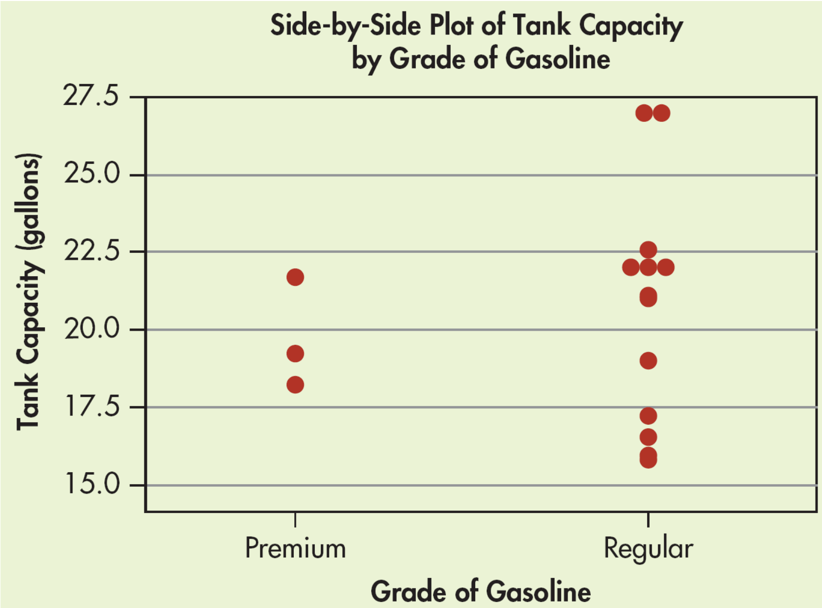
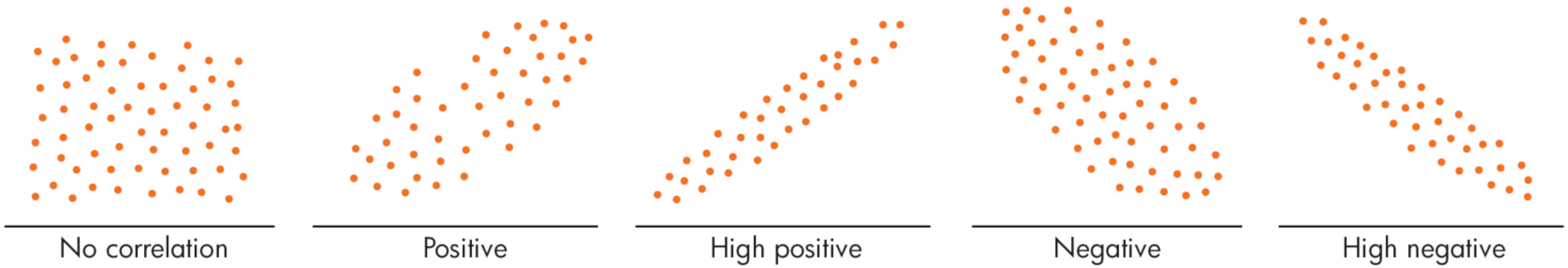


Figure 3.7

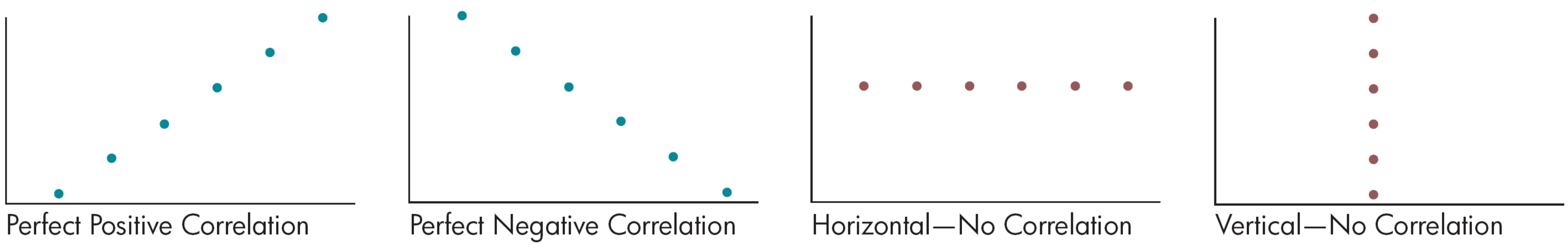
Linear Correlation

Scatter diagrams do not always appear in one of the forms shown in Figures 3.9 and 3.10.



Scatter Diagrams and Correlation

Figure 3.9

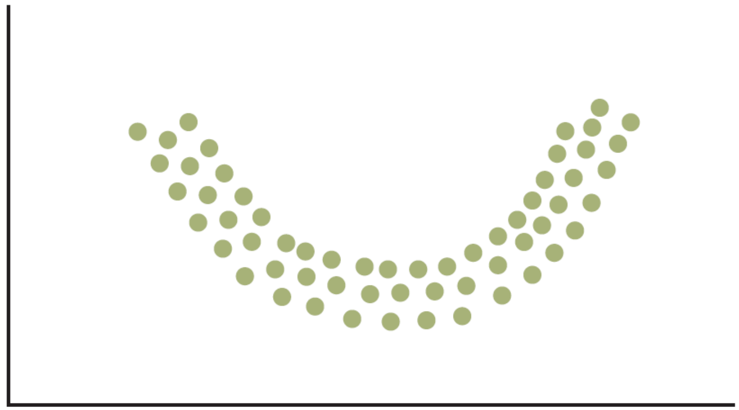


Ordered Pairs Forming a Straight Line

Figure 3.10

Linear Correlation

Sometimes they suggest relationships other than linear, as in Figure 3.11.There appears to be a definite pattern; however, the two variables are not related linearly, and therefore there is no linear correlation.



No Linear Correlation

Figure 3.11

Linear Correlation

The **coefficient of linear correlation**, ***r***, is the numerical measure of the strength of the linear relationship between two variables.

The coefficient reflects the consistency of the effect that a change in one variable has on the other.

The value of the linear correlation coefficient helps us answer the question: Is there a linear correlation between the two variables under consideration?

Linear Correlation

The linear correlation coefficient, *r*, always has a value between –1 and +1.

A value of +1 signifies a perfect positive correlation, and a value of –1 signifies a perfect negative correlation.

If as *x* increases there is a general increase in the value of *y*, then *r* will be positive in value.

For example, a positive value of *r* would be expected for the age and height of children because as children grow older, they grow taller.

Linear Correlation

Also, consider the age, *x*, and resale value, *y*, of an automobile. As the car ages, its resale value decreases. Since as *x* increases, *y* decreases, the relationship results in a negative value for *r*.

The value of *r* is defined by **Pearson's product moment formula:**

Definition Formula

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{(n - 1)s_x s_y} \qquad (3.1)$$

Linear Correlation

Notes

s_x and s_y are the standard deviations of the *x*- and *y*-variables.

To calculate *r*, we will use an alternative formula, formula (3.2), that is equivalent to formula (3.1).

Computational Formula

$$\text{linear correlation coefficient} = \frac{\text{sum of squares for } xy}{\sqrt{(\text{sum of squares for } x)(\text{sum of squares for } y)}}$$

$$r = \frac{SS(xy)}{\sqrt{SS(x)SS(y)}} \qquad (3.2)$$

Linear Correlation

As preliminary calculations, we will separately calculate three sums of squares and then substitute them into formula (3.2) to obtain *r*.

$$\textit{sum of squares for } x = \textit{sum of } x^2 - \frac{(\textit{sum of } x)^2}{n}$$

$$\mathrm{SS}(x) = \sum x^2 - \frac{\left(\sum x\right)^2}{n} \qquad \textbf{(2.8)}$$

We can also calculate:

$$\textit{sum of squares for } y = \textit{sum of } y^2 - \frac{(\textit{sum of } y)^2}{n}$$

Linear Correlation

$$\text{SS}(y) = \sum y^2 - \frac{\left(\sum y\right)^2}{n} \qquad \textbf{(3.3)}$$

$$\textit{sum of squares for xy} = \textit{sum of xy} - \frac{(\textit{sum of x})(\textit{sum of y})}{n}$$

$$\text{SS}(xy) = \sum xy - \frac{\sum x \sum y}{n} \qquad \textbf{(3.4)}$$

Example 5 – *Calculating the Linear Correlation Coefficient, r*

In Mr. Chamberlain's physical-fitness course, several fitness scores were taken. The following sample is the numbers of push-ups and sit-ups done by 10 randomly selected students:

(27, 30) (22, 26) (15, 25) (35, 42) (30, 38)

(52, 40) (35, 32) (55, 54) (40, 50) (40, 43)

Example 5 – *Calculating the Linear Correlation Coefficient, r* cont'd

Table 3.10 shows these sample data, and Figure 3.5 shows a scatter diagram of the data.

Student	1	2	3	4	5	6	7	8	9	10
Push-ups, x	27	22	15	35	30	52	35	55	40	40
Sit-ups, y	30	26	25	42	38	40	32	54	50	43

Data for Push-ups and Sit-ups **[TA03-10]**

Table 3.10

Find the linear correlation coefficient for the push-up/sit-up data.

Example 5 – *Solution*

First, we construct an extensions table (Table 3.12) listing all the pairs of values (x, y) to aid us in finding x^2, xy, and y^2 for each pair and the five column totals.

Student	Push-ups, x	x^2	Sit-ups, y	y^2	xy
1	27	729	30	900	810
2	22	484	26	676	572
3	15	225	25	625	375
4	35	1,225	42	1,764	1,470
5	30	900	38	1,444	1,140
6	52	2,704	40	1,600	2,080
7	35	1,225	32	1,024	1,120
8	55	3,025	54	2,916	2,970
9	40	1,600	50	2,500	2,000
10	40	1,600	43	1,849	1,720
	$\sum x = 351$ sum of x	$\sum x^2 = 13{,}717$ sum of x^2	$\sum y = 380$ sum of y	$\sum y^2 = 15{,}298$ sum of y^2	$\sum xy = 14{,}257$ sum of xy

Extensions Table for Finding Five Summations **[TA03-10]**

Table 3.12

Example 5 – *Solution*

cont'd

Second, to complete the preliminary calculations, we substitute the five summations (the five column totals) from the extensions table into formulas (2.8), (3.3), and (3.4), and calculate the three sums of squares:

$$SS(x) = \sum x^2 - \frac{(\sum x)^2}{n} = 13{,}717 - \frac{(351)^2}{10} = 1396.9$$

$$SS(y) = \sum y^2 - \frac{(\sum y)^2}{n} = 15{,}298 - \frac{(380)^2}{10} = 858.0$$

$$SS(xy) = \sum xy - \frac{\sum x \sum y}{n} = 14{,}257 - \frac{(351)(380)}{10} = 919.00$$

Example 5 – *Solution*

cont'd

Third, we substitute the three sums of squares into formula (3.2) to find the value of the correlation coefficient:

$$r = \frac{SS(xy)}{\sqrt{SS(x)SS(y)}}$$

$$= \frac{919.0}{\sqrt{(1396.9)(858.0)}}$$

$$= 0.8394$$

$$= \mathbf{0.84}$$

Linear Correlation

The value of the linear correlation coefficient helps us answer the question: Is there a linear correlation between the two variables under consideration?

When the calculated value of r is close to zero, we conclude that there is little or no linear correlation. As the calculated value of r changes from 0.0 toward either +1.0 or –1.0, it indicates an increasing linear correlation between the two variables.

Linear Correlation

From a graphic viewpoint, when we calculate *r*, we are measuring how well a straight line describes the scatter diagram of ordered pairs.

As the value of *r* changes from 0.0 toward +1.0 or –1.0, the data points create a pattern that moves closer to a straight line.

Understanding the Linear Correlation Coefficient

Understanding the Linear Correlation Coefficient

The following method will create (1) a visual meaning for correlation, (2) a visual meaning for what the linear coefficient is measuring, and (3) an estimate for *r*.

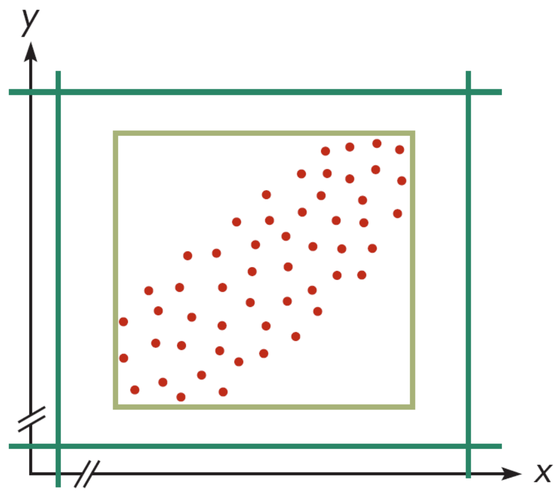
The method is quick and generally yields a reasonable estimate when the “window of data” is approximately square.

Note
This estimation technique does not replace the calculation of *r*. It is very sensitive to the “spread” of the diagram. However, if the “window of data” is approximately square, this approximation will be useful as a mental estimate or check.

Understanding the Linear Correlation Coefficient

Procedure

1. Construct a scatter diagram of your data, being sure to scale the axes so that the resulting graph has an approximately square "window of data," as demonstrated in Figure 3.12 by the light green frame.



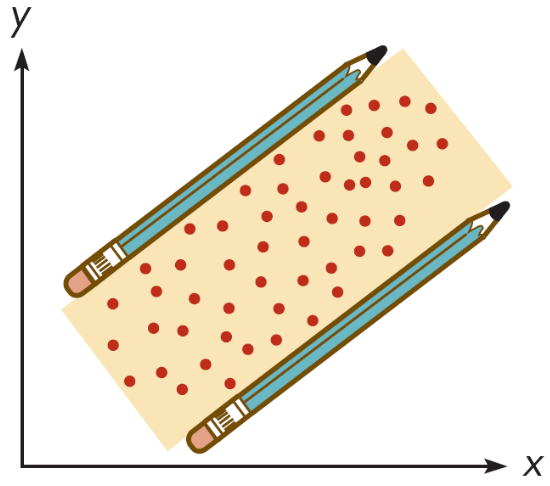
The Data Window

Figure 3.12

The window may not be the same region as determined by the bounds of the two scales, shown as a green rectangle on Figure 3.12.

Understanding the Linear Correlation Coefficient

2. Lay two pencils on your scatter diagram. Keeping them parallel, move them to a position so that they are as close together as possible while having all the points on the scatter diagram between them. (See Figure 3.13.)



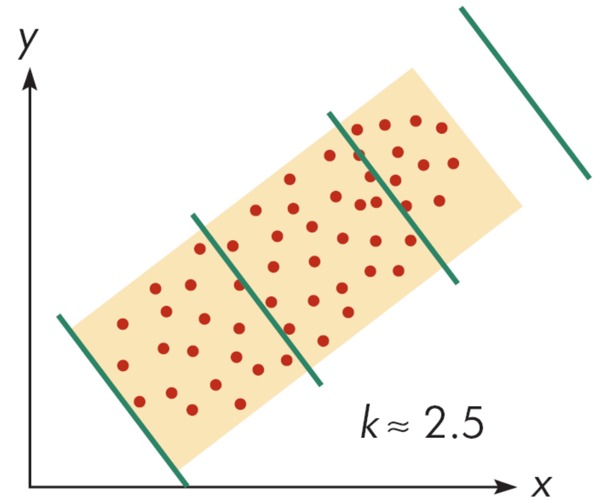
Focusing on Pattern

Figure 3.13

Understanding the Linear Correlation Coefficient

3. Visualize a rectangular region that is bounded by the two pencils and that ends just beyond the points on the scatter diagram. (See the shaded portion of Figure 3.13.)

4. Estimate the number of times longer the rectangle is than it is wide. An easy way to do this is to mentally mark off squares in the rectangle. (See Figure 3.14.) Call this number of multiples *k*.



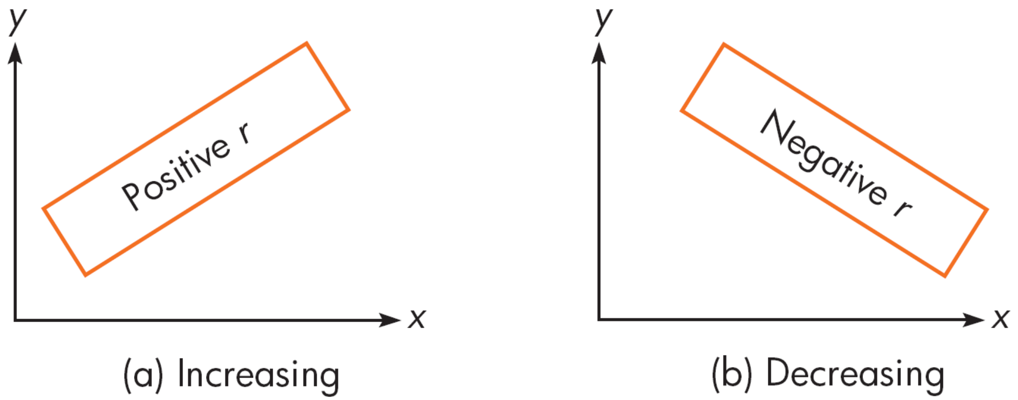
Finding *k*

Figure 3.14

Understanding the Linear Correlation Coefficient

5. The value of *r* may be estimated as $\pm\left(1 - \frac{1}{k}\right)$.

6. The sign assigned to *r* is determined by the general position of the length of the rectangular region. If it lies in an increasing position, *r* will be positive; if it lies in a decreasing position, *r* will be negative (see Figure 3.15).



(a) Increasing Position; (b) Decreasing Position

Figure 3.15

Understanding the Linear Correlation Coefficient

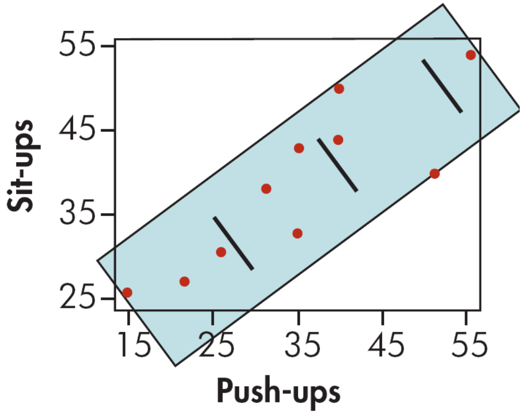
If the rectangle is in either a horizontal or a vertical position, then *r* will be zero, regardless of the length–width ratio.

Let's use this method to estimate the value of the linear correlation coefficient for the relationship between the number of push-ups and sit-ups.

Understanding the Linear Correlation Coefficient

As shown in Figure 3.16, we find that the rectangle is approximately 3.5 times longer than it is wide—that is, $k \approx 3.5$—and the rectangle lies in an increasing position. Therefore, our estimate for r is

$$r \approx +\left(1 - \frac{1}{3.5}\right) \approx +0.70$$



Push-ups versus Sit-ups for 10 Students

Figure 3.16

Causation and Lurking Variables

Causation and Lurking Variables

As we try to explain the past, understand the present, and estimate the future, judgments about cause and effect are necessary because of our desire to impose order on our environment.

The **cause-and-effect relationship** is fairly straightforward. You may focus on a situation, the *effect* (e.g., a disease or social problem), and try to determine its *cause(s),* or you may begin with a *cause* (unsanitary conditions or poverty) and discuss its *effect(s).*

Causation and Lurking Variables

To determine the cause of something, ask yourself why it happened. To determine the effect, ask yourself **what** happened.

Lurking variable A variable that is not included in a study but has an effect on the variables of the study and makes it appear that those variables are related.

A good example is the strong positive relationship shown between the amount of damage caused by a fire and the number of firefighters who work the fire.

Causation and Lurking Variables

The "size" of the fire is the lurking variable; it "causes" both the "amount" of damage and the "number" of firefighters.

If there is a strong linear correlation between two variables, then one of the following situations may be true about the relationship between the two variables:

1. There is a direct cause-and-effect relationship between the two variables.

2. There is a reverse cause-and-effect relationship between the two variables.

Causation and Lurking Variables

3. Their relationship may be caused by a third variable.

4. Their relationship may be caused by the interactions of several other variables.

5. The apparent relationship may be strictly a coincidence.

Causation and Lurking Variables

Remember that a strong correlation does not necessarily imply causation.

Here are some pitfalls to avoid:

1. In a direct cause-and-effect relationship, an increase (or decrease) in one variable causes an increase (or decrease) in another. Suppose there is a strong positive correlation between weight and height.

   Does an increase in weight *cause* an increase in height? Not necessarily. Or to put it another way, does a decrease in weight *cause* a decrease in height?

Causation and Lurking Variables

Many other possible variables are involved, such as gender, age, and body type. These other variables are called *lurking variables*.

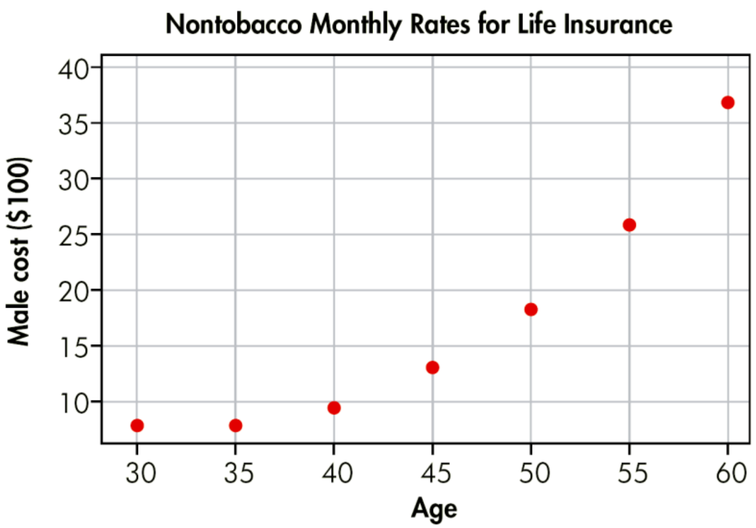
2. Don't reason from *correlation* to *cause*: Just because all people who move to the city get old doesn't mean that the city *causes* aging. The city may be a factor, but you can't base your argument on the correlation.

3. Don't reason from *correlation* to *cause:* Just because all people who move to the city get old doesn't mean that the city *causes* aging. The city may be a factor, but you can't base your argument on the correlation.

Applied Example 6 – *Life Insurance Rates*

Does a high linear correlation coefficient, *r*, imply that the data are linear in nature?

The issue age of the insured and the monthly life insurance rate for non-tobacco users appears highly correlated looking at the chart presented here.



As the issue age increases, the monthly rate for insurance increases for each of the genders.

Applied Example 6 – *Life Insurance Rates* cont’d

	$100,000		$250,000		$500,000	
Issue Age	Male ($)	Female ($)	Male ($)	Female ($)	Male ($)	Female ($)
30	7.96	6.59	11.96	9.13	19.25	12.46
35	8.05	6.56	11.96	9.13	19.57	12.46
40	9.63	7.79	15.22	10.89	23.19	16.47
45	13.14	9.80	22.40	15.44	35.87	24.03
50	18.44	12.42	33.69	21.10	53.81	33.38
55	26.01	15.75	49.22	29.37	87.59	48.06
60	37.10	20.83	74.59	42.05	137.38	69.87

Nontobacco Monthly Rates for Life Insurance **[TA03-13]**

Table 3.13

Let’s consider the issue age of the insured and the male monthly rate for a $100,000 policy. The calculated correlation coefficient for this specific class of insurance results in a value of r = 0.932.

Applied Example 6 – *Life Insurance Rates* cont’d

Typically, a value of r this close to 1.0 would indicate a fairly strong straight-line relationship; but wait. Do we have a linear relationship? Only a scatter diagram can tell us that.

The scatter diagram clearly shows a non-straight-line pattern. Yet, the correlation coefficient was so high. It is the elongated pattern in the data that produces a calculated r so large.

The lesson from this example is that one should always begin with a scatter diagram when considering linear correlation. The correlation coefficient tells only one side of the story!