# CHAPTER 8

## INTRODUCTION TO LINEAR REGRESSION

# Associations Between Variables

Many interesting examples of the use of statistics involve relationships between pairs of variables.

Two variables measured on the same cases are **associated** if knowing the value of one of the variables tells you something that you would not otherwise know about the value of the other variable.

# Relations between two variables

- A **response variable (y)** measures an outcome of a study.

  The response variable represent the dependent variable

- An **explanatory variable (x)** explains or causes changes in the response variable.

  The explanatory variable represent the independent variable.

# Examples

Which variable is the explanatory variable and which is the response variable?

1. Researchers are interested in discovering whether there is a relationship between the engine horsepower and the top speed that the car attains.

2. Whether or not a person likes to sing and whether or not a person likes to dance.

3. The number of pages in a textbook and the cost of a new copy of the textbook.

4. The number of alcoholic drinks consumed and the blood alcohol content.

# Examples

True or False:

1. In a study to determine whether surgery or chemotherapy results in higher survival rates for a certain type of cancer, whether or not the patient survived is one variable, and whether they received surgery or chemotherapy is the other. Treatment type is the explanatory variable here.

2. We wish to find the relationship between room temperature and exam scores for stats students. One good way to do this would be to measure room temperatures one quarter, and exam scores in the next quarter.

# Scatterplot

The most useful graph for displaying the relationship between two **quantitative** variables is a **scatterplot.**

A **scatterplot:**

- Shows the relationship between two quantitative variables measured on the same individuals.
- The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis.
- Each individual corresponds to one point on the graph.

# Scatterplot

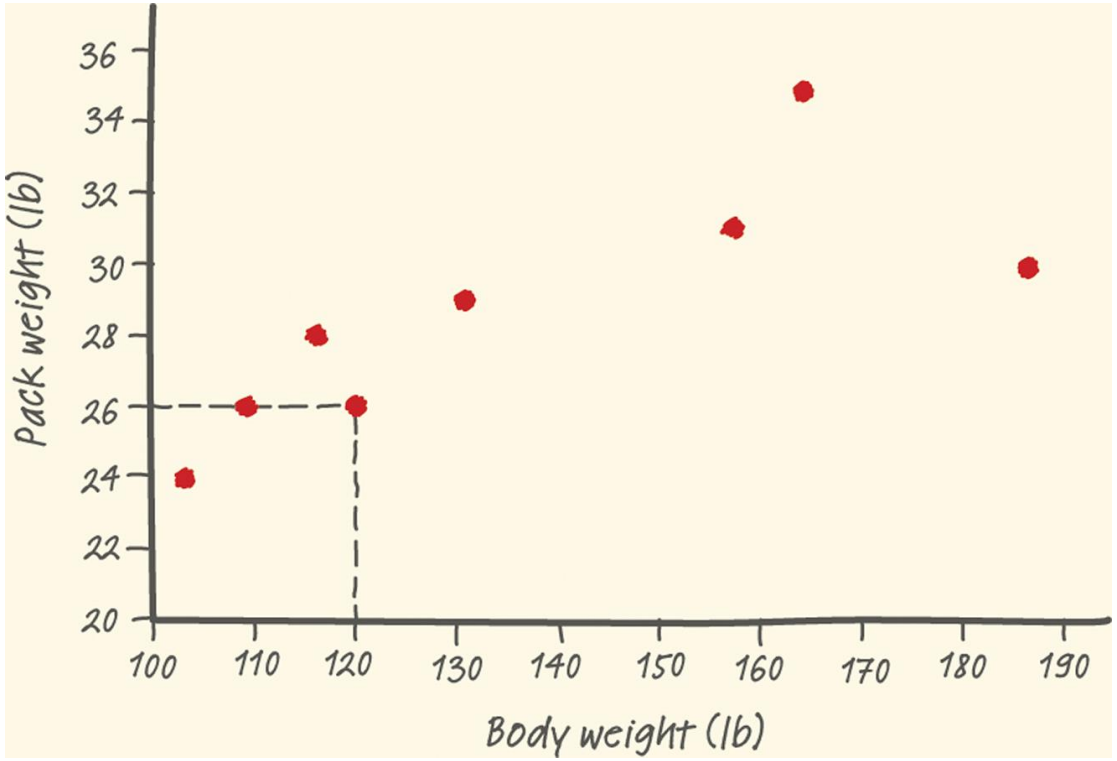
**How to Make a Scatterplot**

1. Decide which variable should go on each axis. If a distinction exists, plot the explanatory variable on the *x* axis and the response variable on the *y* axis.
2. Label and scale your axes.
3. Plot individual data values.

# Example

Make a scatterplot of the relationship between body weight and backpack weight for a group of hikers.

| Body weight (lb) | 120 | 187 | 109 | 103 | 131 | 165 | 158 | 116 |
|---|---|---|---|---|---|---|---|---|
| Backpack weight (lb) | 26 | 30 | 26 | 24 | 29 | 35 | 31 | 28 |



# Interpreting Scatterplots

**How to Examine a Scatterplot**

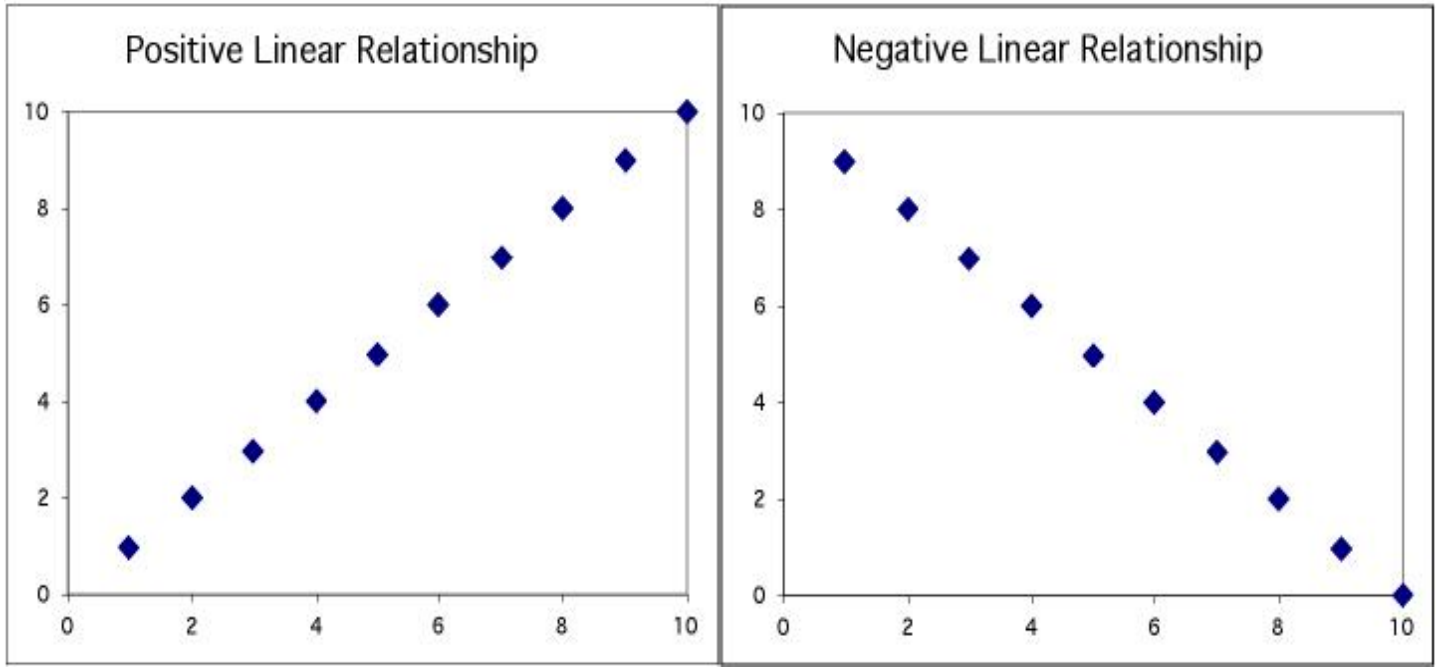
As in any graph of data, look for the *overall pattern* and for striking *deviations* from that pattern.

- You can describe the overall pattern of a scatterplot by the **direction, form,** and **strength** of the relationship.
- An important kind of departure is an **outlier,** an individual value that falls outside the overall pattern of the relationship.

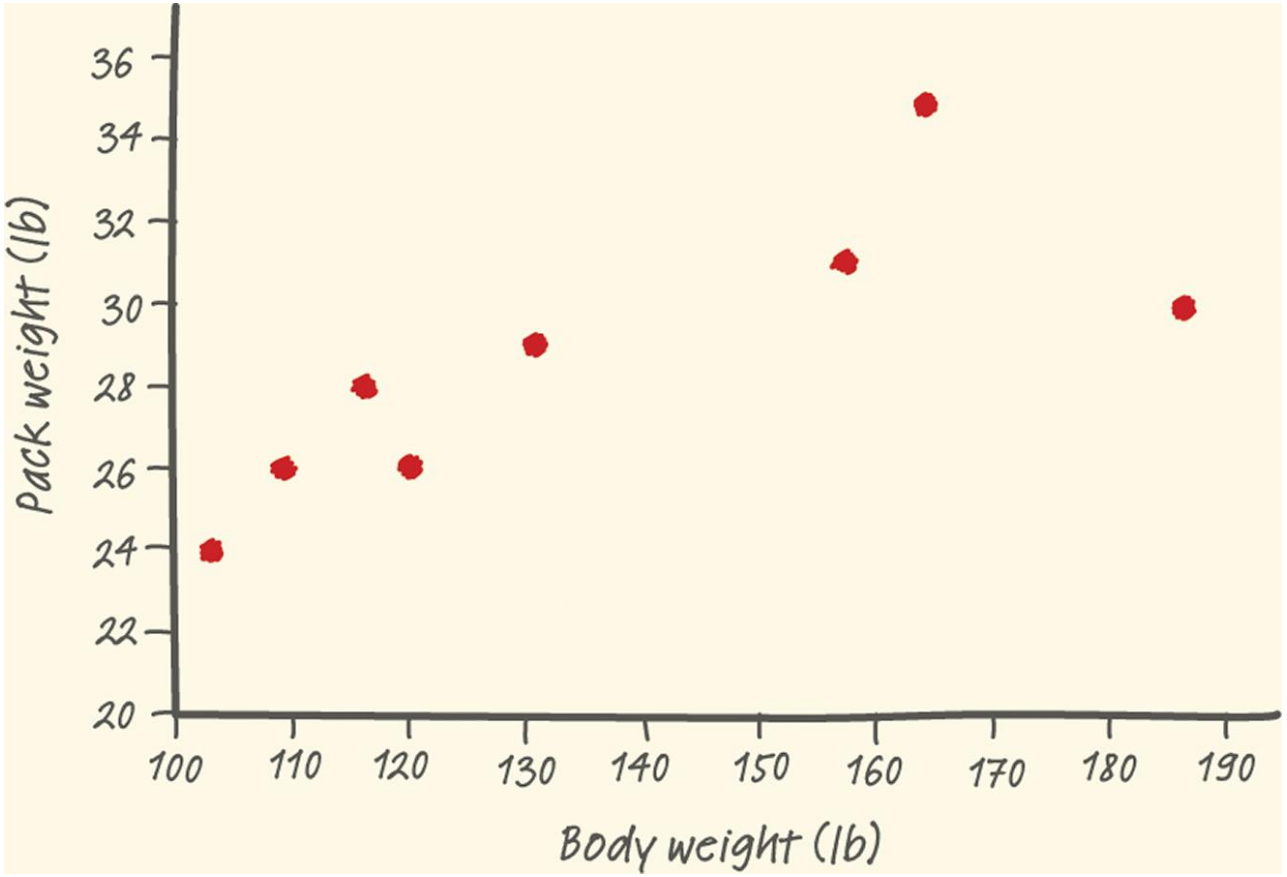
# Interpreting Scatterplots

- Two variables are **positively associated** when above-average values of one tend to accompany above-average values of the other, and when below-average values also tend to occur together.

- Two variables are **negatively associated** when above-average values of one tend to accompany below-average values of the other, and vice-versa.



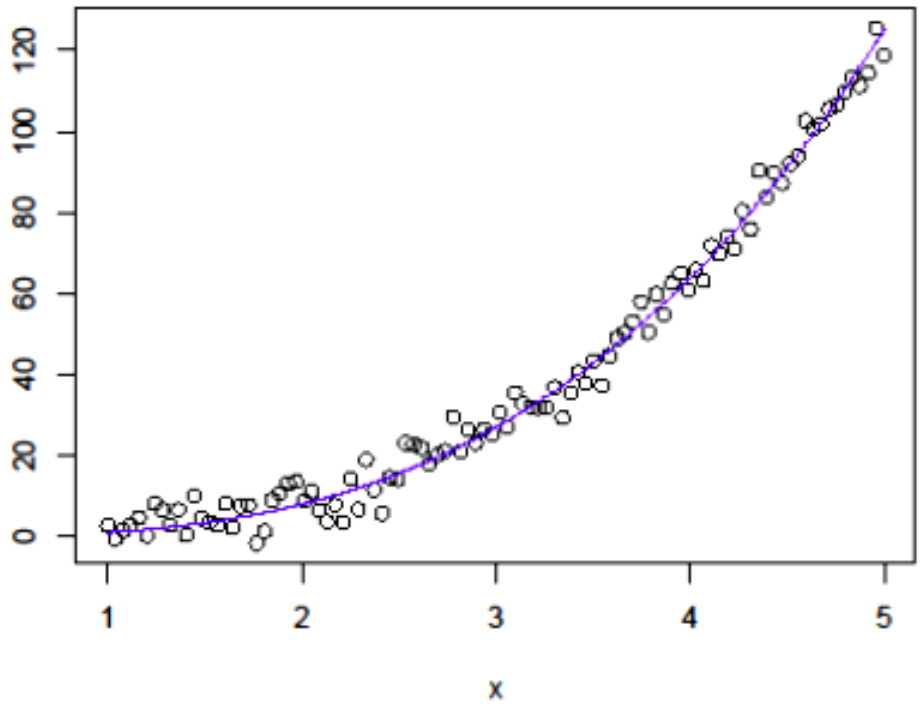
# Interpreting Scatterplots



# Nonlinear Relationships

- There are other forms of relationships besides linear. The scatterplot below is an example of a nonlinear form.

- Note that there is curvature in the relationship between *x* and *y*.



# Cautions with scatterplots:

- The relationship and fit apply to the data that are analyzed.
- We cannot assume that the relationship extends beyond the range of the data.
- Always ask yourself if the relationship that you see makes sense.

# Examples

1. Creating a scatterplot requires two _____ variables.

2. A researcher wants to know if taking increasing amounts of ginkgo biloba will result in increased capacities of memory ability for different students. They administer it to the students in doses of 250 milligrams, 500 milligrams, and 1000 milligrams. The amount of ginkgo will be plotted on the ____ axis.

3. Are the following two variables are positively associated, negatively associated or not associated? How much gas you have in your tank (in gallon) and the cost to fill your tank.

# Measuring Linear Association

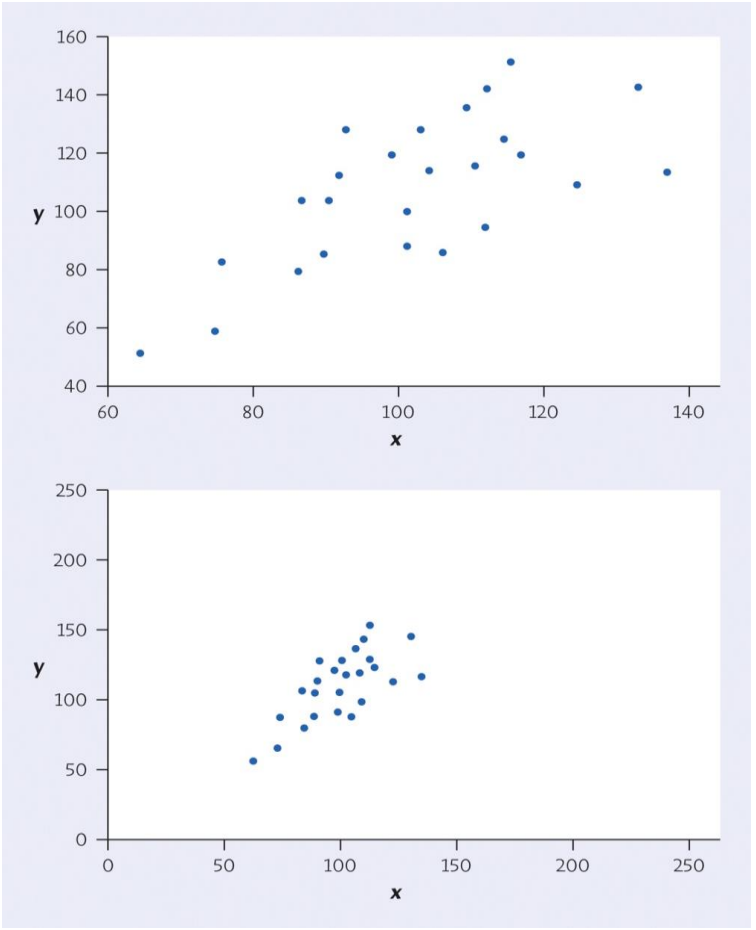
A scatterplot displays the strength, direction, and form of the relationship between two quantitative variables.

Linear (straight-line) relationships:
- Simple pattern
- Strong if the points lie close to a straight line
- Weak if they are widely scattered about a line

Our eyes are not a good judge of how strong a linear relationship is.

Note: Same data – lower plot in larger field

Hence, we need a numerical measure to supplement the graph



# Measuring Linear Association

- The **correlation (*r*)** measures the **direction** and **strength** of the **linear** relationship between two quantitative variables.

- Suppose that we have data on variables x and y for n individuals. The correlation r between x and y is

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

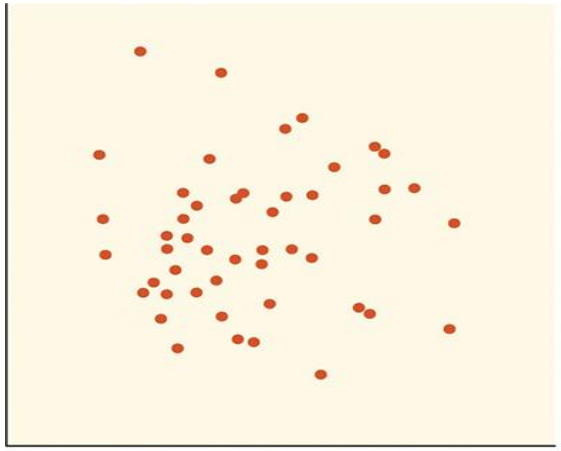
# Measuring Linear Association

We say a linear relationship is strong if the points lie close to a straight line and weak if they are widely scattered about a line. The following facts about *r* help us further interpret the strength of the linear relationship.

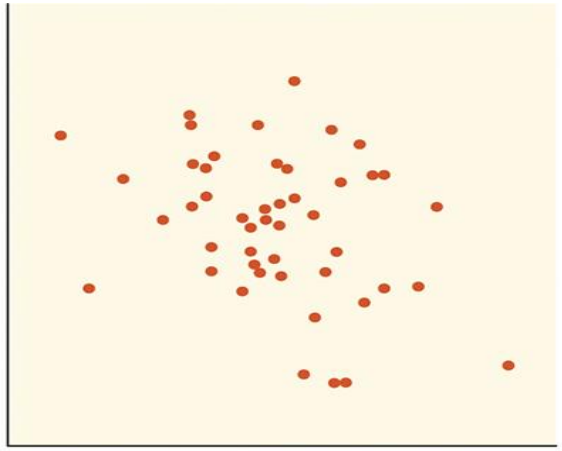
**Properties of Correlation**

- *r* is always a number between –1 and 1.
- $r > 0$ indicates a positive association.
- $r < 0$ indicates a negative association.
- Values of *r* near 0 indicate a very weak linear relationship.
- The strength of the linear relationship increases as *r* moves away from 0 toward –1 or 1.
- The extreme values $r = -1$ and $r = 1$ occur only in the case of a perfect linear relationship.

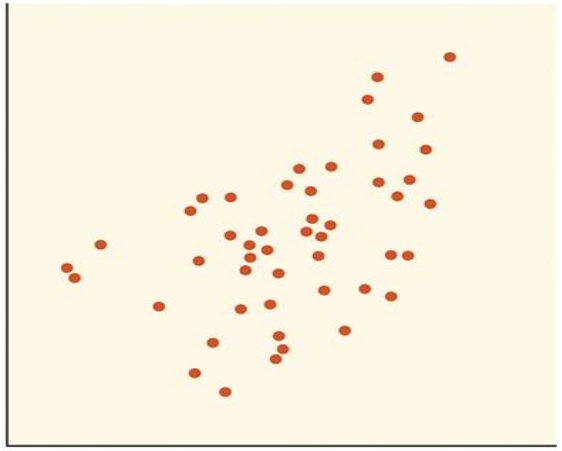
# Correlation



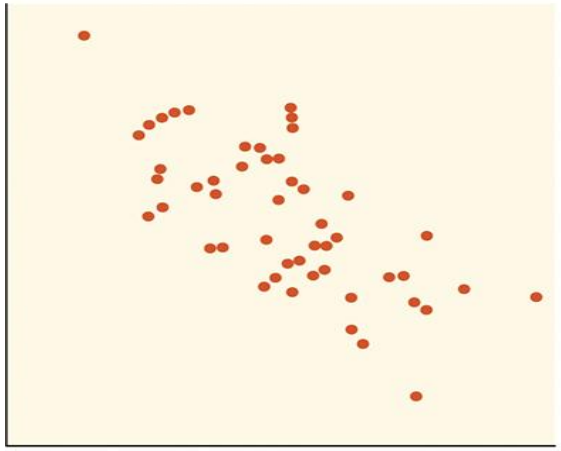
Correlation $r = 0$



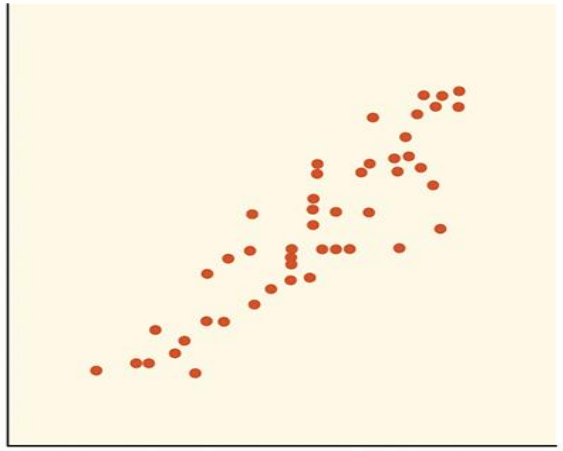
Correlation $r = -0.3$



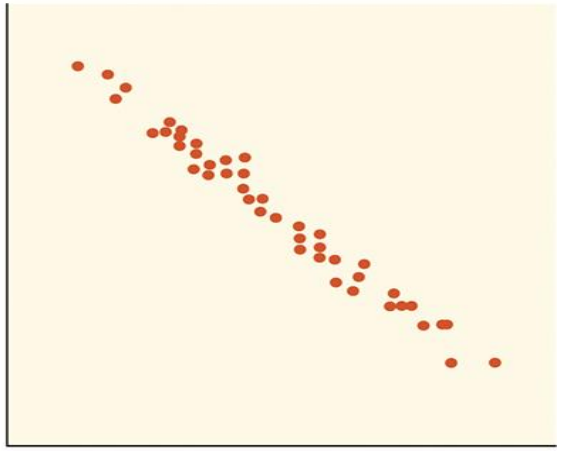
Correlation $r = 0.5$



Correlation $r = -0.7$



Correlation $r = 0.9$



Correlation $r = -0.99$

# Finding the correlation by hand

| x | 6 | 10 | 14 | 19 | 21 |
|---|---|---|---|---|---|
| y | 5 | 3 | 7 | 8 | 12 |

# Properties of Correlation

1. Correlation makes no distinction between explanatory and response variables.

2. $r$ has no units and does not change when we change the units of measurement of $x$, $y$, or both.

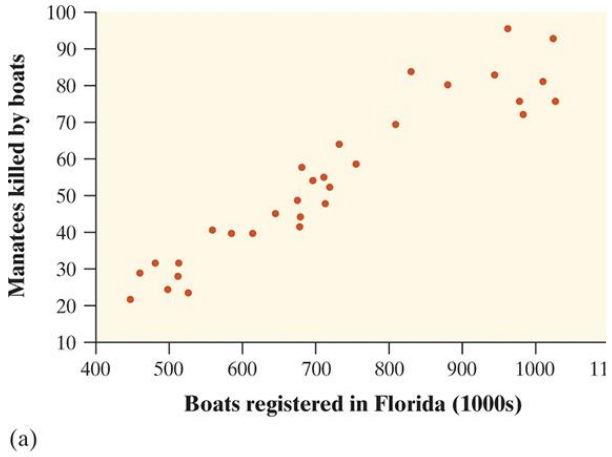
3. Positive $r$ indicates positive association between the variables, and negative $r$ indicates negative association.

4. The correlation $r$ is always a number between –1 and 1.

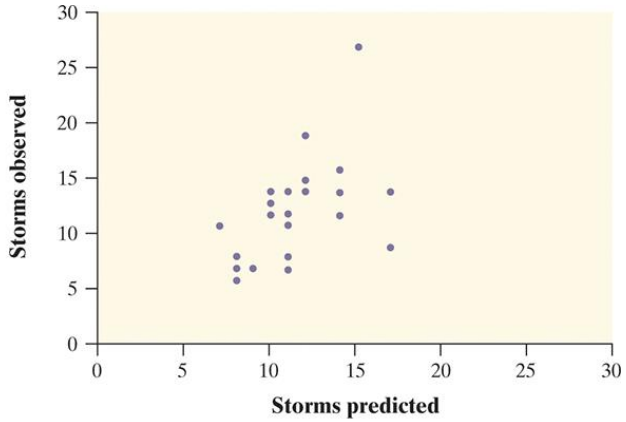
- **Cautions:**
- Correlation requires that both variables be quantitative.
- Correlation *does not describe curved relationships* between variables, no matter how strong the relationship is.
- The correlation *r* is not resistant; it can be strongly affected by a few outlying observations.
- Correlation is not a complete summary of two-variable data.

# Example

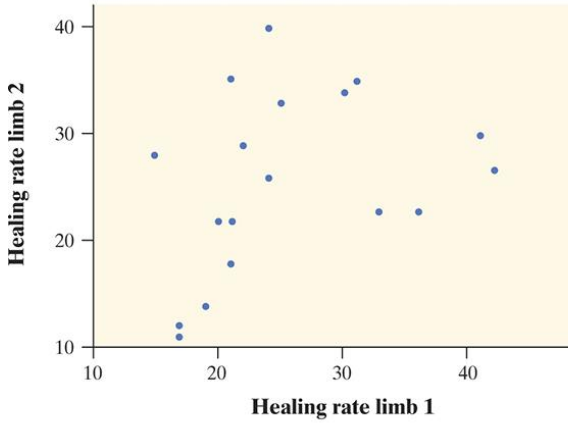
For each graph, estimate the correlation *r* and interpret it in context.



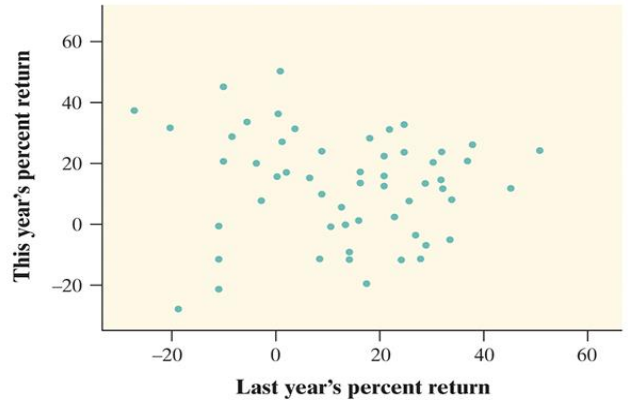

(a)




(b)




(c)




(d)

r = 0.358
r = 0.951
r = -0.081
r = 0.584

# Example

1. The correlation coefficient, r, measures the _____ and strength of a linear relationship.

2. Suppose we know the correlation between IQ and age in years is 0.12. If we change the units on age to months, the new correlation will be _____.

3. True or False:

Before computing the correlation coefficient r , we need to know the mean and standard deviation of both variables.