**Chapter Objectives**

By the end of this chapter, the student should be able to:

- Discuss basic ideas of linear regression and correlation.
- Create and interpret a line of best fit.
- Calculate and interpret the correlation coefficient.
- Calculate and interpret outliers.

Professionals often want to know how two or more numeric variables are related. For example, is there a relationship between the grade on the second math exam a student takes and the grade on the final exam? If there is a relationship, what is the relationship and how strong is it?

In another example, your income may be determined by your education, your profession, your years of experience, and your ability. The amount you pay a repair person for labor is often determined by an initial amount plus an hourly fee.
The type of data described in the examples is bivariate data — “bi” for two variables. In reality, statisticians use multivariate data, meaning many variables.

In this chapter, you will be studying the simplest form of regression, “linear regression” with one independent variable (x). This involves data that fits a line in two dimensions. You will also study correlation which measures how strong the relationship is.

12.1 | Linear Equations

Linear regression for two variables is based on a linear equation with one independent variable. The equation has the form:

\[ y = a + bx \]

where \( a \) and \( b \) are constant numbers.

The variable \( x \) is the independent variable, and \( y \) is the dependent variable. Typically, you choose a value to substitute for the independent variable and then solve for the dependent variable.

**Example 12.1**

The following examples are linear equations.

\[
\begin{align*}
y &= 3 + 2x \\
y &= -0.01 + 1.2x
\end{align*}
\]

**Try It**

12.1 Is the following an example of a linear equation?

\[ y = -0.125 - 3.5x \]

The graph of a linear equation of the form \( y = a + bx \) is a straight line. Any line that is not vertical can be described by this equation.
Example 12.2

Graph the equation $y = -1 + 2x$.

![Graph of the equation $y = -1 + 2x$](image)

**Figure 12.2**

Try It 🔍

**12.2** Is the following an example of a linear equation? Why or why not?

![Graph of a curve](image)

**Figure 12.3**

Example 12.3

Aaron’s Word Processing Service (AWPS) does word processing. The rate for services is $32 per hour plus a $31.50 one-time charge. The total cost to a customer depends on the number of hours it takes to complete the job.

Find the equation that expresses the **total cost** in terms of the **number of hours** required to complete the job.

**Solution 12.3**

Let $x$ = the number of hours it takes to get the job done.
Let $y$ = the total cost to the customer.

The $31.50 is a fixed cost. If it takes $x$ hours to complete the job, then $(32)(x)$ is the cost of the word processing only. The total cost is: $y = 31.50 + 32x$
12.3 Emma’s Extreme Sports hires hang-gliding instructors and pays them a fee of $50 per class as well as $20 per student in the class. The total cost Emma pays depends on the number of students in a class. Find the equation that expresses the total cost in terms of the number of students in a class.

Slope and Y-Intercept of a Linear Equation

For the linear equation \( y = a + bx \), \( b \) = slope and \( a \) = y-intercept. From algebra recall that the slope is a number that describes the steepness of a line, and the y-intercept is the \( y \) coordinate of the point \((0, a)\) where the line crosses the \( y \)-axis.

Figure 12.4 Three possible graphs of \( y = a + bx \). (a) If \( b > 0 \), the line slopes upward to the right. (b) If \( b = 0 \), the line is horizontal. (c) If \( b < 0 \), the line slopes downward to the right.

Example 12.4

Svetlana tutors to make extra money for college. For each tutoring session, she charges a one-time fee of $25 plus $15 per hour of tutoring. A linear equation that expresses the total amount of money Svetlana earns for each session she tutors is \( y = 25 + 15x \).

What are the independent and dependent variables? What is the y-intercept and what is the slope? Interpret them using complete sentences.

Solution 12.4

The independent variable \( (x) \) is the number of hours Svetlana tutors each session. The dependent variable \( (y) \) is the amount, in dollars, Svetlana earns for each session.

The y-intercept is 25 \( (a = 25) \). At the start of the tutoring session, Svetlana charges a one-time fee of $25 (this is when \( x = 0 \)). The slope is 15 \( (b = 15) \). For each session, Svetlana earns $15 for each hour she tutors.

Try It

12.4 Ethan repairs household appliances like dishwashers and refrigerators. For each visit, he charges $25 plus $20 per hour of work. A linear equation that expresses the total amount of money Ethan earns per visit is \( y = 25 + 20x \).

What are the independent and dependent variables? What is the y-intercept and what is the slope? Interpret them using complete sentences.

12.2 | Scatter Plots

Before we take up the discussion of linear regression and correlation, we need to examine a way to display the relation between two variables \( x \) and \( y \). The most common and easiest way is a scatter plot. The following example illustrates a scatter plot.
Example 12.5

In Europe and Asia, m-commerce is popular. M-commerce users have special mobile phones that work like electronic wallets as well as provide phone and Internet services. Users can do everything from paying for parking to buying a TV set or soda from a machine to banking to checking sports scores on the Internet. For the years 2000 through 2004, was there a relationship between the year and the number of m-commerce users? Construct a scatter plot. Let \( x \) = the year and let \( y \) = the number of m-commerce users, in millions.

<table>
<thead>
<tr>
<th>( x ) (year)</th>
<th>( y ) (# of users)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>0.5</td>
</tr>
<tr>
<td>2002</td>
<td>20.0</td>
</tr>
<tr>
<td>2003</td>
<td>33.0</td>
</tr>
<tr>
<td>2004</td>
<td>47.0</td>
</tr>
</tbody>
</table>

Table 12.1

(a) Table showing the number of m-commerce users (in millions) by year.

(b) Scatter plot showing the number of m-commerce users (in millions) by year.

To create a scatter plot:

1. Enter your X data into list L1 and your Y data into list L2.
2. Press 2nd STATPLOT ENTER to use Plot 1. On the input screen for PLOT 1, highlight On and press ENTER. (Make sure the other plots are OFF.)
3. For TYPE: highlight the very first icon, which is the scatter plot, and press ENTER.
4. For Xlist: enter L1 ENTER and for Ylist: L2 ENTER.
5. For Mark: it does not matter which symbol you highlight, but the square is the easiest to see. Press ENTER.
6. Make sure there are no other equations that could be plotted. Press Y = and clear any equations out.
7. Press the ZOOM key and then the number 9 (for menu item "ZoomStat") ; the calculator will fit the window to the data. You can press WINDOW to see the scaling of the axes.

12.5 Amelia plays basketball for her high school. She wants to improve to play at the college level. She notices that the number of points she scores in a game goes up in response to the number of hours she practices her jump shot each week. She records the following data:
Table 12.2

<table>
<thead>
<tr>
<th>X (hours practicing jump shot)</th>
<th>Y (points scored in a game)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>7</td>
<td>22</td>
</tr>
<tr>
<td>9</td>
<td>28</td>
</tr>
<tr>
<td>10</td>
<td>31</td>
</tr>
<tr>
<td>11</td>
<td>33</td>
</tr>
<tr>
<td>12</td>
<td>36</td>
</tr>
</tbody>
</table>

Construct a scatter plot and state if what Amelia thinks appears to be true.

A scatter plot shows the direction of a relationship between the variables. A clear direction happens when there is either:

- High values of one variable occurring with high values of the other variable or low values of one variable occurring with low values of the other variable.
- High values of one variable occurring with low values of the other variable.

You can determine the strength of the relationship by looking at the scatter plot and seeing how close the points are to a line, a power function, an exponential function, or to some other type of function. For a linear relationship there is an exception. Consider a scatter plot where all the points fall on a horizontal line providing a "perfect fit." The horizontal line would in fact show no relationship.

When you look at a scatterplot, you want to notice the overall pattern and any deviations from the pattern. The following scatterplot examples illustrate these concepts.

![Figure 12.6](image1)

(a) Positive linear pattern (strong)  (b) Linear pattern w/ one deviation

![Figure 12.7](image2)

(a) Negative linear pattern (strong)  (b) Negative linear pattern (weak)
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?