12

LINEAR REGRESSION AND CORRELATION

Figure 12.1 Linear regression and correlation can help you determine if an auto mechanic's salary is related to his work experience. (credit: modification of work "USPS commissions local repair-shop for some needed work on its older trucks" by Joshua Rothhaas/ Flickr, CC BY 2.0)

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.

Introduction

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:

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

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.

$$y = 3 + 2x$$
$$y = -0.01 + 1.2x$$

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 $y = -1 + 2x$.

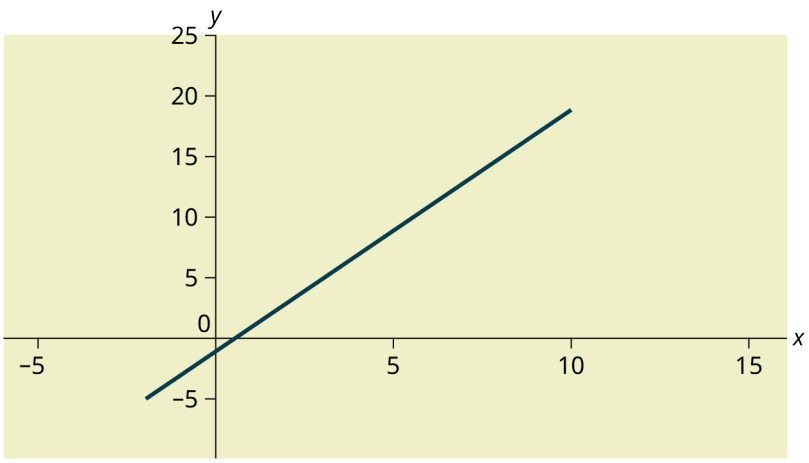

Figure 12.2

TRY IT 12.2

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

Figure 12.3

EXAMPLE 12.3

A local small business completes federal tax returns for customers. 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.

Problem

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

Solution

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 tax return, then (32)(x) is the cost of the tax return processing only. The total cost is: $y = 31.50 + 32x$

TRY IT 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.

(a) (b) (c)

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 $y = 25 + 15x$.

Problem

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

Solution

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

An educational researcher collects data on the vocabulary size of children as a function of age. The data is shown in Table 12.1. Is there a relationship between age and vocabulary size for young children? Construct a scatter plot. Let x = Child's Age, and let y = Vocabulary Size.

Age (years)	Vocabulary Size (number of words)
3	655
4	1098
6	2463
7	3195

Table 12.1