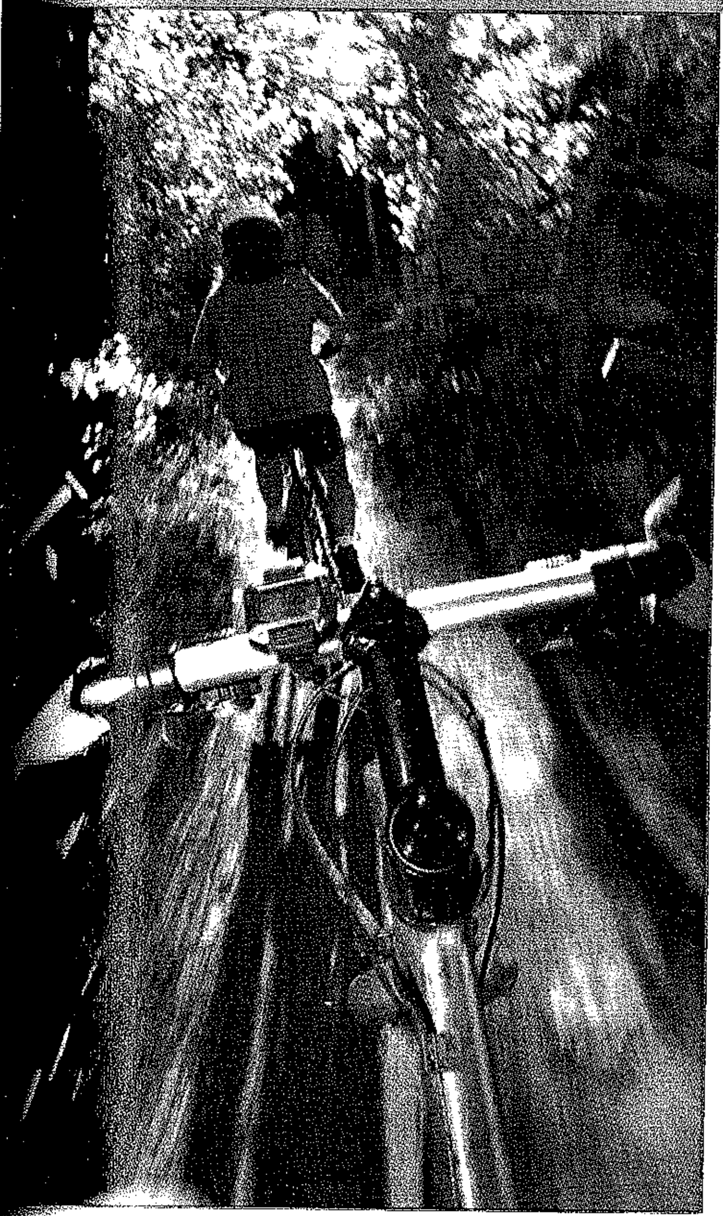


CHAPTER 11

Simple Linear Regression

Contents

Statistics in Action

Can "Dowsers" Really Detect Water?

Where We've Been

We've learned how to estimate and test hypotheses about population parameters based on a random sample of observations from the population. We've also seen how to extend these methods to allow for a comparison of parameters from two or more populations.

☛ Where We're Going

Suppose we want to predict the rainfall at a given location on a given day. We could select a single random sample of n daily rainfalls, use the methods of Chapter 7 to estimate the mean daily rainfall μ, and then use this quantity to predict the day's rainfall. A better method uses information that is available to any forecaster, e.g., barometric pressure and cloud cover. If we measure barometric pressure and cloud cover at the same time as daily rainfall, we establish the relationship between these variables—one that lets us use these variables for prediction. This chapter covers the simplest situation—relating two variables. The more complex problem of relating more than two variables is the topic of Chapter 12.

In Chapters 7–10 we described methods for making inferences about population means. The mean of a population has been treated as a *constant*, and we have shown how to use sample data to estimate or to test hypotheses about this constant mean. In many applications, the mean of a population is not viewed as a constant, but rather as a variable. For example, the mean sale price of residences in a large city last year can be treated as a constant and might be equal to $150,000. But we might also treat the mean sale price as a variable that depends on the square feet of living space in the residence. For example, the relationship might be

$$\text{Mean sale price} = \$30{,}000 + \$60(\text{Square feet})$$

This formula implies that the mean sale price of 1,000-square-foot homes is $90,000, the mean sale price of 2,000-square-foot homes is $150,000, and the mean sale price of 3,000-square-foot homes is $210,000.

What do we gain by treating the mean as a variable rather than a constant? In many practical applications we will be dealing with highly variable data, data for which the standard deviation is so large that a constant mean is almost "lost" in a sea of variability. For example, if the mean residential sale price is $150,000 but the standard deviation is $75,000, then the actual sale prices will vary considerably, and the mean price is not a very meaningful or useful characterization of the price distribution. On the other hand, if the mean sale price is treated as a variable that depends on the square feet of living space, the standard deviation of sale prices for any given size of home might be only $10,000. In this case, the mean price will provide a much better characterization of sale prices when it is treated as a variable rather than a constant.

In this chapter we discuss situations in which the mean of the population is treated as a variable, dependent on the value of another variable. The dependence of residential sale price on the square feet of living space is one illustration. Other examples include the dependence of mean reaction time on the amount of a drug in the bloodstream, the dependence of mean starting salary of a college graduate on the student's GPA, and the dependence of mean number of years to which a criminal is sentenced on the number of previous convictions.

In this chapter we discuss the simplest of all models relating a population mean to another variable, the *straight-line model*. We show how to use the sample data to estimate the straight-line relationship between the mean value of one variable, *y*, as it relates to a second variable, *x*. The methodology of estimating and using a straight-line relationship is referred to as *simple linear regression analysis*.

11.1 PROBABILISTIC MODELS

An important consideration when taking a drug is how it may affect one's perception or general awareness. Suppose you want to model the length of time it takes to respond to a stimulus (a measure of awareness) as a function of the percentage of a certain drug in the bloodstream. The first question to be answered is this: "Do you think an exact relationship exists between these two variables?" That is, do you think it is possible to state the exact length of time it takes an individual (subject) to respond if the amount of the drug in the bloodstream is known? We think you will agree with us that this is *not* possible for several reasons. The reaction time depends on many variables other than the percentage of the drug in the bloodstream—for example, the time of day, the amount of sleep the subject had the night before, the subject's visual acuity, the subject's general reaction time without the drug, and the subject's age would all probably affect reaction time. Even if many variables are included in a model (the topic of Chapter 12), it is still unlikely that we would be able to predict *exactly* the subject's reaction time. There will almost certainly be some variation in response times due strictly to *random phenomena* that cannot be modeled or explained.

If we were to construct a model that hypothesized an exact relationship between variables, it would be called a **deterministic model**. For example, if we believe that y, the reaction time (in seconds), will be exactly one and one-half times x, the amount of drug in the blood, we write

$$y = 1.5x$$

This represents a **deterministic relationship** between the variables y and x. It implies that y can always be determined exactly when the value of x is known. *There is no allowance for error in this prediction.*

If, on the other hand, we believe there will be unexplained variation in reaction times—perhaps caused by important but unincluded variables or by random phenomena—we discard the deterministic model and use a model that accounts for this **random error**. This **probabilistic model** includes both a deterministic component and a random error component. For example, if we hypothesize that the response time y is related to the percentage of drug x by

$$y = 1.5x + \text{Random error}$$

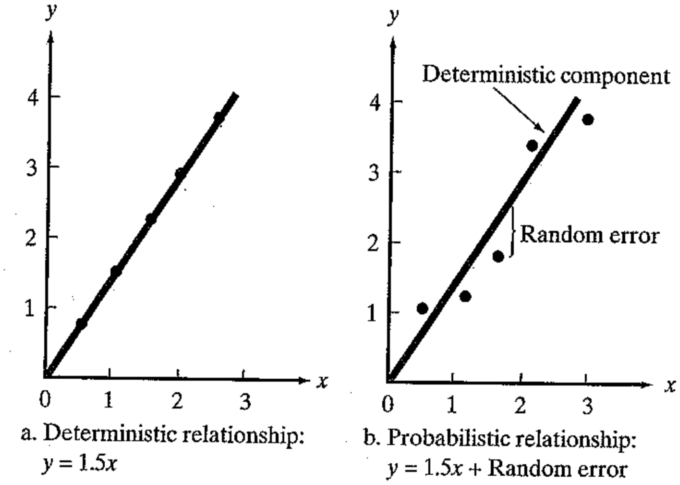
we are hypothesizing a **probabilistic relationship** between y and x. Note that the deterministic component of this probabilistic model is $1.5x$.

Figure 11.1a shows the possible responses for five different values of x, the percentage of drug in the blood, when the model is deterministic. All the responses must fall exactly on the line because a deterministic model leaves no room for error.

Figure 11.1b shows a possible set of responses for the same values of x when we are using a probabilistic model. Note that the deterministic part of the model (the straight line itself) is the same. Now, however, the inclusion of a random error component allows the response times to vary from this line. Since we know that the response time does vary randomly for a given value of x, the probabilistic model provides a more realistic model for y than does the deterministic model.

FIGURE 11.1

Possible Reaction Times, y, for Five Different Drug Percentages, x



a. Deterministic relationship: $y = 1.5x$

b. Probabilistic relationship: $y = 1.5x + \text{Random error}$

General Form of Probabilistic Models

y = Deterministic component + Random error

where y is the variable of interest. We always assume that the mean value of the random error equals 0. This is equivalent to assuming that the mean value of y, $E(y)$, equals the deterministic component of the model; that is,

$E(y)$ = Deterministic component

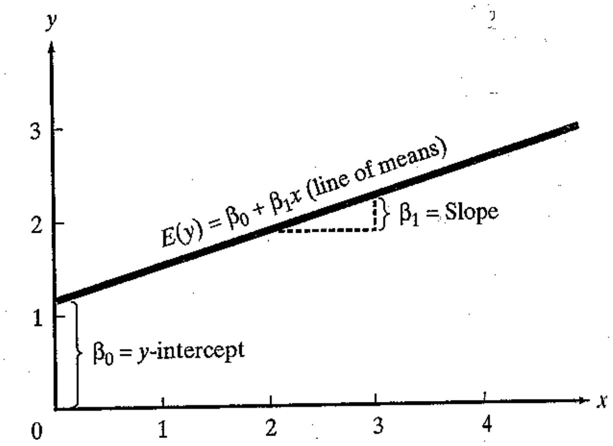
In this chapter we present the simplest of probabilistic models—the **straight-line model**—which derives its name from the fact that the deterministic portion of the model graphs as a straight line. Fitting this model to a set of data is an example of **regression analysis**, or **regression modeling**. The elements of the straight-line model are summarized in the next box.

A First-Order (Straight-Line) Probabilistic Model

$y = \beta_0 + \beta_1 x + \varepsilon$

where y = **Dependent** *or* **response variable** (variable to be modeled)
x = **Independent** *or* **predictor variable** (variable used as a predictor of y)*
$E(y) = \beta_0 + \beta_1 x$ = Deterministic component
ε (epsilon) = Random error component
β_0 (beta zero) = **y-intercept of the line**, that is, the point at which the line intersects or cuts through the y-axis (see Figure 11.2)
β_1 (beta one) = **Slope of the line**, that is, the amount of increase (or decrease) in the deterministic component of y for every 1-unit increase in x. [As you can see in Figure 11.2, $E(y)$ increases by the amount β_1 as x increases from 2 to 3.]

FIGURE 11.2
The Straight-Line Model



*The word *independent* should not be interpreted in a probabilistic sense, as defined in Chapter 3. The phrase *independent variable* is used in regression analysis to refer to a predictor variable for the response y.

In the probabilistic model, the deterministic component is referred to as the **line of means**, because the mean of y, $E(y)$, is equal to the straight-line component of the model. That is,

$$E(y) = \beta_0 + \beta_1 x$$

Note that the Greek symbols β_0 and β_1, respectively, represent the y-intercept and slope of the model. They are population parameters that will be known only if we have access to the entire population of (x, y) measurements. Together with a specific value of the independent variable x, they determine the mean value of y, which is just a specific point on the line of means (Figure 11.2).

The values of β_0 and β_1 will be unknown in almost all practical applications of regression analysis. The process of developing a model, estimating the unknown parameters, and using the model can be viewed as the five-step procedure shown in the next box.

Step 1 Hypothesize the deterministic component of the model that relates the mean, $E(y)$, to the independent variable x (Section 11.1).

Step 2 Use the sample data to estimate unknown parameters in the model (Section 11.2).

Step 3 Specify the probability distribution of the random error term and estimate the standard deviation of this distribution (Sections 11.3 and 11.4).

Step 4 Statistically evaluate the usefulness of the model (Sections 11.5, 11.6, and 11.7).

Step 5 When satisfied that the model is useful, use it for prediction, estimation, and other purposes (Section 11.8).

EXERCISES 11.1–11.9

Learning the Mechanics

11.1 In each case, graph the line that passes through the given points.

a. $(1, 1)$ and $(5, 5)$
b. $(0, 3)$ and $(3, 0)$
c. $(-1, 1)$ and $(4, 2)$
d. $(-6, -3)$ and $(2, 6)$

11.2 Give the slope and y-intercept for each of the lines graphed in Exercise 11.1.

11.3 The equation for a straight line (deterministic) is

$$y = \beta_0 + \beta_1 x$$

If the line passes through the point $(-2, 4)$, then $x = -2, y = 4$ must satisfy the equation; that is,

$$4 = \beta_0 + \beta_1(-2)$$

Similarly, if the line passes through the point $(4, 6)$, then $x = 4, y = 6$ must satisfy the equation; that is,

$$6 = \beta_0 + \beta_1(4)$$

Use these two equations to solve for β_0 and β_1; then find the equation of the line that passes through the points $(-2, 4)$ and $(4, 6)$.

11.4 Refer to Exercise 11.3. Find the equations of the lines that pass through the points listed in Exercise 11.1.

11.5 Plot the following lines:

a. $y = 4 + x$
b. $y = 5 - 2x$
c. $y = -4 + 3x$
d. $y = -2x$
e. $y = x$
f. $y = .50 + 1.5x$

11.6 Give the slope and y-intercept for each of the lines defined in Exercise 11.5.

11.7 Why do we generally prefer a probabilistic model to a deterministic model? Give examples for which the two types of models might be appropriate.

11.8 What is the line of means?

11.9 If a straight-line probabilistic relationship relates the mean $E(y)$ to an independent variable x, does it imply that every value of the variable y will always fall exactly on the line of means? Why or why not?