Theory Construction and Model-Building Skills

Theory Construction and Model-Building Skills

A Practical Guide for Social Scientists

James Jaccard | Jacob Jacoby

Series Editor's Note by David A. Kenny

THE GUILFORD PRESS
New York London

8

Mathematical Modeling

Even if there is only one possible unified theory, it is just a set of rules and equations.

—Stephen W. Hawking (1988)

This chapter describes an approach to theory construction called *mathematical modeling*. Like causal modeling, the approach involves describing relationships between variables, but the emphasis is on describing those relationships using mathematical concepts. Mathematical models can be used in conjunction with causal thinking, as we demonstrate in a later section of this chapter, but social scientists who use mathematical modeling tend not to think in terms of indirect causes, mediated relationships, and moderated relationships in the way that we outlined in Chapter 7. Instead, they focus on thinking about functions and describing relationships mathematically based on functions. They more often than not use nonlinear relationships. Our focus here is not on integrating causal and mathematical modeling as approaches to theory construction. Rather, we merely wish to provide you with an additional tool for your theory construction toolbox, mathematical modeling, as you strive to gain insights into the phenomena you want to study.

Constructing mathematical models can involve complex mathematics that go well beyond the background of many readers of this book. Entire books have been written on mathematical modeling that assume years of study of calculus and formal mathematics. Our treatment must, accordingly, be limited, and we provide only a general sense of building mathematical models and thinking as a math modeler would. However, the chapter should be a good starting point for delving into this approach in more depth vis-à-vis the suggested readings at the end of the chapter.

Mathematical modeling is common in the physical sciences, but it is used less often in the social sciences. Our goal is to provide you with a sense of mathematical modeling as it is pursued in the social sciences. In this chapter we first expose you to basic concepts and terms you will encounter as you read about math models or pursue mathematical modeling. More specifically, we distinguish between categorical, discrete,

and continuous variables; differentiate axioms and theorems; introduce the notion of a function; use linear functions to identify key features of functions; and describe the difference between deterministic and stochastic models. We also provide an intuitive overview of derivatives, differentiation, integrals, and integration in calculus, as well key notions of model identification and metrics. We next describe five commonly used functions in math models: logarithmic functions, exponential functions, power functions, polynomial functions, and trigonomic functions, as well as functions for categorical variables. We conclude our background section by considering ways of transforming and combining functions and building functions for multiple variable scenarios.

Following the presentation of these key concepts, we describe the phases of building a mathematical model and then provide four examples of such models in the social sciences. We then briefly characterize chaos theory and catastrophe theory as influential mathematical models in the social sciences. Our initial discussion may seem a bit fractured as we develop one mathematical concept after another. Be patient. Later sections will pull it all together.

TYPES OF VARIABLES: CATEGORICAL, DISCRETE, AND CONTINUOUS

In Chapter 6 we distinguished between categorical and quantitative variables. A categorical variable has different "levels," "values," or "categories," and there is no special ordering to the categories along an underlying dimension. The categories are merely labels that differentiate one group from another (e.g., "male" or "female" for the variable of gender). In contrast, a quantitative variable is one in which individuals (in the social sciences) are assigned numerical values to place them into different categories, and the numerical values have meaning in that they imply more or less of an underlying dimension that is of theoretical interest.

Mathematical modelers make distinctions between discrete quantitative variables and continuous quantitative variables. A *discrete variable* is one in which there are a finite number of values between two values. For example, for the number of children in a family, there is a finite number of values, say, between 1 child and 4 children, namely the values of 2 children and 3 children. We do not think of there being 1.5 children or 1.7 children in a family. For a *continuous variable*, however, there is an infinite number of values between any two values. Reaction time to a stimulus is an example of a continuous variable. Even between the values of 1 and 2 seconds, an infinite number of values could occur (e.g., 1.001 seconds, 1.873 seconds, 1.874 seconds).

Whether a variable is classified as discrete or continuous depends on the nature of the underlying theoretical dimension and not on the scale used to measure that dimension. Tests that measure intelligence, for example, yield scores that are whole numbers (e.g., 101, 102); hence the scores are discrete. Nevertheless, intelligence is continuous in nature because it involves a dimension that permits an infinite number of values to occur, even though existing measuring devices are not sensitive enough to make such

fine distinctions. In the reaction time example, the measurement of time can be very precise with modern equipment, but there even is a limit to the precision possible with measures of time. Such limits in the precision of measurement do not make the underlying dimension discrete. Reaction time is continuous in character.

Even though social scientists often must rely on discrete measures of continuous constructs, they build models with those measures as if they were continuous. As long as the measures are comprised of many values, this practice usually is not problematic. It is only when the number of values in the measure of a continuous variable is few that problems can arise and special considerations in the modeling effort need to be made.

The distinction between discrete and continuous variables is important because the strategies used to construct a mathematical model differ depending on whether the quantitative variables are discrete or continuous. We devote most of our attention to the case in which the theorist is working with continuous variables, but we occasionally consider qualitative and discrete variables as well.

AXIOMS AND THEOREMS

The term *axiom* is used in many ways in the social sciences, but in mathematics, an axiom is a mathematical statement that serves as a starting point from which other mathematical statements are logically derived. Axioms are "given." They are not derived through deduction, nor are they the subject of mathematical proofs. They are starting points. By contrast, a *theorem* is a statement that can be logically derived from, or is proven by, one or more axioms or previous statements. The use of these terms and many variants of them (e.g., a proposition, a lemma, a corollary, a claim, an identity, a rule, a law, a postulate, a principle) vary somewhat depending on the branch of mathematics, but the above characterization captures the essence of axioms and theorems as used in mathematical models in the social sciences.

FUNCTIONS

Functions are central to mathematical modeling. A simple analogy for thinking about functions is to think of a machine that you put something into and get something back, based on what you input. For example, you might press a key that inputs the number 3 into a machine and out comes the number 9. You might press another key that inputs the number 5 into the machine and out comes the number 15. The machine in this case represents the function "take the input value and triple it." Functions in math typically involve numbers as inputs and outputs.

Suppose we decide to name our machine *Jack*. We can write the function that the machine performs as follows:

$$\text{Jack}(X) = 3X$$

This equation states that whatever the value of X, the value of "Jack of X" will be triple it. The traditional notation is to name the machine f (for "function") and write it as follows:

$$f(X) = 3X$$

All functions have what are called a domain and a range. The *domain* is the set of possible input values, and the *range* is the set of possible output values. The domain and range often are stated mathematically rather than listed as individual numbers. For example, for the function

$$f(X) = \sqrt{X-3}$$

the domain or possible input values is any number greater than or equal to 3 (because you can not calculate the square root of a negative number), and the range or possible output values is any value greater than or equal to 0.[1] The domain is any number that produces a "meaningful output" and that will not cause the machine to malfunction (e.g., the number 2, which would require us to calculate the square root of –1). A shorthand way that mathematical modelers use to express the domain is "the domain is $\{X|X \geq 3\}$," where the symbol "|" is read as "given that." This expression states that the domain is equal to X, given that X is greater than or equal to 3. This may seem a bit cryptic, but it is an efficient way of stating a domain or a range. For example, I might have a function where the domain is $\{X|X > 0\}$ and the range is $\{Y|Y > 0\}$.

Functions can apply to more than a single input. For example, the function $f(X,Z) = X - Z$ has two inputs, X and Z, and an output that is the difference between them. For example, if $X = 5$ and $Z = 2$, the function $f(X,Z)$ yields the output 3.

Functions are the foundation of mathematical models. When one "maps" a function between Y and X, one attempts to specify what function applied to values on X will produce the values of Y. A central task in mathematical modeling is that of mapping functions.

LINEAR FUNCTIONS

One of the most commonly used functions in the social sciences is the linear function. In this section we describe the nature of linear functions and then use them to illustrate

[1]For pedagogical reasons, we restrict all examples in this chapter to real numbers.

basic issues in building mathematical models. In later sections we consider other functions.

The Slope and Intercept

The Slope

Consider the two-variable example we used in Chapter 6 to develop the general idea of a linear relationship, namely, the number of hours an employee worked, *X*, and the amount of money paid to the employee, *Y*. Assume a scenario where each of four individuals works at a rate of $1 per hour. Their scores are:

	X	*Y*
Individual	(hours worked)	(dollars paid)
1	1	1
2	4	4
3	3	3
4	2	2

The relationship between *X* and *Y* is illustrated in Figure 8.1, which uses a scatterplot with connected dots. As indicated by the straight line on the scatterplot, there is a linear relationship between *X* and *Y*. This relationship can be stated mathematically as

$$Y = X$$

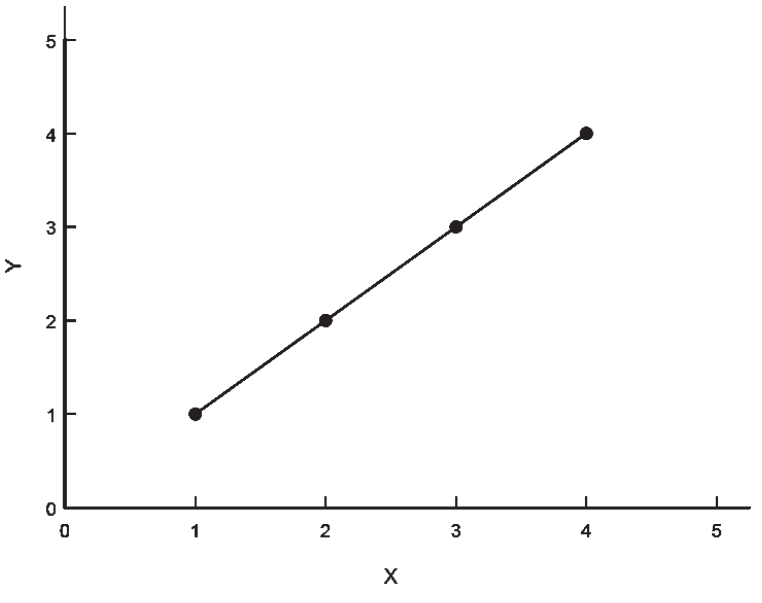

FIGURE 8.1. Linear relationship with slope = 1.

In other words, the number of dollars paid equals the number of hours worked.

Suppose the individuals were not paid $1 per hour, but instead were paid $2 per hour. The scores on *X* and *Y* would be as follows:

Individual	*X* (hours worked)	*Y* (dollars paid)
1	1	2
2	4	8
3	3	6
4	2	4

In this case, the relationship between *X* and *Y* can be stated as

$$Y = 2.00X$$

In other words, the number of dollars paid equals 2 times the number of hours worked. Figure 8.2 presents a scatterplot of these data (line *B*) as well as the data from Figure 8.1 (line *A*). (Line *C* is explained on p. 184.) Notice that we still have a straight line (and, hence, a linear relationship) but, in the case of $2 per hour, the line rises faster than with $1 per hour; that is, the slope of the line is now steeper. Technically, the slope of a line indicates the number of units that variable *Y* changes when variable *X* increases by 1 unit. It is the rate of change in *Y* given a 1-unit increase in *X*. When the wage is $2 per hour, a person who works 1 hour is paid $2, a person who works 2 hours is paid $4,

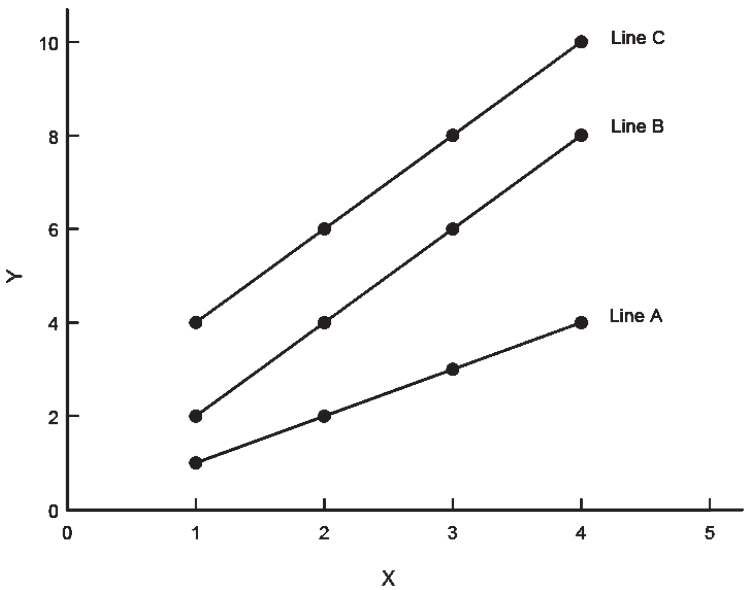

FIGURE 8.2. Example of linear relationships with different slopes or intercepts.

and so on. When X increases by 1 unit (e.g., from 1 to 2 hours), Y increases by 2 units (e.g., from \$2 to \$4). The slope that describes this linear relationship is therefore 2. In contrast, the slope that describes the linear relationship $Y = X$ is 1.0, meaning that as X increases by 1 unit, so does Y. One way in which linear relationships differ is in terms of the slopes that describe them.

The slope that describes a linear relationship can be determined from a simple algebraic formula. This formula involves first selecting the X and Y values of any two individuals. The slope is computed by dividing the difference between the two Y scores by the difference between the two X scores; in other words, the change in Y scores is divided by the change in X scores. Algebraically,

$$b = (Y_2 - Y_1)/(X_2 - X_1) \tag{8.1}$$

where b represents the slope, X_1 and Y_1 are the X and Y scores for any one individual, and X_2 and Y_2 are the X and Y scores for any other individual. In our example, inserting the scores for individuals 1 ($X = 1$, $Y = 2$) and 2 ($X = 4$, $Y = 8$) into Equation 8.1, we find that the slope for line B is

$$b = (2 - 8)/(1 - 4) = 2.00$$

This is consistent with what was stated previously.

The value of a slope can be positive, negative, or 0. Consider the following scores:

Individual	X	Y
1	2	3
2	1	4
3	4	1
4	3	2

Inserting the scores for individuals 2 and 4 into Equation 8.1, we find that the slope is

$$b = (4 - 2)/(1 - 3) = -1.00$$

Figure 8.3 presents a scatterplot of the data for this relationship. The relationship is still linear, but now the line moves downward as we move from left to right on the X axis. This downward direction characterizes a negative slope, whereas an upward direction characterizes a positive slope. A slope of 0 is represented by a horizontal line because the value of Y is constant for values of X.

In sum, a positive slope indicates a *positive* or *direct linear relationship* between X and Y, whereas a negative slope indicates a *negative* or *inverse linear relationship* between X and Y. In the case of a positive relationship, as scores on X *increase*, scores on Y also *increase*. In the case of a negative relationship, as scores on X *increase*, scores on Y

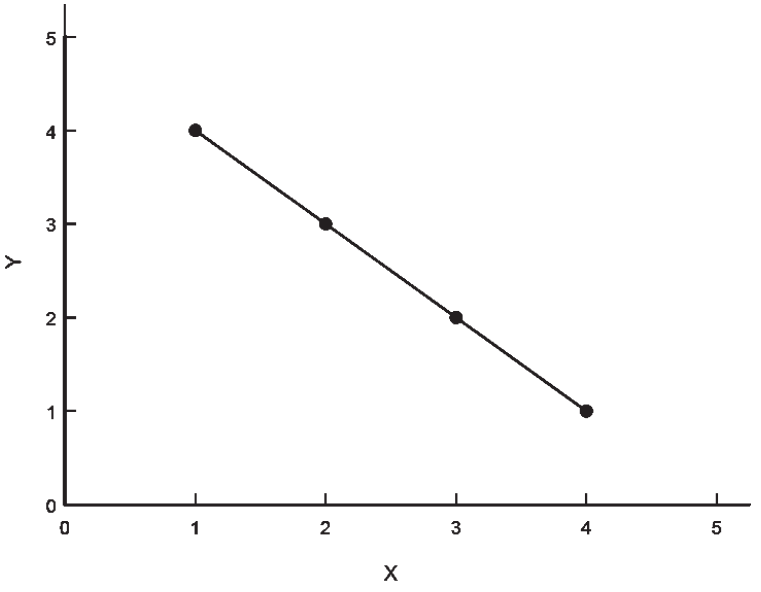

FIGURE 8.3. Example of a negative slope.

decrease. For instance, the slope in Figure 8.3 is –1.00, meaning that for every unit *X* increases, *Y* decreases by one unit.

The Intercept

Let us return to the example where individuals are paid $2 per hour worked. Suppose that in addition to this wage, each individual is given a tip of $1.50. Now the relationship between *X* and *Y* is

$$Y = 1.50 + 2.00X \tag{8.2}$$

Line *C* of Figure 8.2 plots this relationship for the four individuals. If we compute the slope of this line, we find it is 2.00, as before. Notice that lines *C* and *B* are parallel but that line *C* is higher up on the *Y* axis than line *B*. The amount of separation between these two lines can be measured at the *Y* axis, where $X = 0$. When $X = 0$, the *Y* value is 1.50 for line *C* and 0 for line *B*. Thus, line *C* is raised 1.50 units above line *B*.

The point at which a line intersects the *Y* axis when $X = 0$ is called the *intercept*, and its value is denoted by the letter *a*. Another way of thinking about the intercept is that it is the value of *Y* when *X* is zero.

Linear relationships can differ in the values of their intercepts as well as the values of their slopes. The general form of a linear equation is

$$Y = a + bX \tag{8.3}$$

Stated more formally, the linear function is

$$f(X) = a + bX$$

where a and b are constants representing an intercept and slope and X is a variable. A variable, Y, is described by this function if $Y = f(X)$, that is, $Y = a + bX$. Equation 8.3 is called a *linear equation*.

DETERMINISTIC VERSUS STOCHASTIC MODELS

Any linear relationship can be represented by Equation 8.3. A slope and intercept always describe the linear relationship between two variables. Given values of the slope and intercept, we can substitute scores on X into the linear equation to determine the corresponding scores on Y. For example, the linear equation $Y = 1.50 + 2.00X$ tells us that an individual who works for 2 hours is paid \$5.50 because the Y score associated with an X score of 2 is

$$Y = 1.50 + 2.00X = 1.50 + (2.00)(2) = \$5.50$$

An individual who works for 3 hours is paid $Y = 1.50 + (2.00)(3) = \$7.50$, and an individual who works for 4 hours is paid $Y = 1.50 + (2.00)(4) = \$9.50$.

When one variable is a linear function of another, all the data points on a scatterplot will fall on a straight line. However, rarely in the social sciences will we encounter such situations. When two variables only approximate a linear relationship, we need to add a term to the linear equation to accommodate random disparities from linearity. The term is called a *disturbance* or *error term*, yielding the equation

$$Y = a + bX + e$$

where e is the difference between the observed Y score and the predicted score based on the linear function. The errors are assumed to be random rather than systematic because if the errors were systematic, then some meaningful form of nonlinearity would be suggested and would need to be modeled. It is important to keep in mind that e is an unmeasured variable that reflects the disparity between scores predicted by the model and observed scores.

Formal mathematical models do not include an error term when specified at the theoretical level. In this sense, they are deterministic rather than probabilistic. However, when testing mathematical models empirically, it is common for researchers to include an error term because there usually is some random "noise" that creates disparities from model predictions. A common practice is to identify the function that seems appropriate for predicting and understanding a phenomenon and then, in data-based tests of the model, to add an error term to accommodate the hopefully small but random discrep-

ancies that seem inevitable. A model is better if the discrepancies from predictions are trivial and have no practical consequence.

In the world of mathematical models, you will encounter distinctions between deterministic and probabilistic models. A *deterministic model* is one in which there is no random error operating. The model performs the same way for any given set of conditions. In contrast, a *probabilistic model* is one in which some degree of randomness is present. Probabilistic models are also sometimes referred to as *stochastic models*.

MODEL PARAMETERS

Adjustable Parameters and Parameter Estimation

Mathematical models typically include variables that are measured as well as constants whose values can be derived logically or estimated from data. For example, in the linear function

$$f(X) = a + bX$$

there is a variable, X, and two model constants that need to be specified, the intercept, a, and the slope, b. Constants such as the intercept and the slope are called *adjustable parameters* or *adjustable constants*, because their values can be set by the theorist to different values so as to affect the output of the function. For example, we might state that annual income is a function of the number of years of education, where the function is defined as $f(X) = 1{,}000 + 5{,}000X$. If the number of years of education is 2, then output of the function is $1{,}000 + (5{,}000)(2) = 11{,}000$. By contrast, we might state that the function is $f(X) = 2{,}000 + 4{,}000X$. If the number of years of education is 2, then the output of the function is $2{,}000 + (4{,}000)(2) = 10{,}000$. The slope and intercept are adjustable constants that affect the output value of the function as different values of X are substituted into the function.

When one is unsure what the value of the adjustable constants should be, then strategies can be used to estimate their values empirically based on data. For example, in a linear model where the relationship between Y and X is linear, except for the presence of random noise (i.e., $Y = a + bX + e$), a researcher might obtain data for the values of Y and the values of X for a group of individuals and then use traditional least-squares regression methods to estimate the values of the intercept and slope.

Mathematical models differ in the number of adjustable constants they include and in the number of constants that must be estimated from data. Models with many parameter values that must be estimated are less parsimonious and often present greater challenges for testing than models with fewer estimated parameters. When the value of an adjustable parameter is specified a priori by the theorist and not estimated, it is said to be *fixed*. When the value of the adjustable parameter is estimated from data, it is said to

be *estimated*. Thus, you will hear reference to fixed parameters and estimated parameters in math models.

RATES AND CHANGE: DERIVATIVES AND DIFFERENTIATION

Parameters in a mathematical model often are subject to meaningful interpretation. In the linear model, $Y = a + bX$, the slope reflects the predicted change in Y given a 1-unit change in X. It is calculated using Equation 8.1, which we repeat here:

$$b = (Y_2 - Y_1)/(X_2 - X_1)$$

The slope is meaningful because it provides a sense of how much change in Y we can expect, given a change in X. Note that in the linear model, it does not matter where the change occurs on the X continuum. A 1-unit change on X, at the low end of the X continuum, will produce the same amount of change in Y as a 1-unit change in X at the high end of the X continuum. The value of the slope tells us how much change this is.

The slope is, in essence, a rate of change in Y, given a unit change in X. More generally, if we describe the change in Y between any two points as

$$\Delta Y = Y_2 - Y_1$$

and the change in X between those same two points as

$$\Delta X = X_2 - X_1$$

then the rate of change in Y relative to the change in X is the ratio of these

$$\text{Rate of change} = \frac{\Delta Y}{\Delta X} = \frac{Y_2 - Y_1}{X_2 - X_1}$$

which, in this case, is the value of the slope. If $\Delta Y = 4$ and $\Delta X = 2$, then the rate of change of Y relative to a unit change in X is $4/2 = 2$.

The property of equal amounts of change at all points on the X continuum does not apply to nonlinear relationships. Consider the nonlinear relationship between Y and X shown in Figure 8.4. At low values of X, small changes in X result in no change in Y, whereas at high values of X, small changes in X result in large changes in Y. The impact of a 1-unit change in X differs depending on the part of the X continuum at which the change occurs.

When analyzing change, two fundamental concepts from calculus are helpful: derivatives and differentiation. *Derivatives* refer to the concept of instantaneous change, and *differentiation* refers to algebraic methods for calculating the amount of instantaneous change that occurs. Let us explore these concepts in more depth.

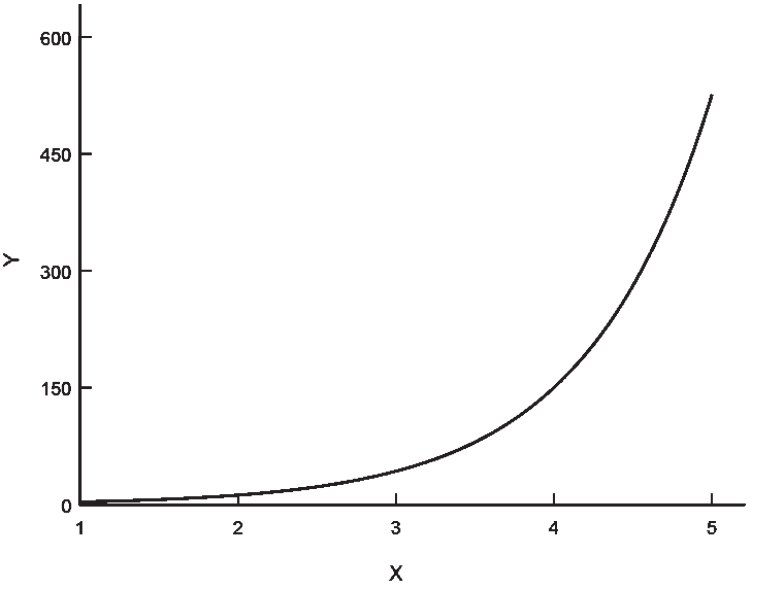

FIGURE 8.4. Nonlinear relationship for derivative example.

Instantaneous Change

Suppose we want to measure the speed of a car driving between two towns, Town A and Town B, that are 120 miles apart. Let Y be the distance traveled by the car. When the car is in Town A and just about to begin its journey, the car has traveled 0 miles, so we set $Y_1 = 0$. When the car reaches Town B, it has traveled 120 miles, so we set $Y_2 = 120$. Now let X be the amount of time the car spends traveling. Before the car leaves Town A, $X_1 = 0$ hours. Suppose when the car finally reaches Town B, the car has been on the road for 2 hours. This means that $X_2 = 2$ hours. Using the logic from above, the rate of change in Y as a function of X is

$$\text{Rate of change} = \frac{(Y_2 - Y_1)}{(X_2 - X_1)} = \frac{\Delta Y}{\Delta X} = \frac{(120 - 0)}{(2 - 0)} = 60 \tag{8.4}$$

or 60 miles per hour. A 1-unit change in time (X, as measured in hours) is associated with a 60-unit change in distance (Y, as measures in miles).

The value of 60 miles per hour represents the average speed of the car during the entire trip. But it is probably the case that the car did not travel at a speed of exactly 60 miles per hour during the entire trip. At times, it probably was driven faster and at other times, slower. Suppose we wanted to know how fast the car was going 15 minutes into the trip. One way of determining this number is to define values for X_1 and Y_1 at 14 minutes and 59 seconds into the trip and then to define X_2 and Y_2 values at 15 minutes and 1 second into trip. We could then apply Equation 8.4 to this more narrowly defined time frame. Although the result would give us a sense of how fast the car was being driven 15

minutes into the trip, it would not tell us how fast the car was being driven at *exactly* 15 minutes into the trip. We want to know at the very instant of 15 minutes into the trip, how fast the car was going, that is, what its rate of change was at that particular instant. It is this concept of instantaneous change to which a derivative refers. The velocity that the car is traveling at an exact point in time maps onto the notion of a derivative.

For a nonlinear relationship such as that in Figure 8.4, it is possible to use differentiation to calculate the instantaneous rate of change in Y at any given value of X. The derivative is the (instantaneous) slope of Y on X at that given point of X. It is analogous to specifying the speed at which a car is being driven at a specific point in time. For some modeling problems, calculating a derivative by the process of differentiation is straightforward. For other problems, it can be quite complex. Methods of differentiation are taught in calculus and need not concern us here. The main point we want to convey is that in many forms of mathematical modeling, rates of change in Y as a function of changes in X are described using the language of derivatives, and it is important that you have a sense of that language.

For linear models, the instantaneous rate of change in Y at some point on the X continuum is the same as the instantaneous rate of change in Y at any other point on the X continuum. By contrast, for the nonlinear relationship in Figure 8.4, the instantaneous rate of change depends on where on the X continuum the change is occurring. In Figure 8.4 the derivative (i.e., instantaneous rate of change) when $X = 1$ is 0.04, whereas when $X = 4$, the derivative is 1.98. We calculated these values using calculus. A common notation for signifying a derivative is dY/dX. A common phrase for describing derivatives is to state "the value of the derivative at $X = 4$ is 1.98." If the derivative has the same value at all points on X (as is the case for a linear relationship), then one refers simply to "the derivative" without specifying the value of X at which the derivative is calculated.

You also may encounter a derivative expressed as a rate of change (ΔY and ΔX), but invoking what is called a limit, perhaps as follows:

$$\lim_{\Delta X \to 0} \frac{\Delta Y}{\Delta X}$$

The left-hand part of this expression contains the abbreviation *lim* (for the word *limit*), and the entire expression describes symbolically the idea of instantaneous change. Specifically, this expression is read as "the change in Y relative to the change in X as the change in X approaches its lower limit of zero" (analogous to the case where we calculated speed at exactly 15 minutes into the trip). The expression is just a way of referring to a derivative in a more formal way.

Second and Third Derivatives

In calculus some functions have higher-order derivatives, such as a second derivative or a third derivative. We will not use second or third derivatives in the mathematical models considered in this chapter, but it will help to have some appreciation for these concepts. As noted above, a derivative refers to a rate of change of one variable (ΔY) relative to the

rate of change of another variable (ΔX) in the context of instantaneous change. In our driving example, the first derivative refers to the speed or velocity with which a car is driven at any given point in time. Suppose the car is driving along and the driver decides to speed up. The result of pressing harder on the accelerator is that the car's velocity (i.e., the first derivative) increases. A second derivative in this case refers to the change in the first derivative that occurred at any given instant. It is analogous to what we commonly call acceleration. When you press the accelerator pedal, you "change" your speed. How much your speed changes at a given instant is the idea of a second derivative. If, in turn, your rate of acceleration changes (e.g., you "let off" the pedal and decelerate), then this maps onto the idea of a third derivative.

In a linear function the value of the second derivative is zero, because there is never a change in the value of the first derivative at different points of *X*. For nonlinear relationships, the value of the second derivative is nonzero at different points on *X*. When reading about mathematical models, in addition to the concept of first derivatives, you may encounter the concepts of second or third derivatives.

In sum, derivatives are useful concepts for describing rates of change in *Y* as a function of *X*. For nonlinear functions, the rate of change in *Y* will differ depending on where on the *X* continuum the change is occurring. A derivative is an index of instantaneous change at a given *X* value. It is a slope, but a special one, namely an "instantaneous" slope. First derivatives are fairly straightforward. Second and third derivatives are a bit more abstract. For those readers familiar with interaction effects in statistics, a second-order derivative is roughly analogous to a two-way interaction and a third order derivative is roughly analogous to a three-way interaction.

DESCRIBING ACCUMULATION: INTEGRALS AND INTEGRATION

When describing mathematical models, many theorists emphasize derivatives, that is, rates of change. There is another concept in calculus that is sometimes emphasized in mathematical models: the *integral*. This concept reflects the amount of something. The process of calculating an integral is called *integration*. To gain a sense of what an integral is, consider the well-known function in statistics of the probability density function for a standard normal distribution. This function, often presented in statistics texts, is the basis for calculating the "area under the curve" in a normal distribution. Figure 8.5 presents a graphical representation of this function, as it often appears in statistics books. The various *X* values on the horizontal axis are standard scores, with a mean of zero and a standard deviation of 1. One can specify any two points in this distribution, say, a value of 0 and a value of 1, and then calculate the area under the curve between these two points. If one scales the total area under the curve to equal a value of 1.00, then the area under the curve between two *X* scores is the proportion of the total area that falls between the two points. For example, the area under the curve between the *X* values of 0 and 1 is 0.3413 (see Figure 8.5). Between the *X* values of 1 and 2, the area under the curve is 0.1359. Between *X* values of –1 and 1, the area under the curve is 0.6826.

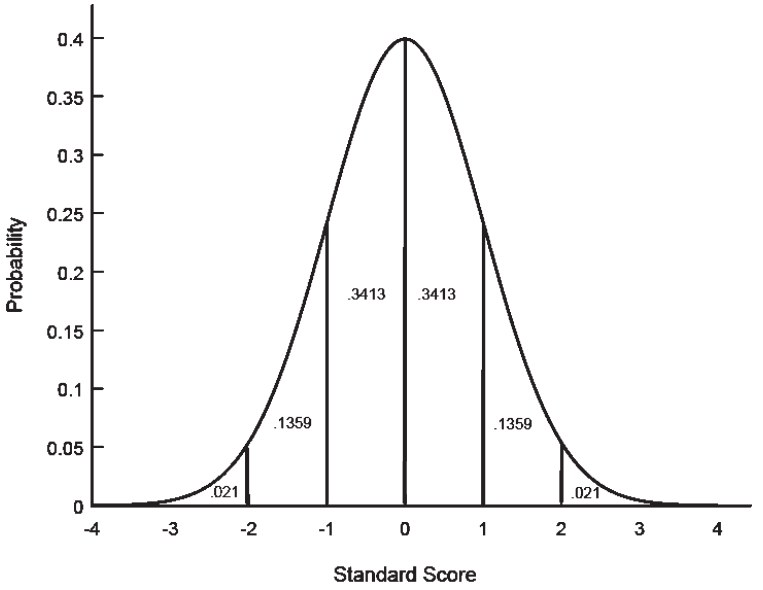

FIGURE 8.5. Area under the curve for a standardized normal distribution.

Graphically, an integral is the area under the curve between two points. The integral for the values 0 and 1 in Figure 8.5 is 0.3413. The integral for the values 1 and 2 is 0.1359. Because it focuses on the area under the curve, one can see that, roughly speaking, an integral refers to "the amount of something."

A common use of integrals in mathematical models is to characterize accumulations, that is, how much of something has accumulated. Many phenomena accumulate. We accumulate money in savings accounts. Frustration accumulates with each stressful event experienced within a short time span. Although the mathematical details of integration are well beyond the scope of this book, when one uses the concept of integrals, one often does so in the context of building models of phenomena that accumulate.

JUST-IDENTIFIED, OVERIDENTIFIED, AND UNDERIDENTIFIED MODELS

Mathematical models vary in their identification status. *Model identification* refers to cases where the values of model parameters must be estimated from data. A *just-identified* model is one for which there is a unique solution (i.e., one and only one solution) for the value of each estimated parameter in the model. Consider an analogy from algebra, where we might be given two equations with two unknowns of the following form:

$$23 = 2X + 3Z$$
$$9 = X + Z$$

For these two equations, there is a unique solution for X and Z: $X = 4$ and $Z = 5$.

An *underidentified* model is one for which there is an infinite number of solutions for one or more of the model parameters. In the equation

$$10 = X + Z$$

there is an infinite number of solutions for X and Z (e.g., $X = 10$ and $Z = 0$ is one solution; $X = 9$ and $Z = 1$ is another solution). Models that have one or more parameters that are underidentified are often problematic.

An *overidentified* model is one for which there is a unique solution for the model parameters, *and* there is more than one feature of the model that can be used to independently estimate the parameter values. Using the algebraic analogy, consider the following three equations:

$$10 = X + Z$$
$$18 = 2X + Z$$
$$12 = X + 2Z$$

There are three equations with a total of two unknowns, and any given pair of equations, no matter which pair, can be used independently to solve for the unknowns. In models for which parameter values must be estimated and the function fit is not perfect (i.e., there is an error term such that the model is stochastic), model parameters that have overidentified status are desirable because one can obtain independent estimates of those parameter values.

In sum, when reading math models, you may encounter references to a model as being just-identified, underidentified, or overidentified. Models that are underidentified are unsatisfactory.

METRICS

When developing mathematical models, theorists give careful consideration to the metric on which the variables in the model are measured, especially when nonlinear modeling is involved. This is because the accuracy of a mathematical model and the inferences one makes can be (but are not always) influenced by the metric of the variables. Depending on the variable metric, a theorist might resort to different functions in the mathematical model to describe the relationships between variables. For example, for the variable time, the model form and parameters introduced into the model might vary depending on whether time is measured in milliseconds, seconds, days, weeks, or years. The nature of metrics poses difficulty for some constructs in the social sciences because the metric used to measure them is arbitrary. For example, when a researcher uses a 10-item agree–disagree scale to measure peoples' attitudes toward religion, the metric might be scored from –5 to +5, or from 0 to 9, or from 1 to 10. In some mathematical

models, the choice of scoring matters a great deal, so an arbitrary metric can create modeling difficulties.

TYPES OF NONLINEARITY

Thus far we have considered a simple mathematical model—the linear model—to introduce several concepts in mathematical modeling. The linear model has two adjustable parameters, a slope and an intercept, that typically are estimated rather than fixed by the theorist. In this section we introduce other functions that are nonlinear in form and that can be used in mathematical models. There are many classes of functions, and we cannot begin to describe them all. Here we focus on describing five major classes of functions (the linear function makes six classes). The idea is to give you a sense of some of the nonlinear functions that can be used to build a math model. After presenting the functions, we describe strategies for modifying and combining them to build even more intricate mathematical representations. As we describe the different functions and the modifications to them that can be made, you will see the wide range of tools available to a math modeler when characterizing relationships between variables.[2]

To describe functions, we often use three concepts: (1) concavity, (2) proportionality, and (3) scaling constants. *Concavity* refers to whether the rate of change on a curve (the first derivative) is increasing or decreasing. A curve that is *concave upward* has an increasing first derivative, and a curve that is *concave downward* has a decreasing first derivative. In terms of proportionality, two variables are proportional to one another when one variable is a multiple of the other. More formally, Y is proportional to X if $Y = cX$, where c is a constant. The value c is called the *constant of proportionality*. For proportionality, doubling X doubles Y, tripling X triples Y, and halving X halves Y. Two variables are said to be *inversely proportional* when there is some constant c for which $Y = c/X$. In this case, doubling X halves Y, tripling X cuts Y by one-third, and halving X doubles Y. *Scaling constants* refer to adjustable parameters in a model that have no substantive meaning but are included to shift a variable from one metric to another. For example, to change meters to centimeters, we multiply the meters by the constant 100, which shifts the metric of length to centimeters. As we describe different functions below, we occasionally do so in terms of the concepts of concavity, proportionality, or scaling constants.

Logarithmic Functions

The logarithmic function (often referred to as the *log function*) has the general form $f(X) = \log_a(X)$, where a is a constant indicating the base of the logarithm. Logs can be calculated for different bases, such as the base 10, the base 2, or the base 8.

[2]Our discussion of functions and the graphic representations of them draws on concepts described by W. Mueller (see *www.wmueller.com/precalculus/index.html*).

The log base 10 of the number 100 is written as $\log_{10}(100)$, where the subscript is the base and the number in parentheses is the number for which you are calculating the log. If n stands for the number for which you are calculating the log, and a is the base of the log, then the log is the solution for b in the equation $n = a^b$. For $\log_{10}(100)$, we solve for b in the equation $100 = 10^b$, so the log base 10 of the number 100 is 2 (because $10^2 = 100$). The value of $\log_5(25)$ is 2 because $5^2 = 25$. Sometimes you will encounter a log expression with no base, such as log(1,000). When this happens, the log is assumed to have a base of 10. So log(1,000) = 3 (because $10^3 = 1{,}000$).

There is a special logarithm, called the natural log, that uses a constant called e as its base. The number e appears in many mathematical theories. Its value is approximately 2.71828. The number e was studied in depth by Leonhard Euler in the 1720s, although it was first studied by John Napier, the inventor of logarithms, in 1614. It has some remarkable mathematical properties (which we will not elaborate on here) and is referred to as Napier's constant. The natural log of a number is signified by the expression ln(n). For example, the natural log of 10 is signified by ln(10). It equals approximately $\log_{2.71828}(10) = 2.302585$.

Figure 8.6 presents sample graphs of log functions. When expressing the relationship between two variables, rather than using a linear function, one might use a log function. Log functions are sometimes used to model growth or change when the change is rapid at first and then slows down to a gradual and eventually almost nonexistent pace (see Figure 8.6a).

Log functions share many common features: (1) The logarithm is undefined for negative values of X (where X is the number for which you are calculating the log); (2) the value of the log can be positive or negative; (3) as the value of X approaches zero, the value of the log approaches negative infinity; (4) when $X = 1$, the value of the log is 0; and (5) as X approaches infinity, the log of X also approaches infinity. For the function $\log_a(X)$, the function output increases with increasing X if $a > 1$ and decreases with increasing X if a is between 0 and 1.

Exponential Functions

The exponential function has the general form $f(X) = a^X$. The function yields output that increases in value with increasing X if $a > 1$ and decreases in value with increasing X if a is between 0 and 1. Figure 8.7 presents examples of common exponential curves. These curves are often used to refer to growth, such as when people say a population is "growing exponentially." With exponential growth or change, the larger a population gets, the faster it grows. For decreasing exponential growth or change, the smaller the population gets, the more slowly it decreases in size. As it turns out, exponential functions are simply the inverse of log functions, so the two functions mirror image each other's properties. For exponential functions, if a is between –1 and 0, then the output value is a damped oscillation as X increases, and if a is < –1, it is an undamped oscillation as X increases (see the later section on trigonometric functions for a discussion of oscillation).

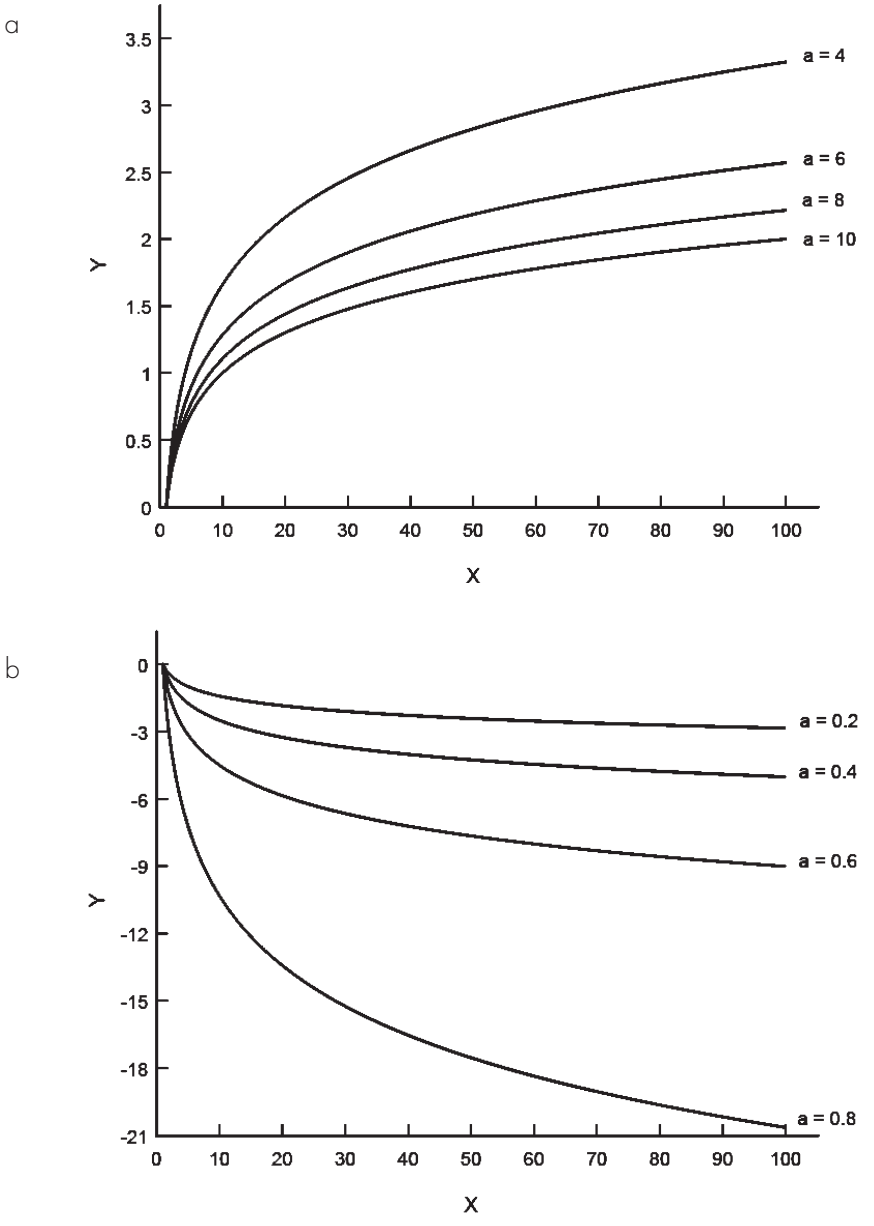

FIGURE 8.6. Graphs of log functions for $\log_a(X)$ with X ranging from 1 to 100. (a) $a > 1$; (b) $0 < a < 1$.

Social scientists often modify the exponential function to create functions that reflect growth or change with certain properties. For example, using the fact that any number raised to the power of 0 is equal to 1, the following equation can be used to describe exponential growth over time

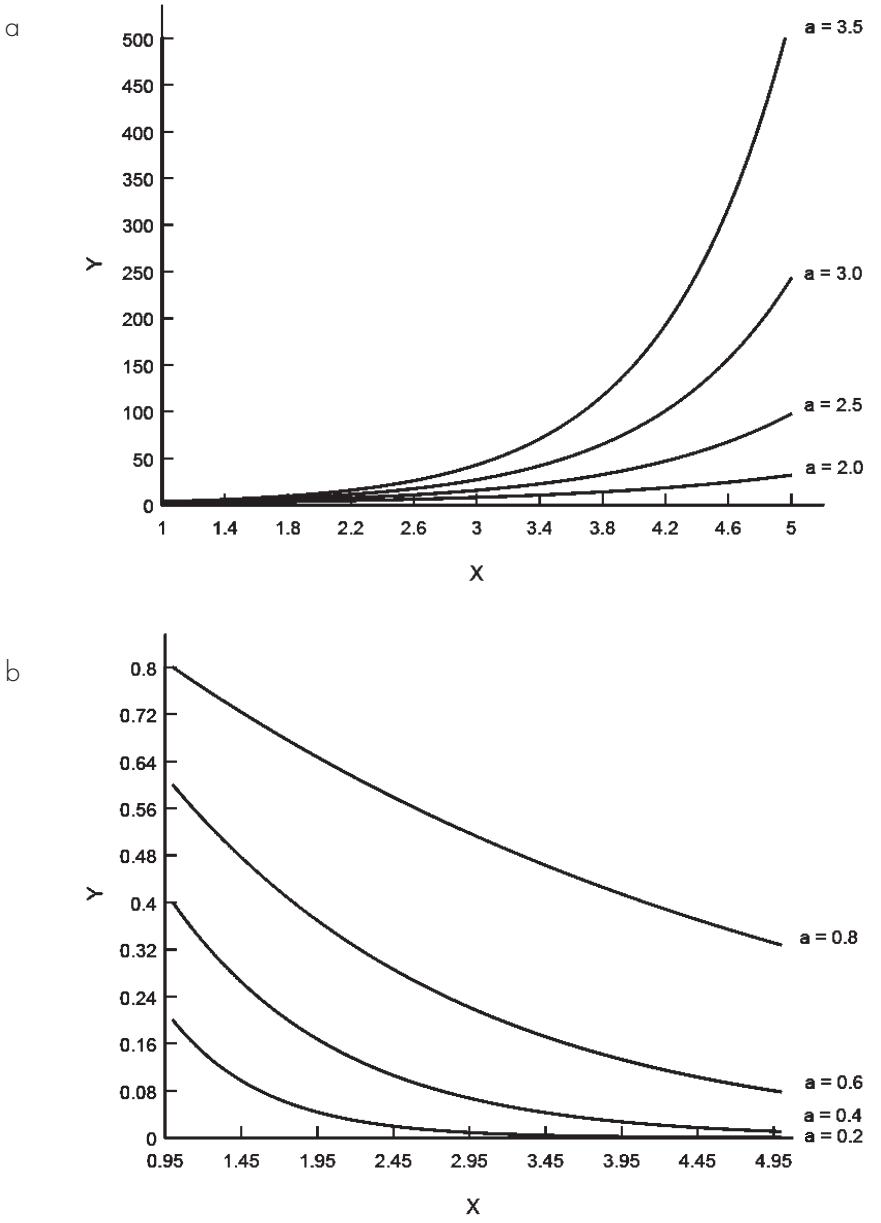

FIGURE 8.7. Graphs of exponential functions for a^X with X ranging from 1 to 5. (a) $a > 1$; (b) $0 < a < 1$.

$$Y = s_0 e^{kX}$$

where Y is the population size at a given point in time, X is the duration in time since a predetermined start time, and s_0, e, and k are constants. In this case, e is Napier's con-

stant. The value of s_0 is fixed at a value equal to the population size at the predetermined start time. Note that when $X = 0$, the population size will equal the population size at the start time (because any number raised to the power of zero is 1.0). For this function, Y increases geometrically with a doubling time equal to $0.6932/k$. A graph illustrating this function appears in Figure 8.8, where the starting size of a population is $s_0 = 5{,}000$, where $k = 0.333$ (yielding a doubling time of just over 2 years), and where X ranges from 0 to 5 years. When expressing the relationship between two variables, rather than using a linear or log function, one might use an exponential function, such as that illustrated in Figure 8.8.

Power Functions

Power functions have the general form $f(X) = X^a$ where a is an adjustable constant. For positive values of X greater than 1, when $a > 1$, the curve will be concave upward, and when a is between 0 and 1, the curve will be concave downward.

Power functions often have a similar shape to exponential and logarithmic functions, with the differences between the functions sometimes being subtle. When the difference is small, it does not matter which function is used to create the model. But differences can exist. Exponential functions increase by multiples over constant input intervals. Logarithms increase by constant intervals over input multiples. Power functions do not follow either of these patterns. Power curves eventually outgrow a logarithm and undergrow an exponential as X increases. A practical example of the function differences is the modeling of the spread of HIV, the virus that causes AIDS. During the

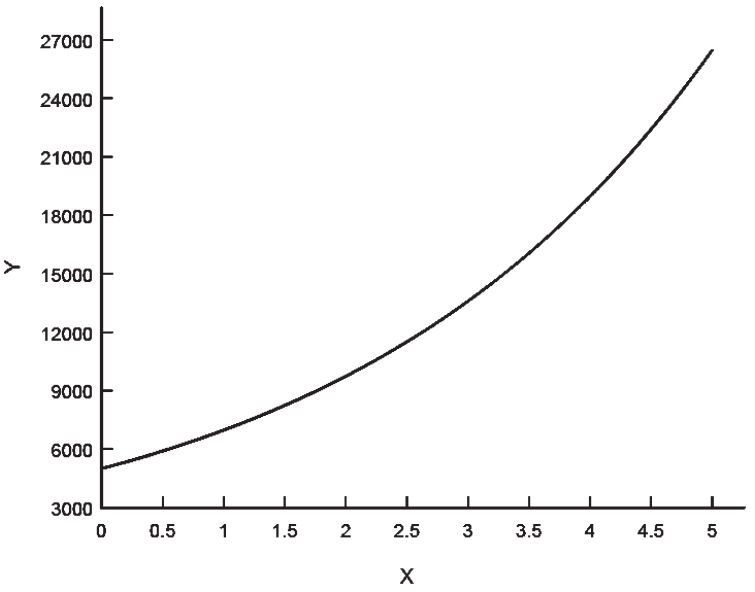

FIGURE 8.8. Exponential growth for $Y = s_0\, e^{kX}$.

early stages of the epidemic, it was thought that the number of HIV cases was growing exponentially, but in later analyses, the function was found to be better mapped by a power function. An exponential model yielded overestimates of the number of cases forecast in future years, which in turn led to overestimates of the required resources to deal with the epidemic (e.g., hospital space, medications; see Mueller, 2006).

Figure 8.9 presents some examples of power functions, and Figure 8.10 plots a power function and an exponential function on the same graph to illustrate some of these properties. When expressing the relationship between two variables, rather than using a linear, log, or exponential function, one might use a power function instead.

Polynomial Functions

Polynomial functions are simply the sum of power functions. The general form of a polynomial function is

$$f(X) = a + bX^1 + cX^2 + dX^3 + \ldots$$

where X continues to be raised to the next highest integer value, and each term has a potentially unique adjustable constant. Polynomials can model data with many "wiggles and turns," but the more wiggles and turns there are, the greater the number of power terms that are required to model it. Notice that when only a single term for X is used with a power of 1, the polynomial model reduces to a linear model. The adjustable constant a is typically viewed as a scaling constant. Adding one term to the linear model (i.e., adding the term cX^2) allows the model to accommodate a curve with one bend. A polynomial model with three terms ($a + bX^1 + cX^2 + dX^3$) will accommodate a curve with two bends. A polynomial model with four terms will accommodate a curve with three bends. In general, to accommodate k bends, you need $k + 1$ terms.

The most popular polynomial functions in the social sciences are the quadratic and cubic functions. They are defined as

$$\text{Quadratic: } f(X) = a + bX + cX^2$$
$$\text{Cubic: } \quad f(X) = a + bX + cX^2 + dX^3$$

Figures 8.11 and 8.12 provide an example of each type of curve, and Figure 8.13 provides an example of a polynomial function with eight terms. The quadratic model is effective for modeling U-shaped and inverted-U-shaped relationships as well as J-shaped and inverted-J-shaped relationships. The cubic function is effective for modeling S-shaped curves. In Figure 8.12b we manipulated the scaling constant, a, to create separation between curves so that you can better see the trends. Figure 8.13 illustrates how diverse a "curve" that large polynomials can create. When expressing the relationship between two variables, rather than using a linear, log, exponential, or power function, one might use a polynomial function instead.

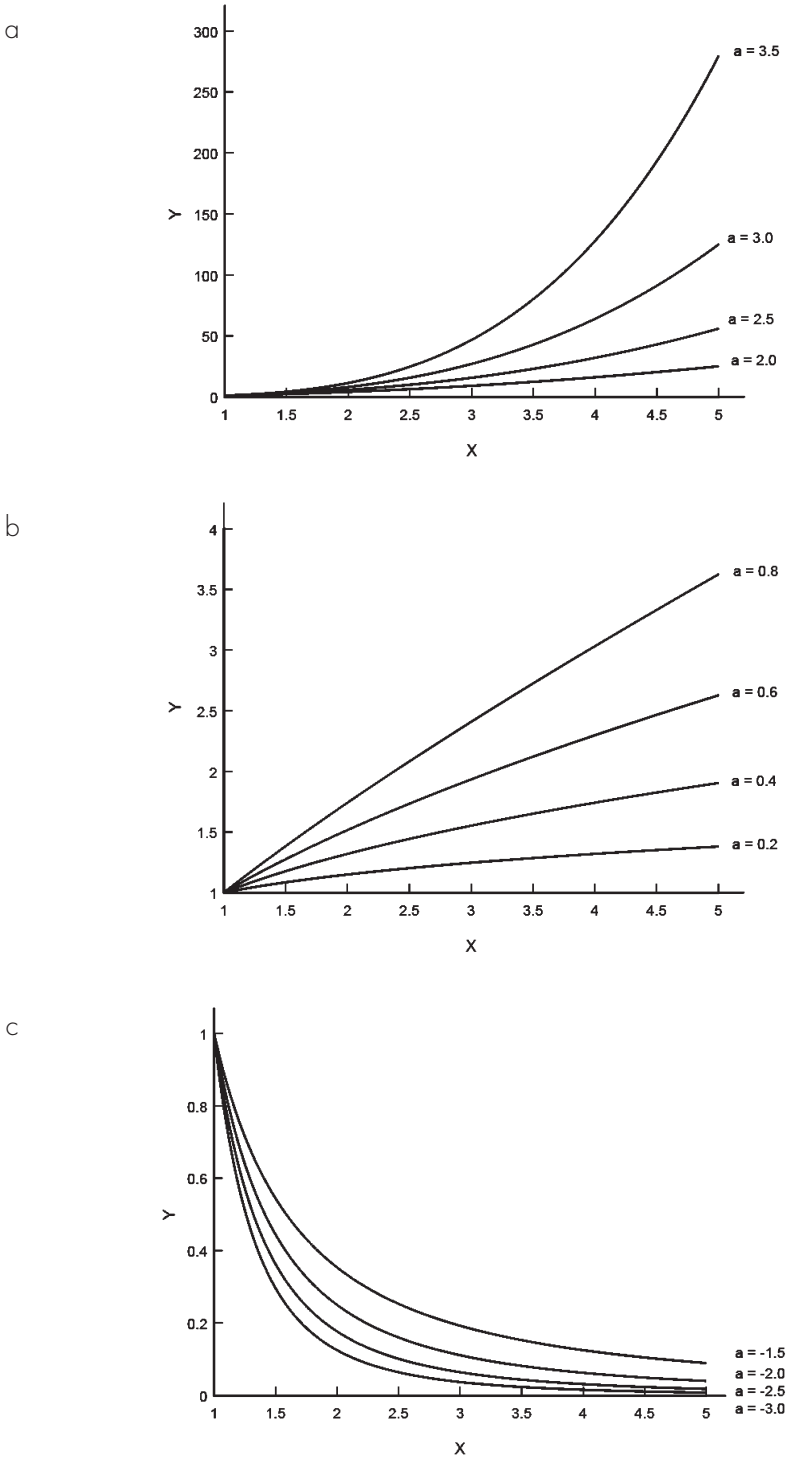

FIGURE 8.9. Graphs of power functions for X^a with X ranging from 1 to 5. (a) $a > 1$; (b) $0 < a < 1$; (c) $a < 0$.

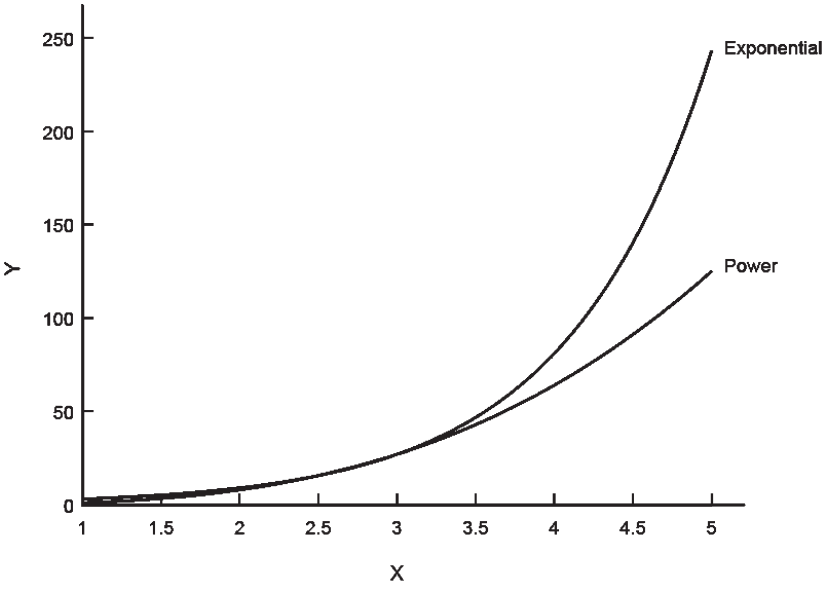

FIGURE 8.10. Example of power and exponential functions.

Trigonomic Functions

Trigonometric functions are typically used to model cyclical phenomena. The two most common functions are the sine function and the cosine function, which have the form $f(X) = \sin(aX)$ and $f(X) = \cos(aX)$, where a is a constant, *sin* is the sine, and *cos* is the cosine. The sine and the cosine functions repeat the values of their outputs at regular intervals as X increases. Simple transformations of the sine and cosine functions can reproduce many forms of periodic behavior. For example, some people have suggested that rhythmic cycles, called biorhythms, reflect active and passive phases in the physical aspects of everyday life. The phases of biorhythms are modeled using a sine function of the form $f(X) = \sin(.224*X)$, where X is the number of days since a baseline index is taken. Output values range from 1 to –1, with positive values indicating increasingly high energy and negative values indicating increasingly low energy. Figure 8.14 plots the output values for 120 days, starting at day 0. As noted earlier, certain types of cyclical phenomena also can be modeled using exponential functions with negative values of a in the expression $f(X) = a^X$.

Choosing a Function

In sum, there are a wide range of functions available to the math modeler for describing the relationship between variables, including linear functions, logarithmic functions, exponential functions, power functions, polynomial functions, and trigonometric functions, to name a few. We have only scratched the surface of the many strategies a

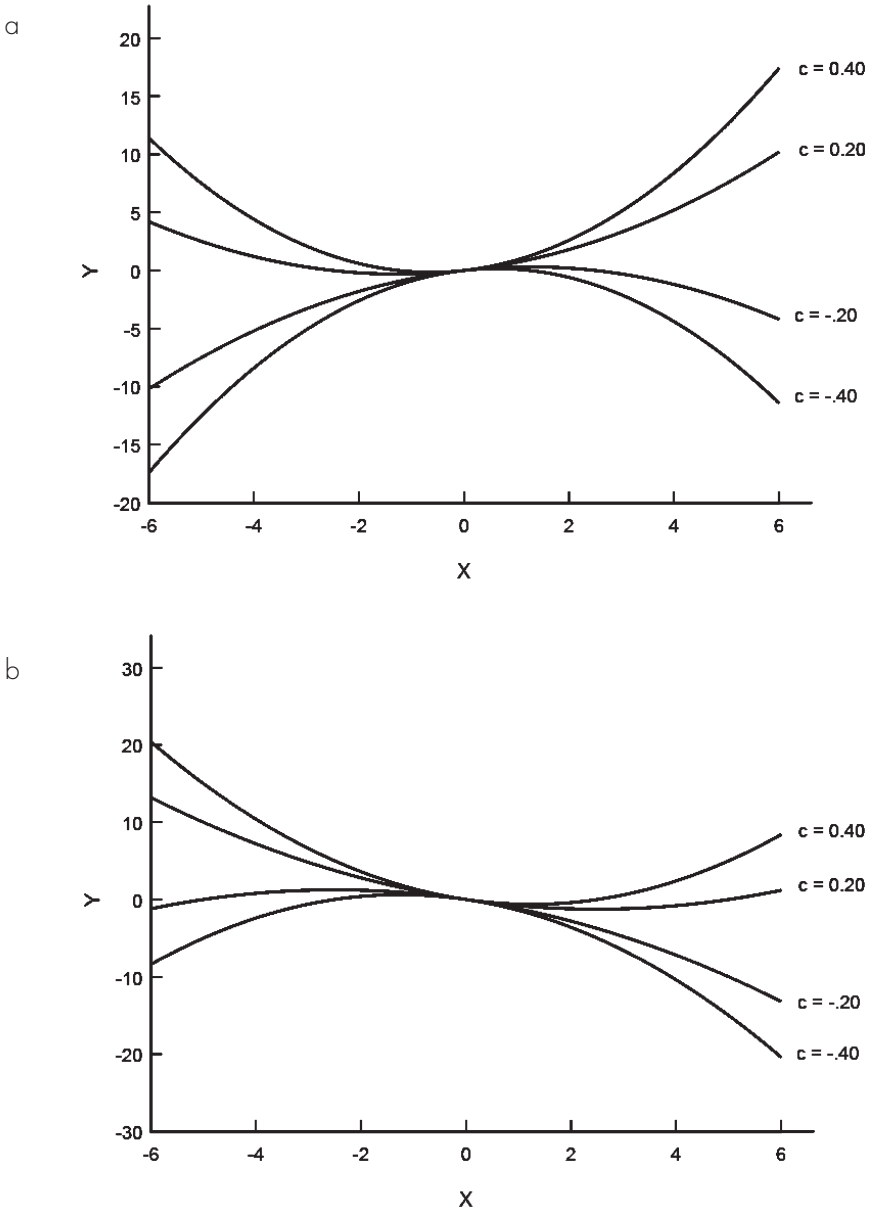

FIGURE 8.11. Quadratic functions. (a) Function $a + bX + cX^2$, with $a = 0$, $b = .5$; (b) function $a + bX + cX^2$, with $a = 0$, $b = -1$.

mathematical modeler can use. As you become familiar with functions and the curves they imply, you should be able to make informed choices about modeling relationships between variables. Mathematical modelers sometimes select functions for their models a priori, based on logic, and other times they make decisions about appropriate model

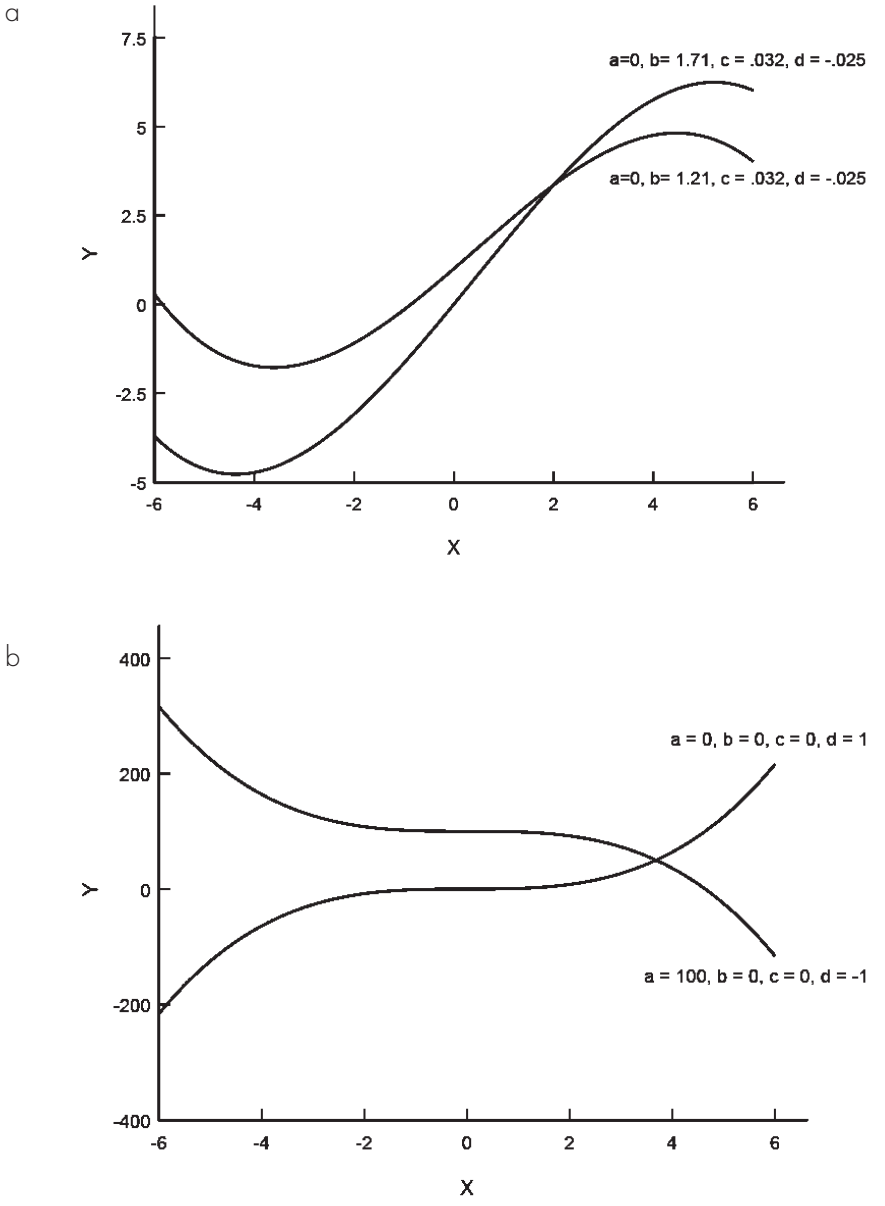

FIGURE 8.12. Cubic functions. (a) Function for $a + bX + cX^2 + dX^3$; (b) additional functions for $a + bX + cX^2 + dX^3$.

functions after collecting data and scrutinizing scatterplots. In the latter case, the model chosen and the values of the adjustable parameters are still subjected to future empirical tests, even though preliminary data are used to gain perspectives on appropriate functional forms. You can gain perspectives on the curves implied by different functions

by creating hypothetical data and applying the different functions to them. We provide information on how to do this using the statistical package SPSS in Appendix 8A to this chapter, and also provide information about other graphics software.

FUNCTIONS FOR CATEGORICAL VARIABLES

Thus far we have considered only functions involving quantitative variables, but functions also can be specified for categorical variables. Consider as a simple example the relationship between whether or not someone uses an umbrella as a function of whether or not it is raining. The relationship between these two categorical variables is expressed as follows

$$f(x) = \begin{cases} \text{umbrella}, & \text{if } x = \text{raining} \\ \text{no umbrella}, & \text{if } x = \text{not raining} \end{cases}$$

where one uses an umbrella if it is raining and one does not use an umbrella if it is not raining.

Sometimes mathematical modelers create quantitative representations of categorical variables and then analyze the quantitative translations using the quantitative functions described earlier. For example, one could specify a mathematical function relating the probability of carrying an umbrella to the probability of it raining, with both variables

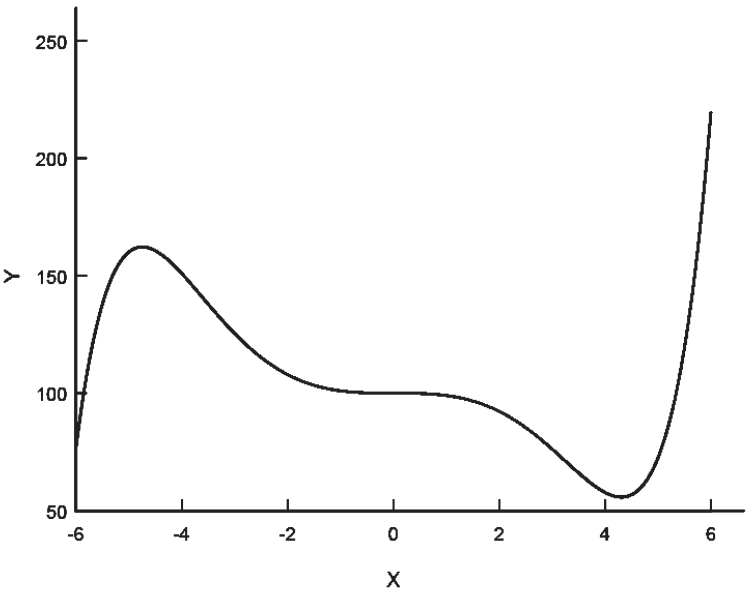

FIGURE 8.13. Polynomial function with seven terms.

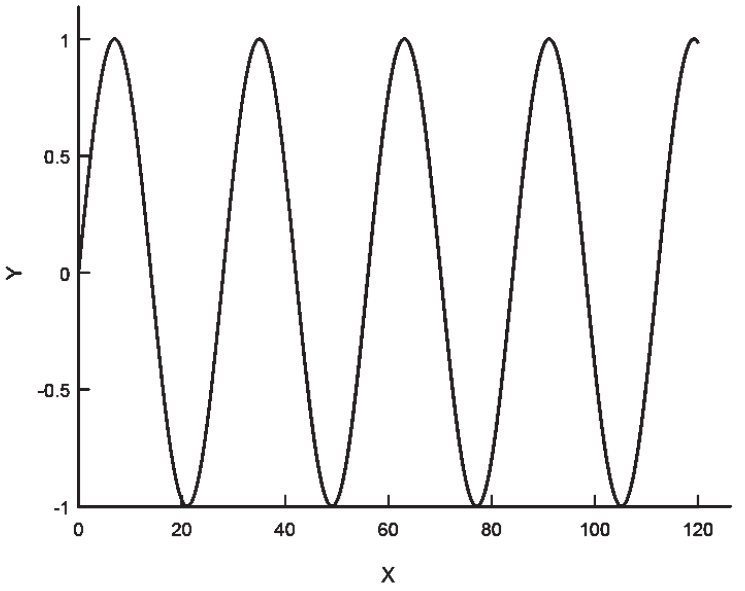

FIGURE 8.14. Sine function.

differing on a probability continuum of 0 to 1.0. The function might then be expressed as an exponential function, as in Figure 8.7a.

In some cases, functions involving categorical and quantitative variables are stated in terms of a table of values rather than symbolically. For example, suppose we specify whether someone is a Democrat or Republican as a function of scores on a 7-point index (e.g., response to a rating scale) of how conservative or liberal he or she is. The scale consists of integers ranging from –3 to +3, with increasingly negative scores signifying more conservativeness, increasingly positive scores signifying more liberalness, and the score of zero representing a neutral point. The function $Y = f(X)$ might be stated as

X	*Y*
–3	Republican
–2	Republican
–1	Republican
0	Democrat
1	Democrat
2	Democrat
3	Democrat

In this representation, the person is said to be a Republican if he or she has a value of –1, –2, or –3. Otherwise, the person is a Democrat.

Another approach to representing a function with a categorical variable is to use a graph. For example, the liberal–conservative and party identification function might be expressed as in Figure 8.15.

ADVANCED TOPICS: MANIPULATING AND COMBINING FUNCTIONS

One creative aspect of mathematical modeling is deriving new functions from old functions so as to create models that are better suited to describing the relationship between variables. We saw hints of this for polynomial functions (which combine power functions). Another class of functions, which we did not discuss, divides one polynomial function by a second polynomial function rather than summing polynomials. These are called *rational functions*. We provide illustrations of manipulating and combining functions here to show the flexibility available to the math modeler.

Function Transformations

One way of modifying functions is to add adjustable parameters to them. Given a function $f(x)$, one can add or subtract an adjustable parameter, a, after the rule described by $f(X)$ is applied: that is, $f(X) \pm a$. This has the effect of shifting the output values upward (in the case of addition) or downward (in the case of subtraction). These transformations are called *vertical shifts*. The output of a function also can be multiplied by the parameter a after the rule described by $f(X)$ is applied, that is, $a \times f(X)$. This transformation vertically stretches (when $a > 1$) or squeezes (when $a < 1$) the graph of the function. Such transformations are called *vertical stretches* or *vertical crunches*. Another possibility is to add or subtract a from $f(X)$ before the rule described by $f(X)$ is applied: that is, $f(X + a)$ or $f(X - a)$. These transformations typically move the graph of the function left when adding a positive value of a or right when subtracting a positive value of a. Such transformations are called *horizontal shifts*. Finally, one can multiply X before the rule described by $f(X)$ is applied; that is, $f(aX)$. These transformations horizontally stretch (when $a < 1$) or squeeze (when $a > 1$) the graph of the function. Such transformations are

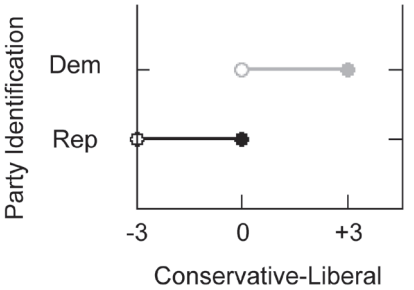

FIGURE 8.15. Graphical representation of a function with a qualitative variable.

called *horizontal stretches* or *horizontal crunches*. Coupled with the possibility of forming inverses for many functions, mathematical modelers have considerable flexibility in manipulating traditional functions with the use of vertical shifts, horizontal shifts, vertical stretches, vertical crunches, horizontal stretches, and horizontal crunches. If you begin your modeling efforts with a traditional function that is approximately correct in form, then transformations such as the above allow you to fine-tune the form of the curve to your problem. An example of this is the classic bounded exponential model, which we now consider.

Recall that the exponential function is $f(X) = a^X$. A simple set of modifications to this function produces what is called a *bounded exponential model*. This has the form

$$Y = a + (b - ce^{-X})$$

where a, b, and c are adjustable constants and e is Napier's constant. The term ce^{-X} is essentially an exponential function where $a = e$ and the exponent is multiplied by an adjustable constant, c. This creates a decaying exponential curve, which is then subtracted from a fixed upper bound or limit reflected by the value of b. As the decaying exponential dies out, the difference from b rises up to the bound. The parameter a is a scaling constant. This kind of function models growth that is limited by some fixed capacity. Figure 8.16 presents an example of this curve, as well as a traditional exponential curve.

Combining Functions

Another strategy that math modelers use is to combine functions. A popular function in the social sciences is the logistic function. It has the general form $f(X) = c/(1 + ae^{-bX})$ where a, b, and c are adjustable constants and e is Napier's constant. A logistic function is a combination of the exponential growth and bounded exponential growth functions that were illustrated in Figure 8.16. In the logistic function, exponential growth occurs when the function outputs for X are small in value. However, this turns into bounded exponential growth as the function outputs approach their upper bound. A logistic function is plotted in Figure 8.17. Note the shapes of the curve to the right and left of the broken line in Figure 8.17 and compare these with the curve shapes in Figure 8.16. The result of combining the exponential growth and the bounded exponential growth functions is an S-shaped curve. The logistic function is a special case of a broader function known as the *sigmoid function*, which generates curves having an S shape.

Combining multiple functions using processes such as those described for the logistic function is another tool available to math modelers. It is not uncommon for a theorist to break the overall relationship into a series of smaller component segments, specify a function to reflect each segment, and then assemble the component functions into a larger whole in one way or another.

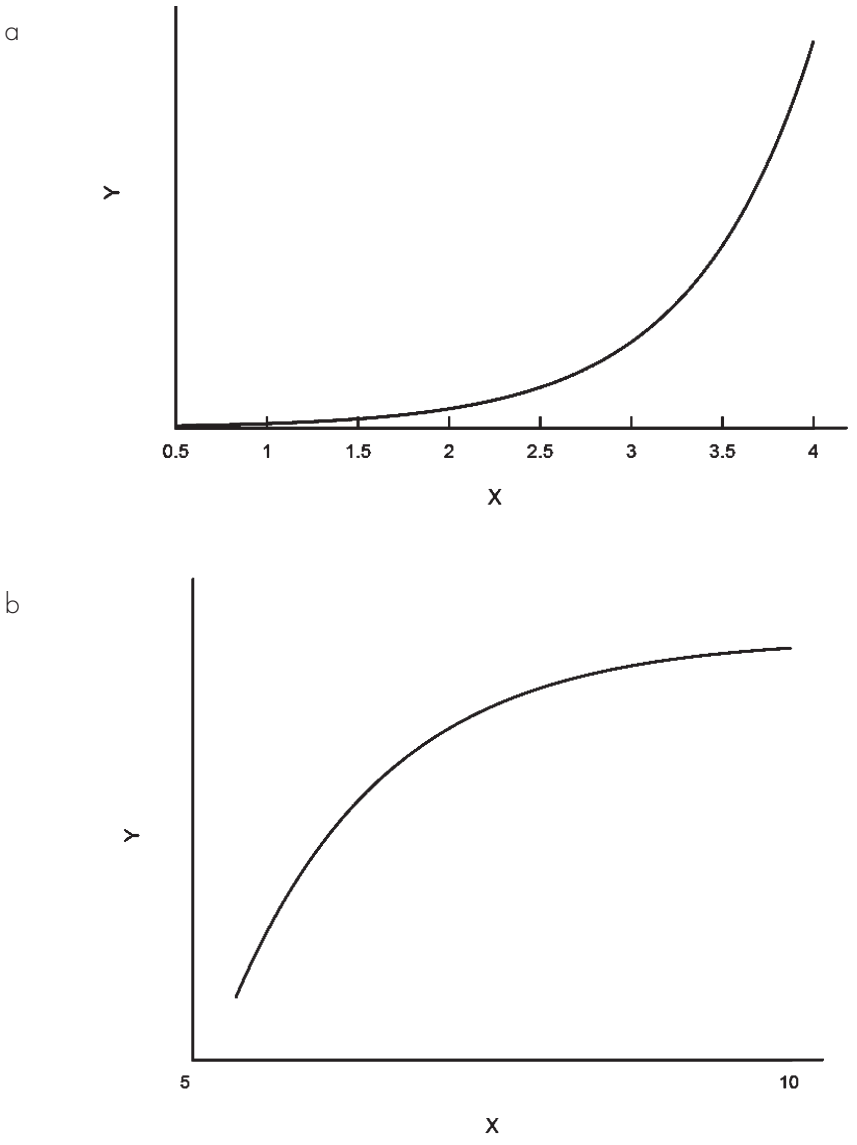

FIGURE 8.16. Exponential and bounded exponential model. (a) Exponential function; (b) bounded exponential function.

In sum, functions can be manipulated with adjustable constants in a variety of ways and subjected to vertical and horizontal stretching and crunching. Functions also can be combined to form even more complex functions (as in the case of the logistic function), and both quantitative and qualitative variables can be modeled. Traditional mathematical modeling opens up a wide range of tools for describing relationships for the theorist to consider.

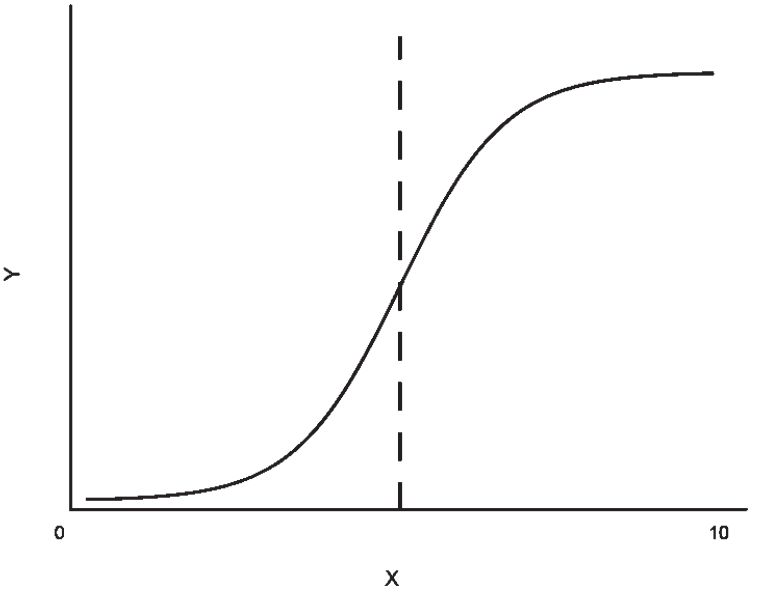

FIGURE 8.17. Logistic function.

MULTIPLE VARIABLE FUNCTIONS

All of the functions we have described use a single input variable. However, functions can involve more than one input variable, and the multiple variables can be combined in a wide variety of ways to yield output. For example, the traditional linear function for a single variable can be extended to include multiple variables (e.g., X and Z) using the following functional form

$$f(X, Z) = a + bX + cZ$$

where a, b, and c are adjustable constants. As another example, a multiplicative function might take the form

$$f(X, Z) = a + bXZ \tag{8.5}$$

where a and b are adjustable constants. Multiplicative models often are used to represent moderated relationships between quantitative variables, as discussed in Chapter 6 (see Jaccard & Turrisi, 2003).

Another example of a multiple variable function that we will make use of later is an averaging function. It takes the general form

$$f(X, Z) = [a/(a + b)]X + [b/(a + b)]Z$$

where a and b are adjustable constants. This model represents function output as a weighted average of X and Z. To see that the function captures a simple arithmetic average, set the values of a and b to 1. This produces

$$\begin{aligned} f(X, Z) &= [1/(1 + 1)]X + [1/(1 + 1)]Z \\ &= (1/2)X + (1/2)Z \\ &= (X + Z)/2 \end{aligned}$$

By allowing a and b to take on nonequal values (e.g., $a = 1$ and $b = 4$), one obtains a "weighted" average rather than a simple arithmetic average, such that the Z will contribute more to the average than X. We will take advantage of this property later when we apply math modeling to a substantive area.

PHASES IN BUILDING A MATHEMATICAL MODEL

Math modelers typically use four phases to construct a mathematical model. First, the modeler identifies the variables that will be included in the model and identifies the metrics on which the variables are measured. Textbooks on mathematical modeling tend to view the variables and metrics as givens and devote little attention to how the variables and metrics are selected. Of course, this is a nontrivial issue, and how one chooses the variables to include is the subject of much of this book. Second, the modeler thinks carefully about the variables, the metrics, and the relationships between the variables, and poses a few candidate functions that might capture the underlying dynamics. He or she might think about the implications of the functions and what predictions to make at both moderate and extreme input values. Eventually, a working function is settled upon, typically a function that includes several adjustable constants. Sometimes the values of the adjustable constants are logically determined, and the modeler fixes the constants at those values. More often than not, the values of the adjustable constants are estimated from data. Third, the modeler collects empirical data, estimates values of the adjustable constants from the data if necessary, and examines the degree of fit between the output values of the function and the values observed in the real world. At this point, if performance of the model is unsatisfactory, a new function might be tried or the original function might be modified to accommodate the disparities. Fourth, given revisions of the function, the model is applied to a new set of data to determine how well the revised model performs. If the model does a good job of reproducing observations in the real world and if the model makes conceptual sense, it will be selected as the model of choice.

This, of course, is an oversimplification of the process that unfolds in building math models, and there are many variants of it that depend on the parameters of the task at hand. Our main point is that building math models is usually a dynamic process that involves much more than simply specifying a function.

AN EXAMPLE USING PERFORMANCE, ABILITY, AND MOTIVATION

Educational researchers have long argued that performance in school is a function of two factors: a student's motivation to perform well and his or her ability to perform well. This relationship is often expressed in the form of a multiplicative model, as follows:

$$\text{Performance} = \text{Ability} \times \text{Motivation} \tag{8.6}$$

The basic idea is that if a student lacks the cognitive skills and capacity to learn, then it does not matter how motivated he or she is; school performance will be poor. Similarly, a student can have very high levels of cognitive skills and the ability to learn, but if the motivation to work and attend to the tasks that school demands is low, then performance will be poor. The multiplicative relationship reflects this dynamic because, for example, if motivation is zero, then it does not matter what a person's score on ability is—his or her performance will always equal zero. Similarly, if ability has a score of zero, it does not matter what a person's motivation score is—his or her performance will always equal zero. Although this makes intuitive sense, the dynamics might be different from those implied by Equation 8.6, as we will now illustrate.

Our first step is to specify the metrics of the variables involved, since they do not have natural metrics. Performance in school might be indexed for individuals using the familiar grade-point average metric that ranges from 1.0 (all F's) to 4.0 (all A's), with decimals rounded to the nearest tenth (e.g., 2.1, 3.5). Ability might be indexed using a standard intelligence test that has a mean of 100 and a standard deviation of 15. Motivation might be indexed using a 10-item scale that asks students to agree or disagree with statements such as "I try hard in school" and "Doing my best in school is very important to me." A 5-point agree–disagree rating scale (1 = strongly disagree, 2 = moderately disagree, 3 = neither agree nor disagree, 4 = moderately agree, and 5 = strongly agree) provides the range of possible responses. The responses to each item are summed to yield an overall score from 10 to 50, with higher scores indicating higher levels of motivation.

Note that none of the metrics takes on a value of zero. Hence, the dynamic of having "zero" ability or "zero" motivation discussed above cannot occur. Indeed, one might question whether there is such a thing as "zero" intelligence (i.e., a complete absence of intelligence). Is a psychological zero point on this dimension even possible? Suppose we decide that although a complete absence of intelligence is not theoretically plausible, a complete absence of motivation to do well in school is plausible. One way of creating a motivation metric with a zero point is to subtract a score of 10 from the original motivation metric. Before this operation, the motivation metric ranged from 10 to 50. By subtracting 10 from the metric, it now ranges from 0 to 40, which includes a zero point.

However, there is a problem with this strategy. Just because we can mathematically create a zero score on the motivation scale by subtracting 10 from it, this does not mean that the score of zero on the transformed scale reflects a complete absence of motivation on the underlying dimension of motivation. What evidence do we have that this is indeed the case? Perhaps a score of zero on the new metric actually reflects a somewhat

low level of motivation but not a complete absence of it. The issue of mapping scores on a metric onto their location on the underlying dimension they represent is complex, and consideration of how to accomplish this is beyond the scope of this book. We will work with the original metric of 10–50 and not make explicit assumptions about where on the underlying motivation dimension these scores locate individuals. We suspect that, based on the content of the items, students who score near 50 are very highly motivated to perform well, and students who score near 10 are very low in (but not completely devoid of) motivation to perform well. But a separate research program is required to establish such assertions (Blanton & Jaccard, 2006a).

Suppose that a student has a score of 100 on the IQ test and a score of 30 on the motivation test. Using Equation 8.6, multiplying the ability score by the motivation score, we obtain 100 × 30 = 3,000, and we would predict a GPA of 3,000! Of course, this is impossible because a student's GPA can range only from 1.0 to 4.0. We need to introduce one or more adjustable constants to Equation 8.6 to accommodate the metric differences and to make it so that a predicted GPA score falls within the 1.0–4.0 range. For example, if we let *P* stand for performance, *A* for ability, and *M* for motivation, then we can allow for the subtraction of a constant from the product to make an adjustment in metric differences, modifying Equation 8.6 as follows

$$P = (A)(M) + a$$

where *a* is an adjustable constant whose value is estimated from data. Note, for example, if $a = -2{,}997$, then this is the same as subtracting 2,997 from the product of *A* and *M*. But perhaps subtracting a constant is not enough to account for the metric differences. For example, a score of 120 on the IQ test coupled with a score of 50 on the motivation test would yield a product value of 6,000, and subtracting a value of 2,997 from it would still produce a nonsensical GPA. A second scalar adjustment we might use is to multiply the product term by a fractional adjustable constant, which yields the general equation

$$P = b(A)(M) + a$$

where *b* is a second adjustable constant (in this case, *b* would be a fraction) designed to deal further with the metric differences. Its value also is estimated from data. The terms on the right-hand side of this equation can be rearranged to yield

$$P = a + b(A)(M) \tag{8.7}$$

If you compare Equation 8.7 with Equation 8.3, you will note that Equation 8.7 is simply a linear function, so performance is assumed to be a linear function of the product of (*A*)(*M*). Not only do the constants *a* and *b* take into account the different metrics, but the value of *b* also provides substantive information as well; namely, it indicates how much change in performance (GPA) one expects given a 1-unit increase in the value of the product term (*A*)(*M*).

Figure 8.18 plots the relationship between performance and motivation at three different levels of ability based on Equation 8.7, where values of a and b have been empirically determined from data collected for a sample of 90 students. In this example, $a = -2.0$ and $b = .0015$. The slope of P on M for any given value of A is bA. There are several features of this plot worth noting. First, note that the effect of motivation on performance is more pronounced as ability increases. This is evident in the steeper slope ($bA = .165$) for the two variables when the ability score is 110 as compared with the slope when the ability score is 100 ($bA = .150$), and, in turn, as compared to the slope when the ability score is 90 ($bA = .135$). These differences in slope may seem small but they are probably substantial. For example, when the ability score is 110, a 10-unit change in motivation is predicted to yield a (.165)(10) = 1.65-unit change in GPA; when the ability score is 100, a 10-unit change in motivation is predicted to yield a (.150)(10) = 1.50-unit change in GPA; when the ability score is 90, a 10-unit change in motivation is predicted to yield a (.150)(10) = 1.50-unit change in GPA.

Second, note that at each of the different levels of ability (90, 100, and 110), the relationship between motivation and performance is assumed to be linear. Is this a reasonable assumption? Perhaps not. Perhaps the relationship between performance and motivation at a given ability level is better captured by an exponential function in the form of one of the curves in Figure 8.7a. For example, when motivation is on the low end of the motivation metric, increasing it somewhat may not have much impact on performance—it will still be too low to make a difference on performance. But at higher levels of the motivation metric, increasing it will have an impact on performance. This dynamic is captured by the exponential functional forms illustrated in Figure 8.7a. Or

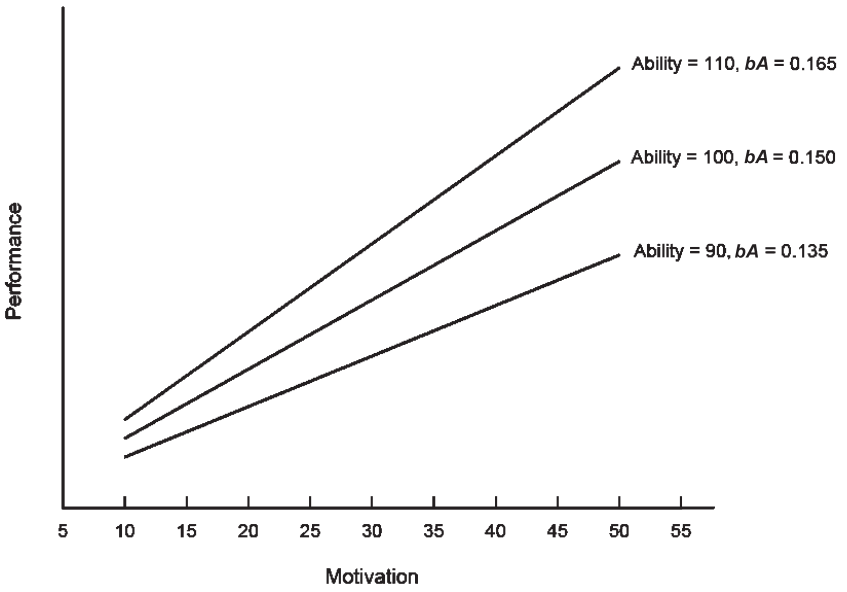

FIGURE 8.18. Example for Performance = Ability × Motivation.

perhaps a power function in the form of one of the curves in Figure 8.9a is applicable. Power functions have the same dynamic as the exponential function in Figure 8.7a, but they "grow" a bit more slowly. Or perhaps an S-shaped function in the form of the curve in Figure 8.17 applies, with floor and ceiling effects on performance occurring at the low and high ends of motivation, respectively.

The multiplicative model specified by Equation 8.6 assumes what is called a *bilinear interaction* between the predictor variables; that is, it assumes that the relationship between the outcome and one of the predictors (in this case, motivation) is always linear no matter what the value is of the other predictor (in this case, ability). To be sure, the value of the slope for the linear relationship between P and M differs depending on the value of A (as noted earlier), but the function form is assumed to be linear. One can modify the model to allow for a nonlinear relationship between performance and motivation at different levels of ability, say, in accord with a power function, as follows

$$P = a + b(A)(M^c) \tag{8.8}$$

where c is an adjustable constant whose value is estimated from data. This model allows for the possibility of a function form like those of Figure 8.9a.

Another notable feature of Figure 8.18 is that at the lowest value of motivation, there is a small degree of separation between the three different lines. The amount of separation between the lines reflects the differences in the effect of ability (at values of 90 vs. 100 vs. 110) on performance when motivation is held constant at the same value. But perhaps the amount of separation should be a bit more or a bit less than what is modeled in Figure 8.18. Equation 8.9 can be further modified to allow for a different amount of separation between the lines than what Equation 8.8 implies, as follows:

$$P = a + b(A)(M^c) + dA \tag{8.9}$$

where d is an adjustable constant whose value is estimated by data. The logic of adding this term is developed in Appendix 8B and is not central to our discussion here. The main points we want to emphasize are the following:

1. The rather simple theoretical representation in Equation 8.6 has nontrivial conceptual ramifications by specifying that the relationship between performance and the predictor variables is captured by the dynamics of a bilinear interaction when, in fact, the interaction may have a different functional form.
2. When building a mathematical model, the metrics of the variables usually have to be addressed (although our next example illustrates a case where this is not necessary).
3. There may be multiple features of the model (e.g., the separation between curves at different levels of one of the component terms as well as the shape of these curves) that must be specified that are not always apparent in simple representations such as Equation 8.6.

The fact is that the often presented model of Performance = Ability × Motivation is poorly specified, and applying principles of mathematical modeling helps to produce a better-specified theory that makes implicit assumptions explicit and highlights complexities that should be taken into account. Appendix 8B develops modeling strategies for this example in more detail and illustrates a substitution principle for building mathematical models. For more discussion of the assumptions of bilinear interactions, see Jaccard and Turrisi (2003).

AN EXAMPLE USING COGNITIVE ALGEBRA

Another example of using mathematical models to represent social phenomena involves models of cognitive algebra. This example illustrates how the implications of a mathematical representation can be pursued without recourse to such things as adjustable constants and complex modeling of data.

Suppose we describe the personal qualities of a political candidate to a person that he or she has not heard of by providing the person with three pieces of information. Suppose that the three pieces of information are all quite positive (e.g., the candidate is said to be honest, smart, and empathic). For purposes of developing this example, suppose we can characterize how positive each piece of information is considered to be using a metric that ranges from 0 to 10, with higher numbers reflecting higher degrees of positivity. We refer to the positivity of a piece of information as P_k, where k indicates the specific piece of information to which we are referring: P_1 refers to the perceived positivity of the first piece of information, P_2 refers to the perceived positivity of the second piece of information, and P_3 refers to the perceived positivity of the third piece of information. Suppose we want to predict how favorable a person will feel toward the candidate based on these three pieces of information. If we let F refer to a person's overall feeling of favorability toward the candidate, with higher values indicating higher levels of favorability, then one model that describes the impact of the information is the following:

$$F = P_1 + P_2 + P_3 \tag{8.10}$$

This model is a simple summative function that specifies that the overall feeling of favorability toward the candidate is the sum of the judged positivity of each individual piece of information (we ignore, for the moment, the metric of F and the issue of adjusting for metric differences). Equation 8.10 can be stated in more general form using summation notation as follows:

$$F = \sum_{i=1}^{k} P_i$$

where k is the number of pieces of information, in this case 3.

Now suppose that instead of a summative function, an averaging function is operat-

ing. That is, the overall feeling of favorability is the *average* of the positivity of the information presented rather than the sum of it. In this case, Equation 8.11 becomes

$$F = (P_1 + P_2 + P_3)/3 \tag{8.11}$$

and this can be represented more generally in summation notation as

$$F = \left(\sum_{i=1}^{k} P_i \right) \Big/ k \tag{8.12}$$

What are the implications of specifying the function as being summative versus averaging in form? It turns out, they are considerable. Let's explore the summation model first. Suppose a person judges the positivity values of the three pieces of information as 8, 8, and 8, respectively. The overall feeling of favorability toward the candidate will be 8 + 8 + 8 = 24. Now suppose we describe a second candidate to this person using the same three pieces of information but we add a fourth descriptor to them (cunning), that is judged to have a positivity value of 4. According to the summation model, the overall feeling of favorability toward this new candidate will be 8 + 8 + 8 + 4 = 28, and the person will prefer the second candidate to the first candidate. Psychologically, it is as if the second candidate brings all the same qualities as the first candidate (i.e., P_1, P_2, and P_3) and then "as a bonus," you get a fourth positive attribute as well (P_4). Hence, the person prefers the second candidate to the first candidate.

Now consider instead the averaging function. The overall feeling toward the first candidate is predicted to be (8 + 8 + 8)/3 = 8.0 and the overall feeling toward the second candidate is said to be (8 + 8 + 8 + 4)/4 = 7.0. In the averaging model, exactly the reverse prediction is made in terms of candidate preference; namely, the person now will prefer the first candidate to the second candidate. Psychologically, the first candidate has nothing but very positive qualities, whereas the second candidate has very positive qualities but also some qualities that are only somewhat positive. The person prefers the first candidate, who has nothing but very positive qualities, to the second candidate, who has very positive qualities but also moderately positive qualities.

Which function better accounts for the impressions people form? It turns out that this can be evaluated in a simple experiment in which two candidates would be described, one with three very positive qualities (Candidate A) and a second with three very positive qualities and a fourth moderately positive quality (Candidate B). Participants would then be asked to indicate which of the two candidates they prefer. The summation model predicts that participants should prefer Candidate B to Candidate A, whereas the averaging model predicts that participants should prefer Candidate A to Candidate B. One can differentiate the two models empirically by conducting the above experiment and determining which candidate tends to be preferred. This is a simple experiment without complex modeling. If the results showed that people tend to prefer Candidate A to Candidate B, then this would be consistent with (but not proof of) a summative process rather than an averaging process. If the results showed that people tended to prefer Candidate B to Candidate A, then this would be consistent with (but not

proof of) an averaging process rather than a summative process. Which process operates has implications for the design of political campaigns and advertising strategies to sell products. For example, if an advertising campaign adds to a person's cognitions a moderately positive piece of information about a product that is already quite positively evaluated, in the case of the averaging model, the advertisement should backfire and lower evaluations of the target product, thereby adversely affecting sales.

The literature on impression formation has extended these simple models of "cognitive algebra" to more complex model forms. For example, it is almost certainly the case that some information is more important to people in forming impressions than other information. As such, it makes sense to weight each piece of information by its importance to the individual. Equation 8.10 can be modified to include such weights, as follows:

$$F = w_1P_1 + w_2P_2 + w_3P_3 \tag{8.13}$$

where w_i is the importance of information i to the individual. Note that Equation 8.10 is a special case of Equation 8.13, namely the case where $w_1 = w_2 = w_3 = 1$. Expressed in summation notation, Equation 8.13 can be represented as

$$F = \sum_{i=1}^{k} w_iP_i \tag{8.14}$$

For the averaging model, introducing importance weights yields the following:

$$F = w_1P_1 + w_2P_2 + w_3P_3/(w_1 + w_2 + w_3) \tag{8.15}$$

Note that Equation 8.10 is a special case of Equation 8.15, namely the case where $w_1 = w_2 = w_3 = 1$. Equation 8.15 can be restated using summation notation as

$$F = \sum_{i=1}^{k} w_iP_i \Big/ \sum_{i=1}^{k} w_i \tag{8.16}$$

By extending the logic of algebraic models to the domain of "cognitive algebra" (which uses the premise that mental operations can be modeled by simple algebra), a great many insights into human information processing have been gained. Some of this research has involved simple experiments that pit competing predictions of different algebraic models against one another, whereas other research has taken the path of more complex math modeling with adjustable constants, error terms, and the like.

Parenthetically, the research literature finds support for both the summation and averaging models. In some contexts, people average the implications of information, whereas in other contexts, they sum it. There also are individual differences in these tendencies, with some people tending to average information in general whereas others tend to sum it in general. There are contexts for which simple summation or averaging models do not hold, and more complex combinatorial models are required to capture the

integration dynamics. Interested readers are referred to Anderson (1981) and Fishbein and Ajzen (1975).

AN EXAMPLE USING ATTITUDE CHANGE

As a third example of a mathematical model, we consider a model of attitude change from the communication literature that was developed by Fishbein and Ajzen (1975). The model concerns the case where a source is trying to persuade the recipient of a persuasive message to change his or her belief in something. A belief is conceptualized as a subjective probability that ranges from 0 to 1.00, much like a probability in mathematics. For example, people might believe with a probability of 0.20 that they will contract lung cancer if they smokes cigarettes. Or people might believe with a probability of 0.30 that a particular brand of toothpaste is the best for fighting tooth decay. In the model there are three probabilities that are of interest: (1) the subjective probability that the recipient holds prior to receiving the persuasive message, P_0, (2) the position that the recipient perceives the source takes in his or her persuasive message, also reflected by a subjective probability, P_S, and (3), the subjective probability of the recipient *after* hearing the persuasive message, P_1. For example, the recipient might have an initial belief corresponding to a subjective probability of 0.20, perceive the source as arguing that the target belief should have a subjective probability of 0.70, and after hearing the arguments of the source, the recipient revises his or her subjective probability to be 0.60. These three variables, P_0, P_S, and P_1, are measured variables in the theoretical system.

The amount of belief change that occurs is the difference in subjective probabilities before and after the message, or $P_1 - P_0$. It is the central outcome variable. Fishbein and Ajzen were interested in understanding factors that impact how much belief change occurs, so they constructed a mathematical model to reflect the underlying dynamics. Let *BC* represent belief change and be formally defined as $P_1 - P_0$. Fishbein and Ajzen begin by assuming that the amount of belief change that occurs is a function of the discrepancy between the recipient's initial position and the perceived position of the source—that is, $P_S - P_0$. If a source argues in favor of the exact same position of the recipient, then $P_S - P_0 = 0$, and no belief change will occur. It is only when the source takes a position that is discrepant from the recipient's that belief change can occur. The more discrepant the position taken by the source relative to the recipient's initial position, the greater the potential for belief change. We thus begin with a simple model based on a difference function:

$$BC = (P_S - P_0) \tag{8.17}$$

Not everyone will accept the arguments in a persuasive message. People differ in the likelihood that they will accept a message, with some people having a low probability of message acceptance, others having a moderate probability of message acceptance, and still others having a high probability of message acceptance. Fishbein and Ajzen

introduced a parameter into the model to reflect the probability that a recipient would accept the arguments of a message; this parameter is signified by P_A. Equation 8.17 thus becomes

$$BC = P_A(P_S - P_0) \tag{8.18}$$

with the constraint that P_A must range from 0 to 1.0 to reflect a probability metric. If a person completely accepts the message, then P_A = 1.00 and the amount of belief change will equal the discrepancy between the recipient's initial position and the position the recipient perceives the source as taking. If a person completely rejects the message, then P_A = 0.00 and there is no belief change. If the person is somewhat accepting of the source's message (i.e., P_A is somewhere between 0.00 and 1.00) then the amount of belief change is proportional to P_A.

Next, Fishbein and Ajzen address factors that impact the probability of acceptance of a message. One important factor is how discrepant the message is from the recipient's initial position. In general, people are more likely to accept messages that argue in favor of their existing beliefs as opposed to messages that argue against their existing beliefs. If we let D represent the absolute discrepancy between the recipient's initial position and the perceived position of the source (i.e., $D = \left|P_S - P_0\right|$), then the probability of acceptance can be modeled as

$$P_A = (1 - D) \tag{8.19}$$

Note that when D = 0, the source is arguing the same position that the recipient already believes and the probability of acceptance is 1.00. As the source's message becomes increasingly discrepant from the recipient's initial position, the probability of acceptance decreases to a minimum of 0.00.

Fishbein and Ajzen recognized that there are factors that can facilitate the acceptance of a message independent of message discrepancy. For example, if the source is a trustworthy and credible person, the exact same message may be more likely to be accepted than if the source is untrustworthy or lacks credibility. Fishbein and Ajzen introduced an adjustable constant to reflect these facilitating conditions, which they labeled f. Equation 8.19 was modified to appear as

$$P_A = (1 - D)^{1/f} \tag{8.20}$$

with the constraint that f be greater than 0. Fishbein and Ajzen thus use a power function to capture the underlying dynamics, where $1/f$ is an adjustable constant. Figure 8.19 presents sample curves for the probability of acceptance as a function of D at different values of f. Note that when f = 1, the relationship between the probability of acceptance and message discrepancy is linear with an intercept of 0 and a slope of 1. As f exceeds 1, the probability of acceptance increases rapidly at lower levels of discrepancy and remains high even as message discrepancy increases. As f decreases in value from 1, the

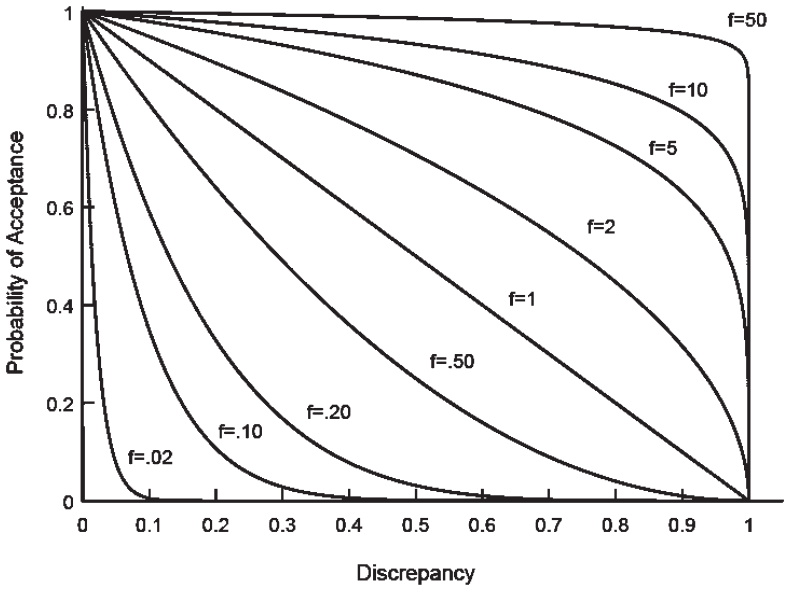

FIGURE 8.19. Fishbein and Ajzen model for probability of acceptance.

probability of message acceptance decreases rapidly at lower levels of discrepancy and remains low as message discrepancy increases.

Equations 8.18 and 8.20 can be combined to yield a single equation. Starting with Equation 8.19, we have

$$BC = P_A(P_S - P_0)$$

Since $PA = (1 - D)^{1/f}$, we can substitute the right-hand side of Equation 8.20 for PA, which yields

$$BC = (1-D)^{1/f}\,(P_S - P_0)$$

and since $D = |P_S - P_0|$, further substitution yields

$$P_1 = (1 - (|P_S - P_0|))^{1/f}\,(P_S - P_0)$$

The belief that a person has after hearing a persuasive message can be further specified by recognizing that $BC = (P_1 - P_0)$, so that if we subtract P_0 from both sides of the equation, we obtain

$$BC = [(1-(|P_S - P_0|)^{1/f}\,(P_S - P_0)] - P_0 \qquad (8.21)$$

Equation 8.21 is a mathematical model that predicts the belief that someone holds

after hearing a persuasive message. Although it may appear a bit intimidating to the mathematically uninitiated, it is based on reasonable communication principles and is reasonably precise in the functional forms it posits. The model makes use of observed measures as well as adjustable constants and incorporates a power function. In empirical applications, $P0$, $P1$, and P_S are measured variables and f is an adjustable constant whose value is estimated from data. The value of f is expected to vary across contexts where factors that facilitate message acceptance vary. For further discussion of this model and its implications, see Fishbein and Ajzen (1975).

AN EXAMPLE USING A TRADITIONAL CAUSAL MODEL

Another example of mathematical modeling in the social sciences is captured by an approach called structural equation modeling (SEM). Although some scientists do not think of structural equation models as mathematical models, they have all the characteristics of mathematical models as described in this chapter. To be sure, they are stochastic rather than deterministic, but their essence is mathematical in nature. The causal model we represent mathematically is presented in the path diagram in Figure 8.20. The model includes disturbance terms (because it is stochastic). We use generic labels for the variables for ease of notation. For this example, we assume that all of the relationships are linear, which is a typical assumption in SEM applications. Each endogenous variable is assumed to be a linear function of all variables that have an arrow

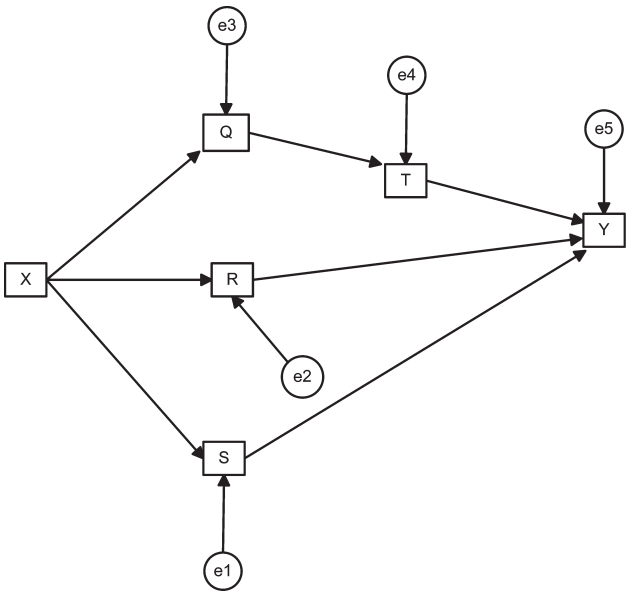

FIGURE 8.20. Causal model.

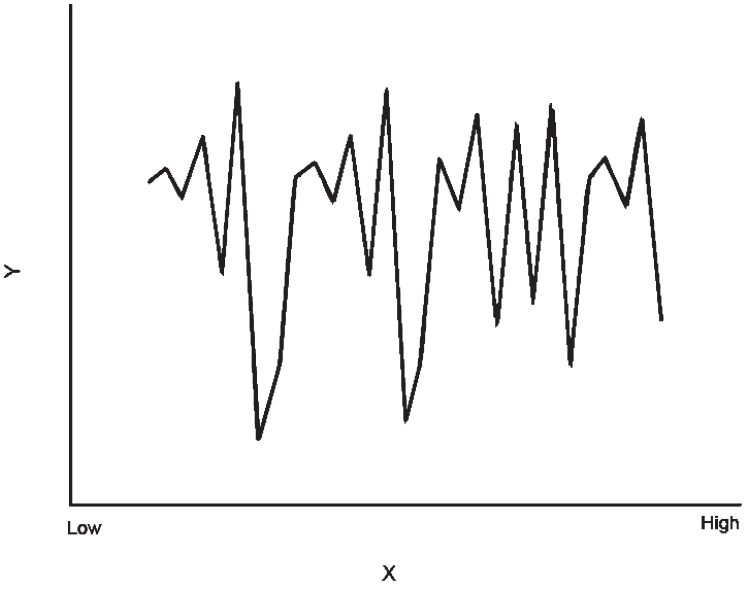

FIGURE 8.21. Chaos theory example.

going directly to it. The model can be expressed as a set of linear equations that are as follows:

$$
\begin{aligned}
Y &= a_1 + b_1T + b_2R + b_3S + e_5 \\
T &= a_2 + b_4Q + e_4 \\
Q &= a_3 + b_5X + e_3 \\
R &= a_4 + b_6X + e_2 \\
S &= a_5 + b_7X + e_1
\end{aligned}
$$

where a_1 through a_5 are adjustable constants representing intercepts, b_1 through b_7 are adjustable constants representing slopes, and e_1 through e_5 are error (or disturbance) terms. The equations yield a model that is overidentified, although constraints must be introduced for estimating the parameters in the presence of the error terms, and other ancillary modeling details must be attended to as well (see Bollen, 1989).

The adjustable constants for the slopes in this model reflect the predicted change in the outcome variable given a 1-unit change in the variable associated with the constant. The one qualification to this statement is for the equation with multiple variables in the linear function. For this equation (where the outcome variable is *Y*), the slope adjustable constant associated with a given variable in the function is the predicted change in the outcome variable (*Y*) given a 1-unit change in the variable *holding constant all other variables in the linear function*. In practice, data on each of the variables would be collected and the model would be fit to the data to determine if it could account for the observed data. The data would be used to estimate the values of the adjustable constants so as to

maximize model fit. If the fit is reasonable, then values of the adjustable constants are subjected to meaningful interpretation.

This example illustrates another strategy in mathematical modeling when dealing with multiple variables: the modeler creates a systems of equations rather than a single equation to represent the multivariate dynamics.

CHAOS THEORY

An area of mathematical modeling that is receiving increased attention in the social sciences is that of chaos theory. In normal parlance, *chaos* refers to disarray. In the field of chaos theory, this also is true but something systematic is thought to underlie the chaos; what appears chaotic actually has a systematic function generating it. The task of the theorist is to map this function.

Chaos theory is typically applied to changes in systems over time, with the state of a system at time $t + 1$ being impacted by the state of the system at some previous time, t. As an example, consider the simple function

$$X_{t+1} = 1.9 - X_t^2 \tag{8.22}$$

where $t + 1$ refers to the time period following time t. For example, perhaps the time interval in question is a week and suppose that the value of X at time t is 1. Then applying Equation 8.22, the value of X 1 week later (i.e., at time $t + 1$) should be $1.9 - 1^2 = 0.9$. At week 2, this input value is substituted into the right-hand side of Equation 8.22 and the result is the predicted value of X at week number 3. It is $1.9 - .9^2 = 1.09$. To predict the value at week 4, the previous value is again substituted into the right-hand side of Equation 8.22 and the result is $1.9 - 1.09^2 = 0.712$. And so on. The pattern of data is plotted in Figure 8.21, which plots the value of X at each week in a series of weeks. The pattern appears to be unsystematic and chaotic with large swings in values. But note that the underlying process is anything but haphazard. The data were the result of a clearly specified and simple function (Equation 8.22). There was no random error in the system. Rather, the "disarray" was systematically generated. The task of the chaos theorist is to identify patterns that appear to be chaotic and to find the function that generates that "chaos."

In math modeling the term *difference equation* refers to the case where a variable at time t is a function of a variable at time t - 1. If the variable at time t is a function of the immediately preceding point in time, it is called a *first-order difference equation*. If it is predicted from time $t - 2$, it is called a *second-order difference equation*. If it is predicted from time $t - 3$, it is a *third-order difference equation*. And so on.

Chaotic modeling tends to require precise measurement, and results can be dramatically influenced by the slightest "noise" or measurement error in the system. Current analytic methods for chaos models tend to require large numbers of observations. Although chaos theory is typically applied to the analysis of systems across time, the

BOX 8.1. Reading Mathematical Models

When reading mathematical models, social scientists with more limited mathematical backgrounds sometimes feel intimidated by the presence of equations. Because equations make clear and unambiguous statements about the presumed relationships between variables, you should embrace equations, not avoid them. When confronted with an equation that seems complex, here are some things you can do to help work your way through it. First, make a list of the variables in the equation and a list of the adjustable constants. Make sure that each of the variables in your list is clearly defined and that the metrics of the variables are specified. Second, determine if the equation contains any of the major functions we discussed. For example, is a power function present? Is an exponential function present? Is a logistic function present? Once you recognize a familiar function form and you have a sense of the family of curves associated with it, then the substantive implications of the equation should start to become apparent. Remember, the fundamental form of the function can be altered using transformations, so be sensitive to the presence of a function that has a transformation imposed on it. Sometimes the function is "disguised" by the adjustable constants attached to it. Third, for each adjustable constant, think about what it is accomplishing and why it was included in the equation. Is it just a scaling factor, or does it have substantive interpretation, like a slope in a linear relationship? Finally, you can use your favorite statistical package (e.g., SPSS) or graphics software to apply the equation to hypothetical data you generate and then examine the curve graphically and see what happens to it as you change values of the adjustable constants or change the hypothetical data used to generate it in systematic ways. Also, keep in mind the conditional nature that multiplicative functions imply; that is, when you see the multiplication of two variables in an equation, then the size of the derivative (i.e., the size of the effect) of one of the variables in the product term is dependent on the value of the other variable in the product term.

If you encounter mathematical symbols with which you are not familiar, then you can usually find their meaning on the Internet. Below are some commonly encountered symbols. A useful website for learning about many areas of mathematics at many different levels is called "Ask Dr. Math": *mathforum.org/dr.math*.

Common Symbols That Reflect Important Numbers

π = the ratio of the circumference to the diameter of a circle, the number 3.1415926535 . . .
e = the natural logarithm base, the number 2.718281828459 . . .
γ = the Euler–Mascheroni constant, the number 0.577215664901 . . .
ϕ = the golden ratio, the number 1.618033988749 . . .
∞ = infinity

cont.

Symbols for Binary Relations

= means "is the same as"
≠ means "is not equal to"
< means "is less than"
≤ means "is less than or equal to"
> means "is greater than"
≥ means "is greater than or equal to"
± means "plus or minus"
≅ means "is congruent to"
≈ means "is approximately equal to"
≃ means "is similar to"
≒ means "is nearly equal to"
∝ means "is proportional to"
≡ means "absolute equality"

Symbols from Mathematical Logic

∴ means "therefore"
∵ means "because"
∋ means "under the condition that"
⇒ means "logically implies that"
⇔ means "if and only if"
∀ means "for all"
∃ means "there exists"

Symbols Used in Set Theory

⊂ means "this set is a subset of"
⊃ means "this set has as a subset"
∪ is the union of two sets and means "take the elements that are in either set"
∩ is the intersection of two sets and means "take the elements that are in either set"
∅ refers to the empty set or null set and means "the set without any elements in it"
∈ means "is an element of"
∉ means "is not an element of"

Symbols for Operations

n! means "the factorial of"
$\sum$ means "the sum of"
$\prod$ means "the product of"
^ means "to the power of"
$\int$ means "the integral of"

cont.

Additional Notations

Greek letters are used to refer to population parameters, Roman, usually italic, letters are used to refer to sample statistics.

A number raised to .5 or to ½ is the same as the square root of the number. A number raised to the power of –1 is the same as the inverse of the number.

properties of space and distance can be used in place of time. Thus, theorists often distinguish between *temporal chaos* and *spatial chaos*. Temporal chaos models that focus on discrete time intervals (e.g., every 10 years; at 3-, 6-, and 12-month intervals) are called *discrete time models*, and those that use time continuously are called *continuous time models*.

A wide range of phenomena is potentially chaotic in nature, including epidemics, economic changes, the stock market, and the mental state of depression, to name a few. However, it is controversial as to whether a truly chaotic system can be isolated in the real world, in the sense described in this chapter (i.e., with a stable, generating function underlying the chaos).

Variants of chaos theory include, among other things, attempts to identify limits of predictable versus unpredictable patterns of data. For example, air flow over the wing of an airplane might be smooth and predictable when the wing is placed at low angles facing the wind. However, the air flow becomes chaotic and unpredictable at larger angles. One could attempt to determine the largest angle that permits smooth air flow, thereby yielding some understanding of this "chaotic system."

The technical aspects of chaos theory are well beyond the scope of this book. However, the theory represents an interesting application of mathematical modeling that promises to have impact in the social sciences in the future.

CATASTROPHE THEORY

Catastrophe theory is another area of mathematical modeling that is receiving attention in the social sciences. A catastrophic event is one where a large and rapid change in a system output occurs even though the system inputs are smooth and continuous. A simple example that captures the idea of catastrophic events is that of increasing the load on a bridge. One can keep adding weight to a bridge and see how the bridge deforms in response to that weight. The deforming of the bridge proceeds in a relatively uniform manner, showing increasing levels of bending. At some critical point, however, additional weight causes the bridge to collapse completely. Phenomena that might be analyzed using catastrophe theory include the occurrence of a nervous breakdown, drug

relapse, divorce, a revolution occurring in a society, a demonstration turning into mob violence, or movement from one developmental stage to another in the context of a stage theory of development.

Catastrophe theory, developed by Rene Thom (1975), postulates seven fundamental mathematical equations to describe discontinuous behavior. Catastrophe theory relates outcome variables to what are called *control variables*, which essentially are explanatory variables. The relationships between the variables are expressed mathematically using nonlinear, dynamic systems that rely on different forms of polynomial functions. The spirit of catastrophe theory can be captured intuitively in a model of aggression in dogs as developed by Zeeman (1976). The behavioral outcome ranges from flight to attack, and the response on this dimension is thought to be a function of two emotions that represent control variables: fear and anger. When fear and anger are at their neutral points, then simple increases in either fear or anger lead to a continuous increase in flight or attack responses, respectively. However, if anger is increased in an already fearful dog, then the potential for a sudden jump from flight to attack can occur. Similarly, if fear is increased in an already angry dog, a sudden jump from attack to flight can occur. The mathematical models developed by Thom and expanded by other mathematicians are designed to model such dynamics. Catastrophe theory represents another area of mathematical modeling that is starting to receive attention from the social sciences.

ADDITIONAL EXAMPLES OF MATHEMATICAL MODELS IN THE SOCIAL SCIENCES

Mathematical models exist in all of the major subdisciplines of the social sciences. Most of the subdisciplines have journals that are devoted exclusively to mathematical modeling (e.g., *Journal of Mathematical Psychology*, *Journal of Mathematical Sociology*, *Journal of Quantitative Anthropology, Marketing Science*). Mathematical models also appear in more mainstream journals, but with less frequency. It is impossible to describe the many areas in which mathematical models have been developed, but in this section, we provide a brief sampling to highlight the diversity of applications.

One area where mathematical models have been prominent is in the analysis of human decision making. This endeavor has involved applications of expected-utility theory, linear regression models, Bayesian probability models, and information theory models, to name a few. The models use mathematics to document both the strengths and limitations of humans as information processors when making decisions. Mathematical models also are prominent in theories of memory, learning, language, bargaining, and signal detectability. Mathematical models have been used extensively in the analysis of social networks involving units such as institutions, communities, elites, friendship systems, kinship systems, and trade networks. Mathematical models of political behavior have explored such issues as voting and fairness. Behavior genetics relies heavily on mathematical decompositions of the effects of unique environmental influences, shared

environmental influences, and genetic influences on human behavior in the context of twin studies. Spatial models are used to analyze residential and neighborhood patterning and the effects of this patterning on a wide range of phenomena. Geostatistical techniques explore spatial autocorrelation structures and then use mathematical models to estimate values of variables across regions. Our list could go on, but hopefully, this provides you with a sense of the diverse areas to which mathematical models have been applied.

EMERGENT THEORY CONSTRUCTION AND MATHEMATICAL MODELS

It may seem heretical to use the terms "grounded/emergent theory" and "mathematical modeling" in the same sentence, but there is no reason why some of the concepts developed in this chapter could not be used within emergent theory frameworks. For example, as one thinks about the conceptual relationships that emerge from qualitative data, are these relationships linear or nonlinear in form? If nonlinear, might they be described by logarithmic functions, exponential functions, power functions, polynomial functions, sine functions, or cosine functions? Could some systematic combination of variables underlie what seems to be chaos? Is their anything to be gained by thinking about qualitative data in terms of the logic of multiplicative modeling or cognitive algebra? And so on.

We noted earlier that mathematical modelers usually give short shrift to how the variables they decide to include in their models are chosen. Certainly an emergent theoretical framework might help them select their variables in informed and creative ways.

We think it would be interesting to have a mathematical modeler and a grounded/emergent theorist work as a multidisciplinary team on a common problem, with the instructions to develop an integrated finished product that they both would "sign off" on. Such a collaboration would undoubtedly yield nontraditional perspectives on matters.

SUMMARY AND CONCLUDING COMMENTS

Mathematical modeling is an elegant framework for constructing theories. The emphasis of mathematical modeling is thinking in terms of functions and how to describe relationships between variables in mathematical terms. Functions specify how input variables should be operated upon mathematically to produce outputs. One of the most commonly used functions in the social sciences is the linear function, which has two adjustable constants, a slope and an intercept. The intercept is the output value of the function when the input X equals zero, and the slope is the change in the output given a 1-unit increase in X. Rarely do data conform to a perfect linear function. Model disparities are accommodated through the addition of disturbance or error terms to models.

Errors are assumed to be random and inconsequential for the purposes at hand. Models without errors are deterministic, and models with errors are stochastic.

Mathematical models vary in their number of adjustable constants and the meaning of those constants/parameters. Some parameters reflect rates of change in function output per unit change in function input. These rates are best captured using the concepts of derivatives and differentiation from calculus. Derivatives refer to the concept of instantaneous change, and differentiation refers to mathematical methods for calculating the amount of instantaneous change that occurs. Integrals focus on "areas under the curve" or accumulation, and integration refers to the methods used to calculate integrals.

Mathematical models also differ in their identification status, with some models being underidentified, some being just-identified, and others being overidentified. A just-identified model is one for which there is a unique solution for each estimated parameter. In an underidentified model there are an infinite number of solutions for one or more of the model parameters. In an overidentified model there is a unique solution for the model parameters and there also is more than one feature that can be used to independently estimate a parameter value. Finally, mathematical models vary in the metrics upon which they rely upon for the input variables. The metrics of variables can affect the type of functions used to describe the relationships between variables and how the parameter variables are interpreted.

Although the assumption of linear relationships is ubiquitous in the social sciences, nonlinear relationships could very well be more common. Five major classes of nonlinear functions are logarithmic functions, exponential functions, power functions, polynomial functions, and trigonometric functions. Logarithms are used to model growth or change, where the change is rapid at first and then slows to a gradual and eventually almost nonexistent pace. Logarithmic models reflect rates of increase that are inversely proportional to the output value of the function. Exponential functions are the inverse of log functions, with the two functions mirroring each other's properties. Power functions have a similar shape to exponential and logarithmic functions, but differ at higher values of the input X. Power curves eventually outgrow a logarithmic function and undergrow an exponential function. Polynomial functions are the sum of power functions and can accommodate phenomena with "wiggles and turns." The more bends there are in a curve, the greater the number of polynomial terms that are needed to reflect those bends. Trigonometric functions are used to model cyclical phenomena, with the most common functions being the sine and cosine functions.

Functions can be manipulated through transformations and can be combined to form new functions. For example, the often used logistic function is a combination of a bounded exponential function and an increasing exponential function. Combining and manipulating functions is a key ingredient to building effective mathematical models. A typical theory construction process involves breaking up the overall process into a series of smaller component processes, specifying a function to reflect each component, and then assembling the component functions into a larger whole.

When functions involve more than one input variable, additional levels of flexibility

and complexity are introduced, as the input variables are combined additively or multiplicatively. With multiple input variables, the theorist often thinks of the function relating each individual input variable to the output variable and then combines the different variables and their functions, while taking into account the synergistic interaction between the input variables. When choosing functions to use in a model, it is advisable not to overparameterize the model or to add parameters that are not subject to meaningful substantive interpretation.

Mathematical modeling represents a sophisticated way of thinking about relationships between variables. The approach is underutilized in the social sciences, and we believe that theory construction efforts can benefit from thinking about phenomena from this perspective.

SUGGESTED READINGS

Abramowitz, M., & Stegun, I. (1974). *Handbook of mathematical functions, with formulas, graphs, and mathematical tables*. New York: Dover.—A classic on a wide range of functions. A bit dated, but still informative.

Aris, R. (1994). *Mathematical modeling techniques*. New York: Dover.—A description of mathematical modeling techniques, with a bent toward the physical sciences.

Bender, E. (1978). *An introduction to mathematical modeling*. Mineola, NY: Dover.—Describing math modeling through hundreds of examples, this book is primarily oriented to the physical and engineering sciences.

Moony, D., & Swift, R. (1999). *A course in mathematical modeling*. New York: Mathematical Associates of America.—A good introduction to mathematical modeling.

Saunders, P. (1980). *An introduction to catastrophe theory*. Cambridge, UK: Cambridge University Press.—A midlevel book on catastrophe theory.

Williams, G. (1997). *Chaos theory tamed*. Washington, DC: Joseph Henry Press.—An excellent presentation of chaos theory.

KEY TERMS

differentiation (p. 187)
integral (p. 190)
integration (p. 190)
deterministic model (p. 186)
probabilistic model (p. 186)
stochastic model (p. 186)
just-identified model (p. 191)
overidentified model (p. 191)
underidentified model (p. 191)
concavity (p. 193)
proportionality (p. 193)
scaling constant (p. 193)
logarithmic function (p. 193)
exponential function (p. 194)
Napier's constant (p. 194)
power function (p. 197)
polynomial function (p. 198)
trigonometric function (p. 200)
rational function (p. 205)
bounded exponential function (p. 206)
logistic function (p. 206)
vertical shift (p. 205)
horizontal shift (p. 205)
vertical crunch (p. 206)
horizontal crunch (p. 206)
sigmoid function (p. 206)
bilinear interaction (p. 213)
chaos theory (p. 222)
discrete time model (p. 225)
continuous time model (p. 225)
temporal chaos (p. 225)
spatial chaos (p. 225)
catastrophe theory (p. 225)
difference equation (p. 222)
first-order difference equation (p. 222)

EXERCISES

Exercises to Reinforce Concepts

1. What is the difference between an axiom and a theorem?
2. What is a function?
3. How do you interpret the value of a slope and intercept in a linear relationship?
4. How do you calculate a slope in a linear relationship? How do you calculate the intercept?
5. Why would you add an error term to a model? How does this relate to the terms *stochastic* and *deterministic* modeling?
6. What is the difference between a derivative and differentiation?
7. What is the difference between a first and second derivative?

8. What is integration?
9. Why are metrics important to consider when constructing a mathematical model?
10. Briefly describe the major types of nonlinear functions.
11. What are the major types of transformations to functions, and what effects do they have?
12. What criteria are used in choosing a function?
13. Briefly characterize chaos theory.
14. Briefly characterize catastrophe theory.

Exercises to Apply Concepts

1. Find an example of a mathematical model in the literature and write a summary of it. Discuss each of the key parameters in the model and what those parameters represent. Develop the model's conceptual and substantive implications.
2. Develop a mathematical model for a phenomenon of interest to you. Begin by identifying your outcome variable and then variables that you believe are related to it. Specify the functions relating the variables and add relevant constants to the equations, as appropriate. Justify conceptually each function and each constant. Decide if the model should be deterministic or stochastic. Start simple and then build complexity into the model, accordingly.
3. Pick a phenomenon of interest to you and try to apply either chaos theory or catastrophe theory to it. Describe the new theory as completely as you can.

Appendix 8A

SPSS Code for Exploring Distribution Properties

This appendix presents syntax from SPSS that can be used to examine curves produced by different functions.

First, open SPSS with the data field blank. We will use the syntax editor. The first step is to create a variable with a large number of cases, say 100,000. This is accomplished with the following syntax:

```
INPUT PROGRAM.
LOOP #I = 1 TO 100000.
END CASE.
END LOOP.
END FILE.
END INPUT PROGRAM.
COMPUTE X = $CASENUM.
EXECUTE.
```

The last entry in the LOOP command (100000) specifies the number of cases to generate. Numbers are generated in a variable called *X*, and these numbers range from 1 to the number of cases generated. These can be transformed to take on any metric you wish. For example, to have them range from 0 to 1, multiply *X* by .00001. To have them range from –5 to +5, multiply by .00001, subtract 0.5, and then multiply the result by 5. And so on.

Next, we compute the function we are interested in graphing. Suppose it is a log to the base 10. SPSS offers numerous built-in functions, and in this case, we use the syntax

```
COMPUTE XX=LG10(X).
GRAPH
/HISTOGRAM=XX.
```

The last two lines construct a histogram of the data, and the shape of the function will be evident from this. You can add adjustable constants and perform various transformations discussed in the chapter, as desired.

The major function commands available in SPSS are arsin, artan, cos, exp, lg10, ln, sin, and sqrt. These are defined in the help menu in SPSS. One also can work with a wide range of statistical functions, including the logistic function. Note that it is possible to calculate a log to any base from the natural log. The logarithm base *a* of any number is the natural logarithm of the number

divided by the natural logarithm of the base. For example, to calculate $\log_2(100)$, evaluate the expression ln(100)/ln(2).

There are a host of graphic software programs (for both PC and Mac) designed for scientists that allow them to graph a wide range of functions easily and quickly. These include CoPlot, DPlot, Sigma Plot, and Grapher. We are fond of DPlot. Other statistical software programs that have good graphics packages are S Plus, R, and Statistica.

Appendix 8B

Additional Modeling Issues for the Performance, Motivation, and Ability Example

This appendix describes details for the example modeling the effects of ability and motivation on performance, where the relationship between performance and motivation is nonlinear instead of linear at a given level of ability. We assume the reader is versed in standard statistical methods and psychometric theory. We illustrate the case first where motivation is assumed to impact performance in accord with a power function, with the shape of the power function changing as a function of ability. Then we mention the case where the relationship between performance and motivation is assumed to be S-shaped, with the form of the S varying as a function of ability.

We build the power function model by first positing that performance is a power function of motivation,

$$P = a + bM^c \tag{A.1}$$

where a and b are adjustable constants to accommodate metrics and c is an adjustable constant to isolate the relevant power curve in light of a and b. According to the broader theory, the effect of motivation on performance varies depending on ability (e.g., when ability is low, increases in motivation will have negligible effects on performance, but when ability is moderate to high, increases in motivation will have a more substantial impact on performance). Stated another way, the shape of the power curve will differ depending on the level of ability of students, such that the value of c is some function of A. In addition, it is likely the case that the adjustable constants a and b vary as a function of A. To simplify matters and to develop the underlying logic, we will assume that c is a linear function of A, that a is a linear function of A, and that b is a linear function of A. This yields the equations

$$\begin{aligned} c &= d + fA \\ a &= g + hA \\ b &= i + jA \end{aligned}$$

where c, d, f, g, h, i, and j are adjustable constants that conform to the respective linear models. Using substitution principles, we can substitute the right-hand side of these equations into A.1, which yields

$$P = (g + hA) + (i + jA)(M)^{(d + fA)}$$

Expanding, we obtain

$$P = (g + hA) + iM^{(d + fA)} + jAM^{(d + fA)}$$

We can rewrite this equation using the more familiar symbols of a and b for adjustable constants in regression analysis:

$$P = a + b_1A + b_2M^{(b_3+b_4A)} + b_5AM^{(b_3+b_4A)}$$

This model can be fit to data and the values of the adjustable constants estimated using nonlinear regression algorithms in SPSS or some other statistical package. The adjustable constants are amenable to interpretation, but we forgo explication of this here. Additional interpretative complications present themselves if the metrics involved are arbitrary, but we do not pursue such matters here either (see Blanton & Jaccard, 2006a, 2006b).

One intuitive way of seeing the implications of the function once the values of the adjustable constants are estimated is to calculate predicted scores that vary M by 1 unit at select values of A. These can be graphed and then subjected to interpretation.

An alternative approach to modeling the data that uses methods that are more familiar to social scientists is to use polynomial regression. In this approach, performance is assumed to be a quadratic function of motivation. Although the full quadratic curve most certainly is not applicable (because it is U-shaped), the part of the curve that forms the right half of the "U" could apply. The model includes adjustable constants to isolate this portion. We begin by writing a model where performance is a quadratic function of motivation

$$P = a_1 + b_1M + b_2M^2 \qquad \text{(A.2)}$$

and the adjustable constants in this equation (the intercept and the regression coefficients) are modeled as being a linear function of ability (we could use a nonlinear function, but for the sake of pedagogy, we assume a linear function), yielding

$$\begin{aligned} a_1 &= a_2 + b_3A \\ b_1 &= a_3 + b_4A \\ b_2 &= a_4 + b_5A \end{aligned}$$

Using the substitution principle, we substitute the right-hand sides of these equations for their respective terms in Equation A.2, which produces

$$P = (a_2 + b_3A) + (a_3 + b_4A)M + (a_4 + b_5A)M^2$$

Expanding this yields

$$P = a_2 + b_3A + a_3M + b_4AM + a_4M^2 + b_5AM^2$$

Rearranging and relabeling the constants to conform to more traditional notation yields the model

$$P = a + b_1A + b_2M + b_3AM + b_4M^2 + b_5AM^2$$

This model can be fit using standard least squares regression.

To model an S-shaped function, one can stay with polynomial regression but extend the logic to a cubic function. The basic idea is to express performance as a cubic function of motivation

$$P = a_1 + b_1M + b_2M^2 + b_3M^3$$

and then to model the adjustable constants as a function of A. Finally, use the substitution method to derive the more complex generating function.

Alternatively, one can use a logistic function to capture the S shape and then model the adjustable constants within it as a function of A. This approach requires the use of nonlinear algorithms in estimating the adjustable constants.