2

Maximum Likelihood Estimation

2.1 CHAPTER OVERVIEW

Maximum likelihood is the go-to estimator for many common statistical models, and it is one of the three major pillars of this book. As its name implies, the estimator identifies the population parameters that are most likely responsible for a particular sample of data. I spend most of the chapter unpacking this statement for analyses with normally distributed outcomes. Not only are such models exceedingly common across many different substantive disciplines, but the normal curve also appears prominently throughout the book as a distribution for missing values. As such, this chapter sets up a lot of later material. For now, I focus on complete-data maximum likelihood analyses, but all the major ideas readily generalize to missing data, and much of Chapter 3 tweaks concepts from this chapter.

The chapter begins with a simple univariate example that illustrates the mechanics of estimation and builds to multiple regression. As you will see, maximum likelihood estimates are equivalent to those of ordinary least squares, as both approaches identify estimates that minimize squared distances to the data points, albeit in different ways. After describing significance tests and corrective procedures for non-normal data, I illustrate estimation for a mean vector and variance–covariance matrix. This multivariate analysis lays the groundwork for missing data handling in models with general missing data patterns. Although I mostly discuss models with analytic solutions for the estimates, I introduce iterative optimization algorithms in this chapter, as they will be the norm with missing data.

2.2 PROBABILITY DISTRIBUTIONS VERSUS LIKELIHOOD FUNCTIONS

Probability distributions and likelihood functions play a prominent role throughout the book, so it is important to introduce these concepts early and establish some recurring

notation. A binary outcome with score values of 0 and 1 provides a simple platform for exploring some key ideas. As the name implies, a **probability distribution** is a mathematical function that describes the relative frequency of different score values. The **Bernoulli distribution** below describes the probability of the two scores:

$$f\left(Y_i \mid \pi\right) = \pi^{Y_i}\left(1-\pi\right)^{(1-Y_i)} = \begin{cases} \pi & \text{if } Y_i = 1 \\ 1-\pi & \text{if } Y_i = 0 \end{cases} \tag{2.1}$$

The function on the left side of the equation says that the probability of a particular score value depends on the unknown population proportion π to the right of the vertical pipe (the pipe means "conditional on" or "depends on"). The right side of the equation gives the rules for computing the two probabilities.

To provide a substantive context, I use the math achievement data on the companion website. Among other things, the data set includes pretest and posttest math achievement scores and academic-related variables (e.g., math self-efficacy, standardized reading scores, sociodemographic variables) for a sample of $N = 250$ students (see Appendix). One of the variables in the data is a binary indicator that measures whether a student is eligible for free or reduced-priced lunch (0 = *no assistance,* 1 = *eligible for free or reduced-price lunch*). Hypothetically, suppose we knew that the true proportion of eligible students in the population is $\pi = .45$. Figure 2.1 displays the probability distribution as a bar graph, and its mathematical description is $f(Y_i|\pi = .45)$. I use generic function notation $f(\cdot)$ throughout the book to represent the height of a distribution or curve at some value on its horizontal axis, so "*f* of something" always refers to vertical elevation. In this example, $f(Y_i|\pi = .45)$ is just a fancy way of referencing the vertical height of the bars in Figure 2.1.

The figure and previous equation highlight the defining feature of a probability distribution: Probabilities must sum to 1. The same is true for continuous probability distributions like the normal curve, where the *area* under the curve must equal 1. We will encounter many different curves and functions throughout the book, not all of which are probability distributions. The likelihood is one important example. Returning to Equation 2.1, the function on the left side of the expression has two inputs inside the parentheses: data values and a parameter. The ordering of the two inputs implies that the data values vary, but the parameter to the right of the vertical pipe (the "conditional on" symbol) functions as a known constant; that is, the probability distribution says how likely certain scores are given an assumed value for π.

After collecting data, the function is "reversed" by treating scores as known and varying the parameter π. The resulting **likelihood function** describes the relative frequency of different parameter values given the observed data. For example, suppose that we collect data from a single student who is eligible for free or reduced-price lunch (i.e., $Y = 1$). Reversing the role of the data and the parameter in the function gives the following likelihood expression:

$$L_i\left(\pi \mid Y_i = 1\right) = \pi^1\left(1-\pi\right)^0 = \pi \tag{2.2}$$

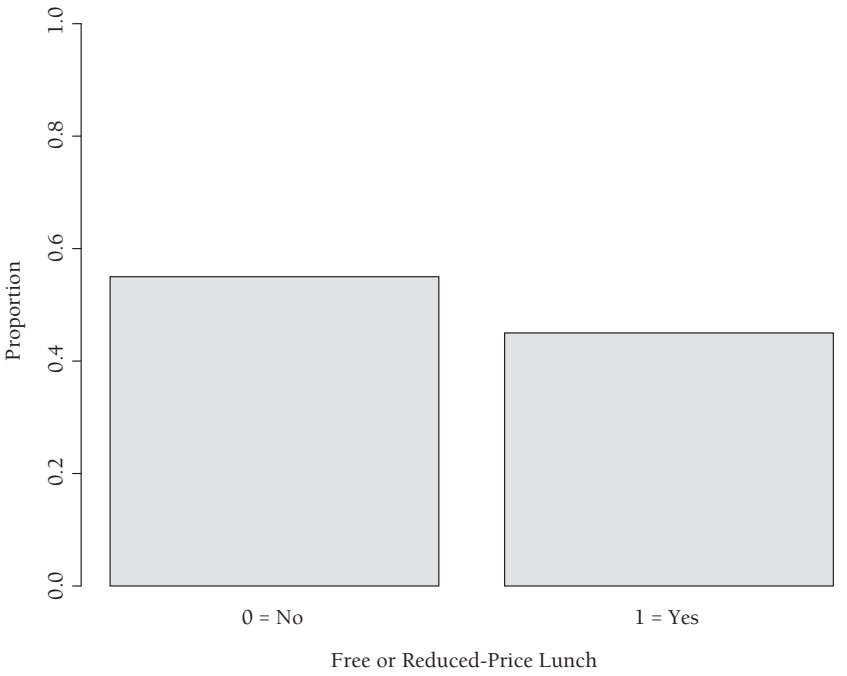

FIGURE 2.1. The probability distribution for a binary variable that measures whether a student is eligible for free or reduced-priced lunch. The bar graph corresponds to a distribution where the true proportion $\pi = .45$.

The left side of the equation now says that the likelihood of a particular parameter value depends on the observed data. Consistent with the previous function notation, "*L* of something" references the height of the distribution at a particular value on the horizontal axis, but the abscissa now reflects all possible values of π between 0 and 1.

To illustrate the effect of reversing the function's arguments, Figure 2.2 graphs the likelihood in Equation 2.2 across the entire range of π. The height of the graph—the likelihood of the parameter given the observed data—quantifies the data's support for every possible value of π. Two points are worth highlighting. First, the probability distribution of the data is discrete, but the likelihood function is a continuous distribution. Second, notice that the function defines a triangle with an area equal to 0.50. Thus, by the previous definition, the likelihood is *not* a probability distribution, because the area under the function does not equal 1. This distinction is important, because it is incorrect to say that $L_i(\pi|Y_i)$ describes the probability of the parameter given the data—that interpretation is reserved for a Bayesian analysis. Rather, you should view likelihood as a function that describes the data's evidence or support for different parameter values. As you will see later in the chapter, the likelihood function provides the mathematical machinery for identifying parameter values that maximize fit to the observed data.

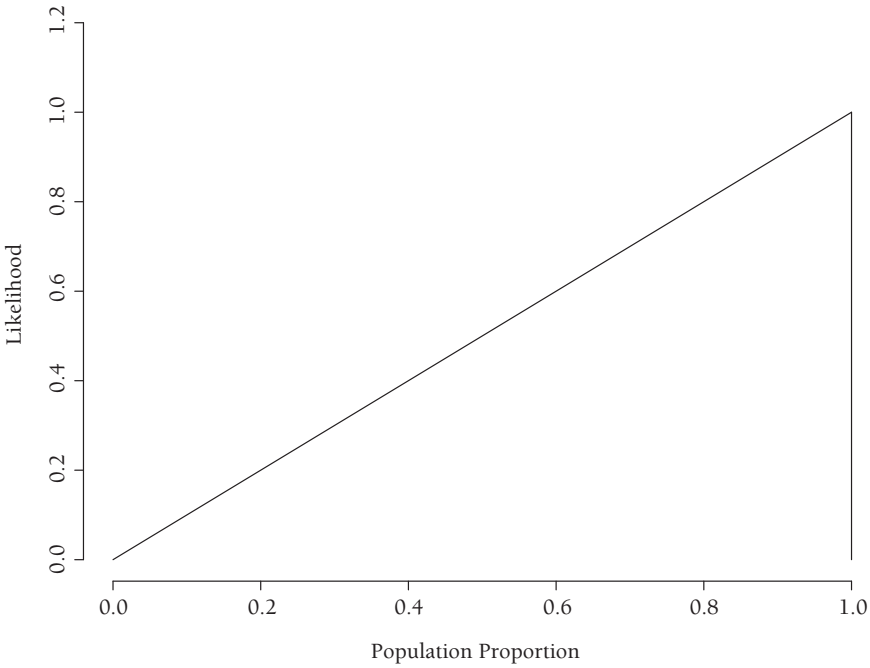

FIGURE 2.2. The likelihood function describing the relative frequency of different parameter values given a single observation where $Y = 1$. The height of the graph—the likelihood of the parameter given the observed data—quantifies the data's support for every possible value of π.

2.3 THE UNIVARIATE NORMAL DISTRIBUTION

The applications of maximum likelihood estimation in this chapter primarily leverage the normal distribution. The normal curve is a reasonable approximation for many continuous variables that we encounter in the behavioral and social sciences, and it also appears prominently later in the book as a latent response distribution for categorical variables (Albert & Chib, 1993; Johnson & Albert, 1999). A univariate analysis example is a useful starting point, because the basic estimation principles from this simple context readily generalize to more complicated analyses. Continuing with the math achievement data, I use the math posttest scores to illustrate how to estimate the mean and variance with maximum likelihood. As you will see, the mechanics of this simple example readily extend to more complex analyses.

To begin, the probability distribution for a normally distributed variable is

$$f\left(Y_i \mid \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{\left(Y_i - \mu\right)^2}{\sigma^2}\right) \tag{2.3}$$

where Y_i is the outcome score for participant i (e.g., a student's math posttest score), and μ and σ^2 are the population mean and variance. To reiterate some important notation,

the function on the left side of the equation can be read as "the relative probability of a score given assumed values for the parameters." Visually, "f of Y" is the height of the normal curve at a particular score value on the horizontal axis. Dissecting the right side of the expression, the kernel inside the exponential function defines the curve's shape. Notice that the main component is a squared z-score that quantifies the standardized distance between a score and the mean. Finally, the fraction to the left of the exponential function is a scaling term that ensures that the area under the curve sums or integrates to 1. This scaling term makes the function a probability distribution.

From the previous section, you know that a probability distribution treats scores as variable and parameters as known constants. To illustrate, assume that the true population parameters are $\mu = 56.79$ and $\sigma^2 = 87.72$ (these happen to be the maximum likelihood estimates). Next, consider two math scores, $Y_1 = 53$ and $Y_2 = 45$. Substituting these scores and the parameter values into Equation 2.3 gives $f(Y = 53|\mu, \sigma^2) = 0.039$ and $f(Y = 45|\mu, \sigma^2) = 0.019$. As seen in Figure 2.3, "f of something" refers to the height of the normal curve at a particular score value on the horizontal axis. Although these vertical coordinates look like probabilities, they are not—the probability of any one score is effectively 0, because the horizontal axis can be sliced into a countless number of infinitesimally small intervals. Rather, the height coordinates represent *relative probabilities*. For example, it is incorrect to say that 3.9% of all students from this population have a

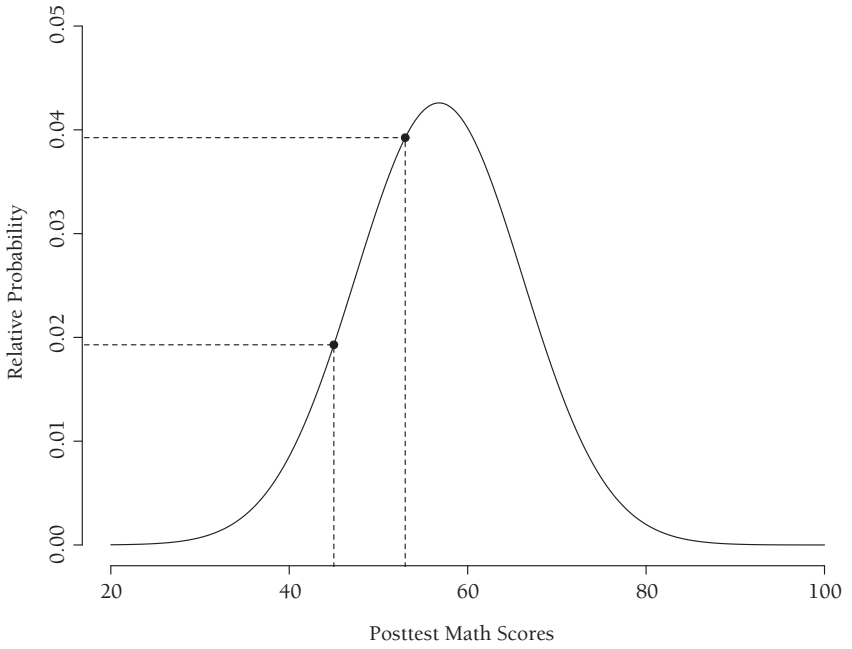

FIGURE 2.3. A normal distribution with parameters $\mu = 56.79$ and $\sigma_2 = 87.72$. The black dots are the relative probabilities for two math scores: $f(Y = 53|\mu, \sigma^2) = 0.039$ and $f(Y = 45|\mu, \sigma^2) = 0.019$.

test score of 53, but you can say that a score of 53 is about twice as likely as a score of 45, because its vertical elevation is twice as high.

The Likelihood and Log-Likelihood Functions

The goal of maximum likelihood estimation is to identify the population parameter values most likely to have produced a particular sample of data. After collecting data, the function is "reversed" by treating scores as known and varying the parameters. The likelihood expression for a single observation is

$$L_i\left(\mu, \sigma^2 \mid Y_i\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{\left(Y_i - \mu\right)^2}{\sigma^2}\right) \tag{2.4}$$

where L_i represents one observation's support for a particular combination of the mean and variance. The likelihood expression might seem like a notational sleight of hand since the right side of the expression is identical to Equation 2.3. However, the notation on the left side of the equation signals an important shift: The probability distribution views scores as hypothetical and parameters as known, whereas likelihood views parameters as hypothetical and scores as known. Applied to the math achievement data, Equation 2.4 quantifies the degree to which one observation from *this* sample supports different values of μ and σ^2.

Identifying the maximum likelihood estimates requires a summary measure that quantifies the entire sample's evidence about the unknown parameter values. From probability theory, the product of individual probabilities describes the joint occurrence of a set of independent events. For example, the probability of flipping a fair coin twice and observing two heads in a row is .50 × .50 = .25. Applying this rule to the individual likelihood expressions gives the sample likelihood function.

$$L\left(\mu, \sigma^2 \mid \text{data}\right) = \prod_{i=1}^{N} L_i\left(\mu, \sigma^2 \mid Y_i\right) \tag{2.5}$$

Extending previous ideas, the likelihood quantifies a particular sample's support for different values of μ and σ^2. Visually, the likelihood function describes a three-dimensional surface with the population mean and variance on the horizontal and depth axes and L as the height of the surface at each unique combination of the two parameters. It is important to reiterate that the likelihood function is *not* a probability distribution, because the area under the surface does not equal 1.

Applying Equation 2.5 to the math data involves multiplying 250 very small numbers, each of which requires many decimals to achieve good precision. As you can imagine, the resulting product is infinitesimally small. Taking the natural logarithm of the relative probabilities provides a more tractable metric. This transformation maps probabilities onto the negative side of the number line, with higher probabilities taking on "less negative" values than lower probabilities. To illustrate, reconsider the pair of math scores and the parameters from the previous example: $Y_1 = 53$, $Y_2 = 45$, $\mu = 56.79$, and $\sigma^2 =$

87.72. Transforming the relative probabilities to the logarithmic scale gives ln(0.039) = –3.24 and ln(0.019) = –3.96. Figure 2.4 shows that –3.24 and –3.96 also represent height coordinates, but the log transformation has changed the normal curve to a parabola. Nevertheless, the conclusion is the same: A score of 53 is more likely than a score of 45.

Working with logarithms changes the structure of the likelihood, because the logarithm product rule says to add rather than multiply the transformed likelihood values (i.e., $\ln(A \times B) = \ln(A) + \ln(B)$). Applying the product rule gives the **log-likelihood function** below.

$$LL\left(\mu,\sigma^2 \mid \text{data}\right) = \sum_{i=1}^{N} \ln\left(L_i\left(\mu,\sigma^2 \mid Y_i\right)\right) = \sum_{i=1}^{N} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{\left(Y_i-\mu\right)^2}{\sigma^2}\right)\right) \tag{2.6}$$

$$= -\frac{N}{2}\ln\left(2\pi\right) - \frac{N}{2}\ln\left(\sigma^2\right) - \frac{1}{2}\left(\sigma^2\right)^{-1}\sum_{i=1}^{N}\left(Y_i-\mu\right)^2$$

Visually, the log-likelihood function defines a three-dimensional surface, the height of which represents the data's support for a unique combination of the parameters. Figure 2.5 shows the likelihood surface for a range of different parameter combinations. To get a better look at the surface, Figure 2.6 is a contour plot that conveys the perspective of a

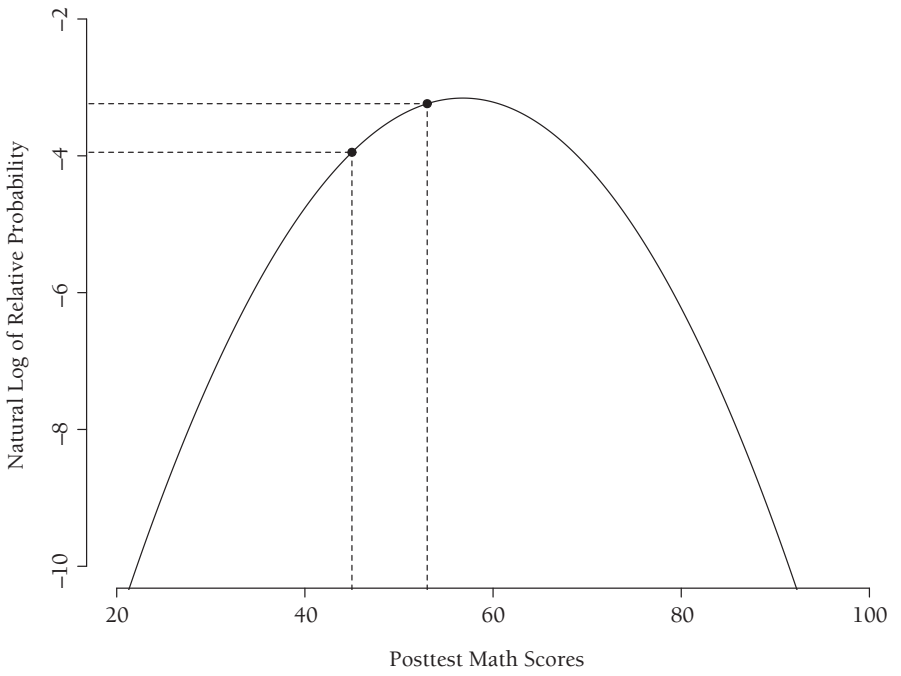

FIGURE 2.4. Natural logarithm of a normal distribution with parameters $\mu = 56.79$ and $\sigma_2 = 87.72$. The black dots represent the natural log of two relative probabilities: ln(.039) = –3.24 and ln(.019) = –3.96.

drone hovering over the peak of the log-likelihood surface, with smaller contours denoting higher elevation (and vice versa). The data's support for the parameters increases as the contours get smaller, and the maximum likelihood estimates are located at the peak of surface, shown as a black dot. The goal of estimation is to identify the parameter values at that coordinate.

As you might have surmised, the log-likelihood value will always be a large negative number, because it sums individual fit values that are themselves usually negative numbers. For example, the peak of the function in the previous figures has a vertical elevation of $LL = -913.999$, and the log-likelihood values decrease (i.e., become more negative) as μ and σ^2 move away from their optimal values for the data. Several factors influence the log-likelihood value (e.g., the sample size, the number of variables, the amount of missing data), and there is no cutoff that determines good or bad fit to the data. However, we can use the log-likelihood to make relative judgments about different candidate parameter values. These relative fit assessments are an integral part of estimation and hypothesis testing.

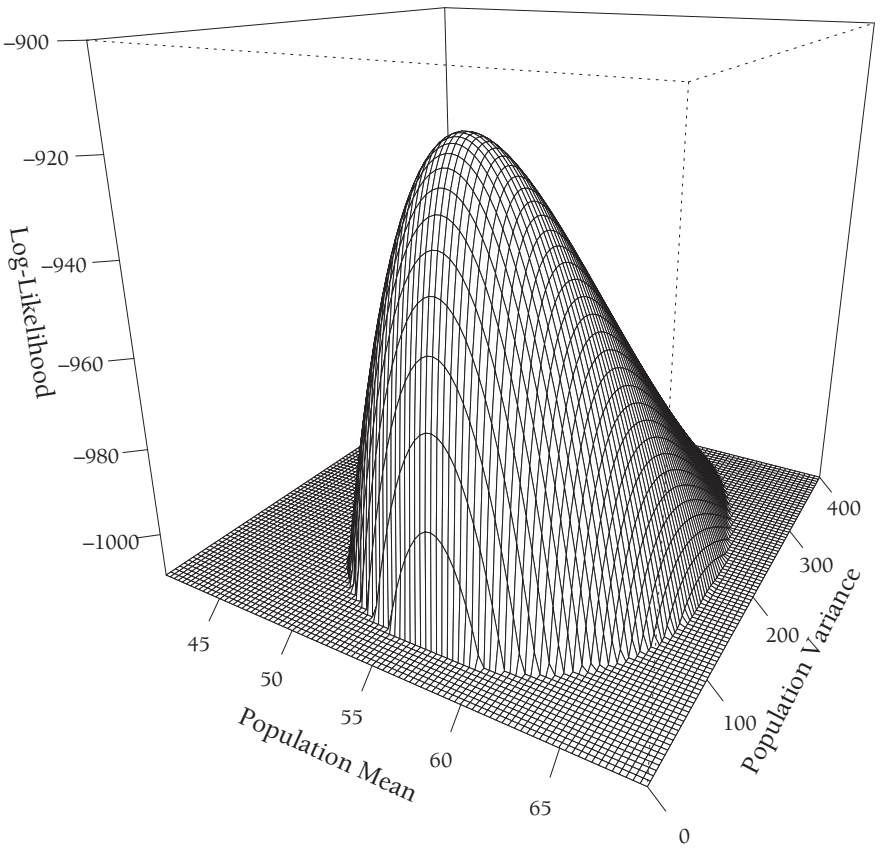

FIGURE 2.5. Bivariate log-likelihood surface for different values of μ and σ_2. The height of the surface represents the data's support for different combinations of the mean and the variance. Note that the floor of the function is located well below the minimum value on the vertical axis.

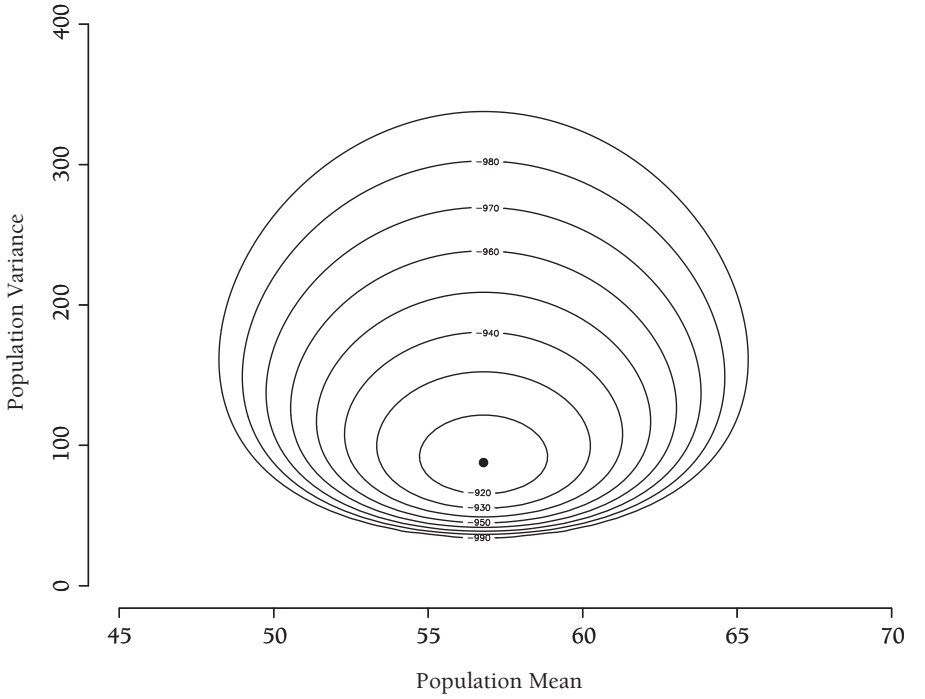

FIGURE 2.6. Contours of the log-likelihood surface at different values of μ and σ_2. The plot conveys the perspective of a drone hovering over the peak of the log-likelihood surface, with smaller contours denoting higher elevation (and vice versa). The height of the surface represents the data's support for different combinations of parameter values, and the maximum likelihood estimates are located at the peak of surface (shown as a black dot).

2.4 ESTIMATING UNKNOWN PARAMETERS

The key take-home message thus far is that "reversing" a probability distribution by treating the observed data as known constants defines a log-likelihood function that measures the data's support for different candidate parameter values. The goal of estimation is to identify the parameter values that maximize the log-likelihood function, as these are the values that garner the most support from the data. Visually, this corresponds to finding the peak of the three-dimensional surface in Figures 2.5 and 2.6. The resulting estimates are optimal in the sense that they minimize the sum of the squared *z*-scores in the normal distribution function. There are three main ways to find the maximum likelihood estimates: (a) a grid search that computes the log-likelihood value for each unique combination of the parameter values, (b) an analytic solution that provides an equation for solving the estimates, and (c) an iterative optimization algorithm. The first approach is usually too unwieldly and inefficient for practical applications, but it is a good starting point for this simple example, because it illustrates important concepts. I describe analytic solutions and optimization algorithms later in the chapter.

To illustrate the mechanics of a grid search, Table 2.1 shows individual and sample log-likelihood values at five different estimates of the population mean (to keep the illustration simple, I held the variance constant at its maximum likelihood estimate). As you might expect, an individual's contribution to the log-likelihood differs across the five estimates, because a given score offers more support for some parameter values than others (i.e., the standardized distances from the scores to the center of the normal curve change with different values of μ). The summary log-likelihood values in the bottom row of Table 2.1 similarly fluctuate as a function of the population mean. As explained previously, the log-likelihood summarizes the data's support for a particular combination of parameter values, such that higher (i.e., less negative) values reflect better fit to the data. If the five means in the table were our only options, we would choose $\hat{\mu} = 57$ as the maximum likelihood estimate, because this parameter value maximizes fit to the sample data (i.e., minimizes the sum of the squared *z*-scores).

Next, I conducted a comprehensive grid search that varied the population mean in tiny increments of 0.01 and plotted the resulting log-likelihood values in Figure 2.7. As you can see, the function resembles a hill or a parabola, with the optimal parameter value located at its peak. This brute-force estimation process revealed that the curve's highest elevation, $LL = -913.999$, is located at $\mu = 56.79$, and no other value of the mean has more support from the data. As such, $\hat{\mu} = 56.79$ is the maximum likelihood estimate of the mean, or the population parameter with the highest probability of producing this sample of 250 math scores. I applied the same grid search procedure to the variance after fixing the mean at its maximum likelihood estimate. Figure 2.8 shows the resulting log-likelihood function. Although the function looks very different—the right skew owes

TABLE 2.1. Individual and Sample Log-Likelihoods at Five Values of μ

Y	μ = 53	μ = 55	μ = 57	μ = 59	μ = 61
63	–3.72601	–3.52081	–3.36120	–3.24720	–3.17880
53	–3.15599	–3.17880	–3.24720	–3.36120	–3.52081
71	–5.00285	–4.61524	–4.27323	–3.97682	–3.72601
53	–3.15599	–3.17880	–3.24720	–3.36120	–3.52081
57	–3.24720	–3.17880	–3.15599	–3.17880	–3.24720
55	–3.17880	–3.15599	–3.17880	–3.24720	–3.36120
59	–3.36120	–3.24720	–3.17880	–3.15599	–3.17880
. . .	. . .	. . .	. . .	. . .	. . .
54	–3.16170	–3.16170	–3.20730	–3.29850	–3.43530
71	–5.00285	–4.61524	–4.27323	–3.97682	–3.72601
49	–3.24720	–3.36120	–3.52081	–3.72601	–3.97682
54	–3.16170	–3.16170	–3.20730	–3.29850	–3.43530
61	–3.52081	–3.36120	–3.24720	–3.17880	–3.15599
51	–3.17880	–3.24720	–3.36120	–3.52081	–3.72601
38	–4.43853	–4.80334	–5.21375	–5.66977	–6.17138
Sums	**–934.4897**	**–918.5749**	**–914.0604**	**–920.9462**	**–939.2323**

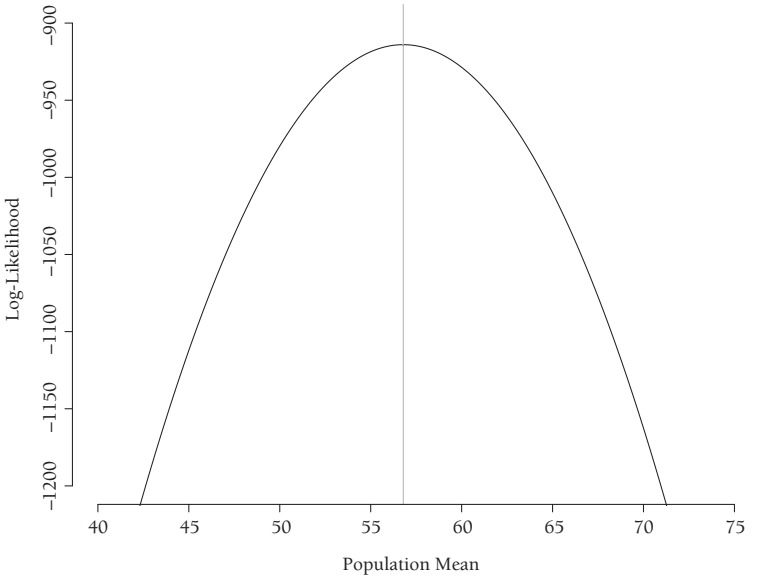

FIGURE 2.7. Likelihood function with respect to the mean, holding the variance constant at its sample estimate. The log-likelihood on the vertical axis represents the data's support for a particular parameter value. The peak of the function is the maximum likelihood estimate of the mean.

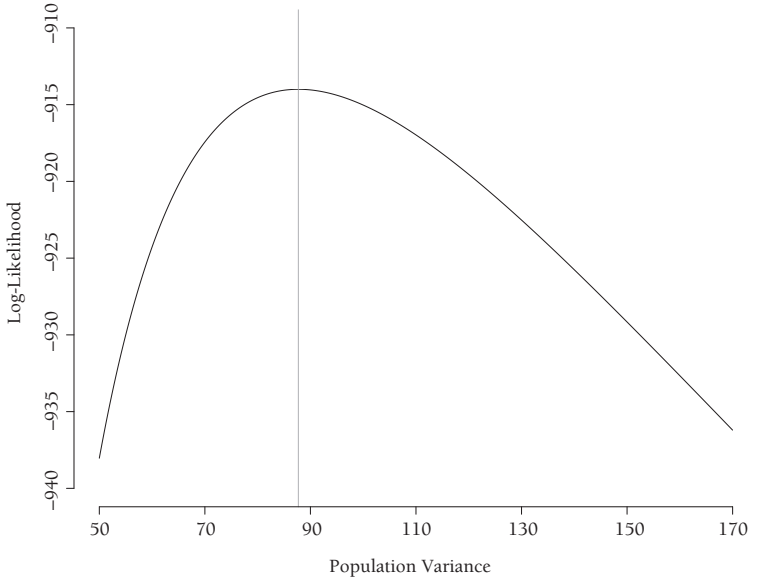

FIGURE 2.8. Likelihood function with respect to the variance, holding the mean constant at its maximum likelihood estimate. The log-likelihood on the vertical axis represents the data's support for a particular parameter value. The maximum likelihood estimate of the variance is located at the peak of the function.

to the fact that the variance is bounded at 0 on the low end—the graph nevertheless displays the data's support for different parameter values. The brute-force grid search revealed that $\hat{\sigma}^2 = 87.72$ is the maximum likelihood estimate of the variance.

2.5 GETTING AN ANALYTIC SOLUTION

You can imagine that a grid search quickly becomes impractical as the number of model parameters increases. A second approach is to derive an equation that gives an analytic solution for the maximum likelihood estimates. Although this strategy has limited applications, the mechanics of getting the solution—in particular, leveraging calculus derivatives—sets the stage for the iterative optimization algorithms that I discuss later in the chapter.

To begin, a first derivative is a slope coefficient. Returning to Figure 2.7, the log-likelihood function is a parabolic curve. Imagine using a magnifying glass to zoom in on the log-likelihood function within a very narrow slice along the horizontal axis. Although the entire function has substantial curvature, magnifying the log-likelihood at a particular point on the curve would reveal a straight line. Thus, you can think of the curved function in Figure 2.7 as stringing together a sequence of very tiny straight lines, the direction and magnitude of which vary as you move from left to right on the horizontal axis. These linear slopes are the first derivatives of the function. To infuse a bit more precision, the first derivative is the slope of a line that is tangent to the function at a particular value on the horizontal axis. To illustrate, Figure 2.9 shows the derivatives at five values of μ. I refer to these slopes as the first derivatives of the log-likelihood function *with respect to the mean,* because the variance (the other unknown quantity in the function) is held constant. First derivatives are central to finding an equation for the maximum likelihood estimates, and they also appear prominently in the iterative optimization algorithms I discuss later in the chapter.

Moving from left to right across Figure 2.9, the derivatives decrease in magnitude (i.e., the slopes flatten) as elevation rises, and the slope is exactly 0 at the function's peak. The fact that the first derivative is 0 at the point on the function directly above the maximum likelihood estimate suggests that we can set the derivative expression to 0 and solve for the unknown parameter. First, we need the derivative equations. I give the expressions below, and introductory calculus resources catalog the differential calculus rules for getting the first derivatives of a function. To begin, the first derivative of the log-likelihood function with respect to the mean (i.e., the linear slopes in Figure 2.9) is as follows:

$$\frac{\partial LL}{\partial \mu} = \left(\sigma^2\right)^{-1} \sum_{i=1}^{N} \left(Y_i - \mu\right) \tag{2.7}$$

In words, the left side of the expression reads "the first derivative of the log-likelihood function with respect to the mean," where ∂ is a common symbol for a derivative, and the fraction denotes the differential operator. Setting the right side of the equation equal to 0 and solving for μ gives the maximum likelihood estimate of the mean.

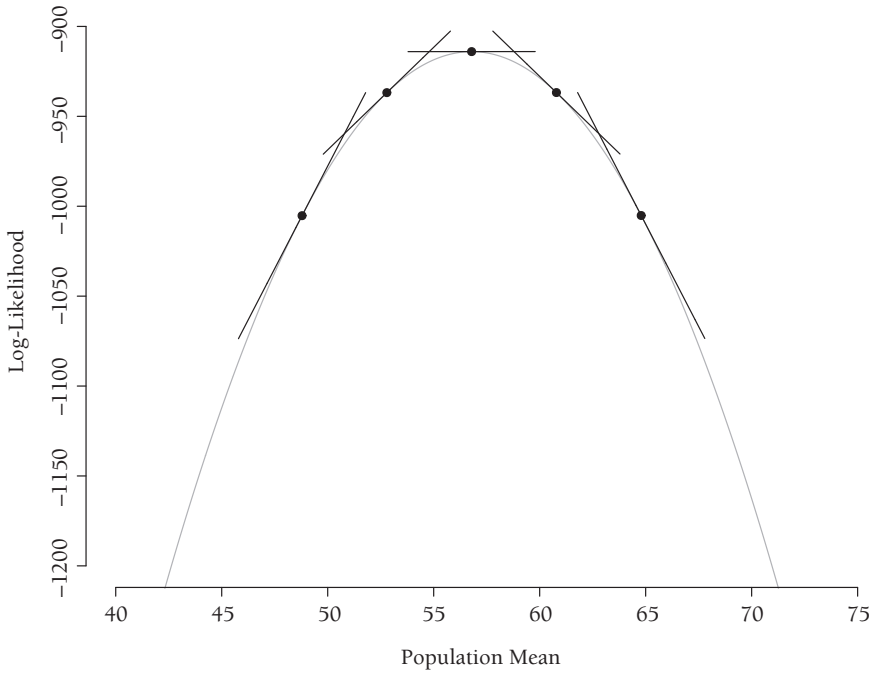

FIGURE 2.9. Likelihood function with respect to the mean, holding the variance constant at its maximum likelihood estimate. The dashed lines represent first derivatives, or slopes of lines tangent to the function at each black dot.

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} Y_i = \bar{Y} \tag{2.8}$$

Notice that $\hat{\mu}$ is identical to the familiar formula for the arithmetic mean. Consistent with the previous grid search, applying the expression to the math posttest scores gives a maximum likelihood estimate of $\hat{\mu} = 56.79$.

Differentiating the log-likelihood function with respect to the variance gives the slopes of tangent lines at different points on the function in Figure 2.8.

$$\frac{\partial LL}{\partial \sigma^2} = -\frac{N}{2}\left(\sigma^2\right)^{-1} + \frac{1}{2}\left(\sigma^2\right)^{-2}\sum_{i=1}^{N}\left(Y_i - \mu\right)^2 \tag{2.9}$$

Again, setting the right side of the equal to 0 and solving for σ^2 gives the maximum likelihood estimate of the variance.

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \mu\right)^2 \tag{2.10}$$

Notice that the maximum likelihood solution has N rather than $N - 1$ in the denominator. We know that applying the equation for the population variance to sample data

gives negatively biased estimates that underrepresent population-level variation, and the same is true for maximum likelihood variance estimates. This bias is an issue in small samples but quickly becomes negligible as N increases. Such is the case with the math achievement data, where the maximum likelihood and unbiased estimates are very similar: $\hat{\sigma}^2 = 87.72$ and $s^2 = 88.07$.

2.6 ESTIMATING STANDARD ERRORS

The log-likelihood function provides a mechanism for estimating standard errors, and this, too, relies on calculus derivatives. The process lends itself well to graphical displays, so I interleave a conceptual description with the technical details. To set the stage, Figure 2.10 shows the log-likelihood functions for two data sets with the same mean but different variance. The solid curve, which is identical to Figure 2.7, corresponds to the math posttest data, and the flatter dashed function comes from a data set with 50% more variance (i.e., $\hat{\sigma}^2 = 131.58$ vs. 87.72).

The curvature of the log-likelihood function (i.e., its steepness or flatness) determines the precision of the maximum likelihood estimate at its peak. To understand why this is the case, recall that the log-likelihood quantifies the data's evidence for different candidate parameter values. Looking at the solid curve in Figure 2.10, you can see that the

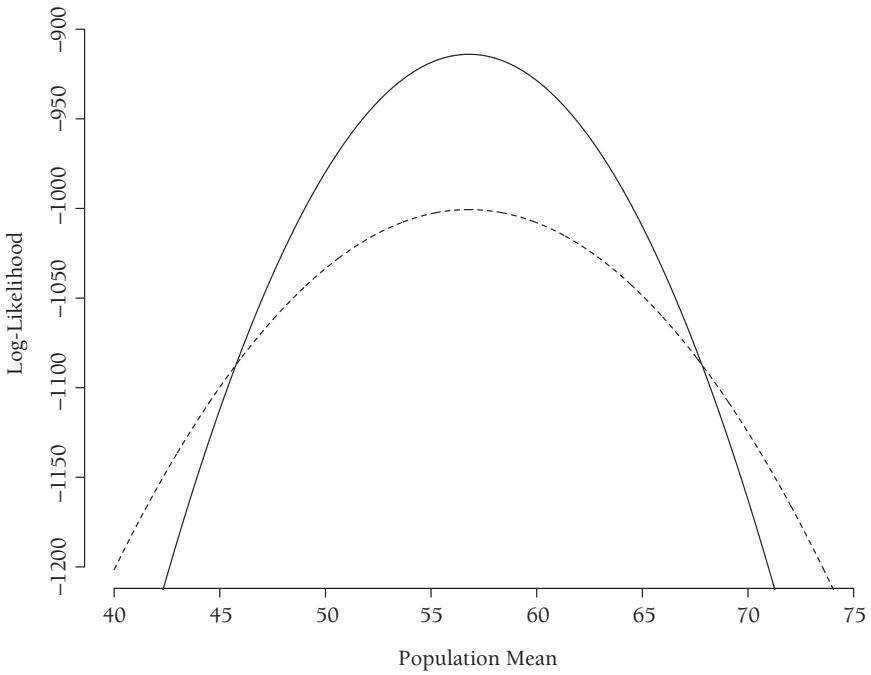

FIGURE 2.10. Log-likelihood functions for two data sets with the same mean but different variance. The solid curve, which is identical to Figure 2.7, corresponds to the math posttest data, and the flatter dashed function comes from a data set with 50% more variance.

data's support for competing parameter values decreases rapidly as μ moves away from its optimal value in either direction. In contrast, the dashed curve is much flatter, meaning that the data provide similar support for a range of parameter values near the peak. As such, the steeper function reflects a more precise estimate with a smaller standard error. This makes intuitive sense if you think about estimation as a hiker trying to climb to the highest possible elevation on a mountain. A climber standing at the top of a steep peak would be very certain about reaching the exact summit, because elevation drops quickly in every direction, whereas a climber standing on a flatter plateau would be less confident about the summit's precise location. To apply this idea to data, we need to figure out how to quantify curvature of the log-likelihood and translate that into a standard error.

Second Derivatives

Measuring curvature and computing standard errors requires the second derivatives of the log-likelihood function. These second derivatives, which are also slope coefficients, have an intuitive visual interpretation. To illustrate, Figure 2.11 displays the first derivatives of the two log-likelihood functions from Figure 2.10. Moving from left to right, the linear slopes along the steep curve vary substantially, changing from large positive values on the left to large negative values on the right. Conversely, the slopes along the

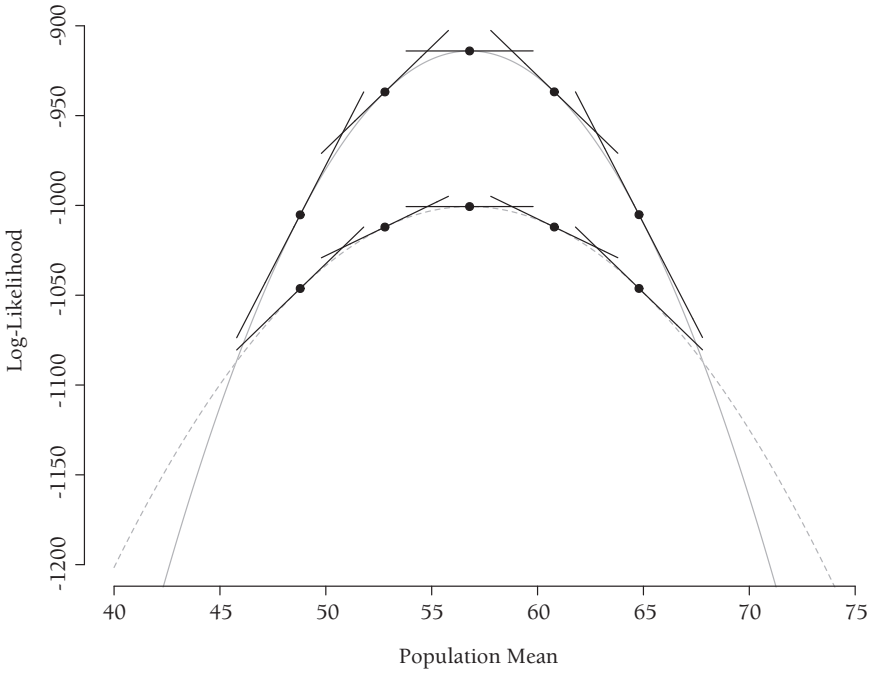

FIGURE 2.11. Log-likelihood functions for two data sets with the same mean but different variance. The straight lines represent first derivatives. The steep function has rapidly changing first derivatives and thus a large second derivative, whereas the flatter function has a smaller second derivative, because its slopes don't change as much near the peak.

flatter curve exhibit less variability, ranging from moderately positive to moderately negative. Mathematically, a second derivative captures the rate at which the first derivative slopes change across the log-likelihood function. For example, the steep function in Figure 2.11 has rapidly changing first derivatives and thus a large second derivative. Conversely, the flatter function has a smaller second derivative, because its slopes don't change as much near the function's peak.

Second derivatives can be confusing, because they are metaquantities that capture the rate of change in the linear slopes; that is, they are equations that give the slope of the slopes. A regression analogy is useful for sorting this out. Returning to Figure 2.9, you can think of the curve as a nonlinear regression line that predicts the log-likelihood at different values of the parameter (i.e., the parameter is the predictor variable, and the log-likelihood is the outcome). The linear term from this regression, which is the first derivative, tells us how much the log-likelihood changes for an infinitesimally small increase in the parameter. The second derivative is also a slope from a regression, but that regression now predicts the *first derivatives* at different values of the parameter (i.e., the parameter is the predictor variable, and the first derivative is the outcome). Because the linear slopes change at a constant rate across the parabolic function, the second derivative reflects the change in the slope for each one-unit increase in the population mean. The regression analogy highlights that first and second derivatives are just the same concept applied to different variables.

To illustrate second derivatives more concretely, reconsider the first derivative slope expression from Equation 2.7. We know that substituting $\hat{\mu} = 56.79$ into the formula (i.e., evaluating the function at the maximum likelihood estimate) returns a slope coefficient of 0. Next, we can use the expression to compute the first derivative after increasing or decreasing the mean by 1 point. Starting with the steep curve in Figure 2.11, substituting $\mu = 55.79$ and 57.79 into the equation gives first derivatives equal to +2.85 and –2.85, respectively. Thus, we can verify that a one-unit increase in the population mean changes the first derivative (i.e., the slope of the log-likelihood at a particular point) by –2.85. This value is the second derivative! Moving to the flatter function, substituting the same two estimates into the equation gives first derivatives equal to +1.90 and –1.90, respectively. A one-unit increase in the population mean now induces smaller changes in the linear slopes, because the log-likelihood function is less peaked. As you can see, larger second derivatives (in absolute value) reflect greater curvature and more precision, whereas smaller second derivatives imply less curvature.

I previously explained that second derivatives are the same concept as a first derivative but applied to a different dependent variable (a function of the original function). As such, getting the second derivatives involves applying differential calculus rules to the slope equations from Equations 2.7 and 2.9. To begin, the second derivative of the log-likelihood function with respect to the mean (i.e., the curvature of the function in Figure 2.7) is as follows:

$$\frac{\partial^2 LL}{\partial \mu^2} = -\frac{N}{\sigma^2} \tag{2.11}$$

Substituting $\hat{\sigma}^2 = 87.72$ (the maximum likelihood estimate) and $N = 250$ into the expression verifies the earlier conclusion that the second derivative equals –2.85. The second

derivative of the log-likelihood function with respect to the variance (i.e., the curvature of the function in Figure 2.8) is as follows:

$$\frac{\partial^2 LL}{\left(\partial\sigma^2\right)^2} = \frac{N}{2}\left(\sigma^2\right)^{-2} - \left(\sigma^2\right)^{-3}\sum_{i=1}^{N}\left(Y_i - \mu\right)^2 \tag{2.12}$$

Substituting $\hat{\sigma}^2 = 87.72$ and $N = 250$ in the expression gives a second derivative equal to –.016. Because the log-likelihood function in Figure 2.8 has multiple bends, the rate of change in the linear slopes is no longer constant going from left to right. Thus, we need to view the second derivative as curvature at the function's peak. Again, you can think of this number (in absolute value) as the estimate's precision.

You probably noticed that the values of the second derivative were both negative. In fact, this is not a coincidence, as the sign of the second derivative signals whether a solution corresponds to the maximum or the minimum of a function. To understand why this is the case, imagine a U-shaped log-likelihood function that is a mirror image of the parabola in Figure 2.7. When applied to a U-shaped function, the first derivative takes on a value of 0 at the *lowest* point on the curve (i.e., the bottom of a valley instead of the peak of a hill). The sign of the second derivative differentiates the minimum and maximum of a function and thus tells us whether an estimate is located at the bottom of a trough or the peak of a hill. To illustrate, imagine traversing a U-shaped function moving from left to right. Contrary to the derivatives displayed in Figure 2.9, the linear slopes from an inverted function change from large negative values to large positive values; that is, a one-unit increase to the parameter increases rather than decreases the slopes, thus giving a *positive* second derivative. Consequently, the fact that the second derivatives were negative is important, because it signals that the estimates are, in fact, located at the peak of the surface.

From Second Derivatives to Standard Errors

With second derivatives in hand, we can now compute standard errors. This process involves three steps: (1) Multiply each derivative by –1, (2) compute its reciprocal, and (3) take the square root. To begin, multiplying the second derivative by –1 gives a quantity known as **information** or **Fisher information** (after statistician Ronald Fisher). This step rescales the derivative so that large positive values reflect greater precision or confidence in the estimate. Second, computing the reciprocal or inverse of information gives the **sampling variance**, or the expected squared difference between the estimate and the true population parameter. Applying these first two steps to the mean and variance gives the following expressions for their sampling variances:

$$\text{var}\left(\hat{\mu}\right) = -\left(-\frac{N}{\hat{\sigma}^2}\right)^{-1} = \frac{\hat{\sigma}^2}{N} \tag{2.13}$$

$$\text{var}\left(\hat{\sigma}^2\right) = -\left(\frac{N}{2}\left(\hat{\sigma}^2\right)^{-2} - \left(\hat{\sigma}^2\right)^{-3}\sum_{i=1}^{N}\left(Y_i - \hat{\mu}\right)^2\right)^{-1} = 2\left(\hat{\sigma}^2\right)^3\left(-N\hat{\sigma}^2 + 2\sum_{i=1}^{N}\left(Y_i - \hat{\mu}\right)^2\right)^{-1} \tag{2.14}$$

Finally, taking the square root of the sampling variance gives the standard error. Notice that the square root of Equation 2.13 is the familiar formula for the standard error of the mean, $\hat{\sigma} \div \sqrt{N}$.

To illustrate standard error computations, reconsider the two log-likelihood functions in Figure 2.11. The steeper curve corresponds to the math achievement data from the companion website, which has a variance $\hat{\sigma}^2 = 87.72$. Substituting this estimate into Equation 2.13 gives a sampling variance equal to $\text{var}(\hat{\mu}) = 0.35$ and a standard error equal to $SE_{\hat{\mu}} = 0.59$. Consistent with the usual interpretation of a standard error, 0.59 is the expected difference between the maximum likelihood estimate and the true population mean, or the standard deviation of estimates from many random samples of size 250. As a comparison, the dashed curve corresponds to a transformed data set with 50% more variance. Substituting $\hat{\sigma}^2 = 131.58$ into Equation 2.13 returns a sampling variance and standard error equal to $\text{var}(\hat{\mu}) = 0.53$ and $SE_{\hat{\mu}} = 0.73$, respectively. These results reinforce the previous conclusion that steeper functions with more curvature reflect greater precision and smaller standard errors.

2.7 INFORMATION MATRIX AND PARAMETER COVARIANCE MATRIX

The log-likelihood function in Equation 2.6 varies as a function of *two* unknowns. Although the univariate analysis allows us to consider each parameter separately, without regard to the other, changes to one parameter generally correlate with changes to another. Returning to the three-dimensional surface in Figure 2.5, the presence of such a correlation implies that curvature or elevation changes along one axis systematically track with elevation changes along the other. Although the mean and variance happen to be uncorrelated in this example, we need to establish a more generalizable recipe for computing standard errors that accounts for potential linkages among the parameters.

Second derivatives are obtained by applying differential calculus rules to the first derivative expressions (e.g., differentiating Equation 2.7 with respect to μ gives Equation 2.11). To get the association between two parameters, you differentiate the first derivative expression for one parameter with respect to a different parameter. For example, to get the covariance between μ and σ^2, you differentiate the slope expression from Equation 2.7 with respect to σ^2 (or equivalently, differentiate the slope expression from Equation 2.9 with respect to μ). The cross-product derivative expression for this example is as follows:

$$\frac{\partial^2 LL}{\partial \mu \partial \sigma^2} = \frac{\partial^2 LL}{\partial \sigma^2 \partial \mu} = -\left(\sigma^2\right)^{-2} \sum_{i=1}^{N} \left(Y_i - \mu\right) \tag{2.15}$$

The left side of the equation reads "first differentiate the log-likelihood with respect to the mean, then differentiate the resulting expression with respect to the variance" (or vice versa).

Next, the second derivatives and the cross-product terms are stored in a symmetric matrix known as the **Hessian**.

$$\mathbf{H}_{\mathrm{O}}(\boldsymbol{\theta}) = \begin{pmatrix} -\frac{N}{\sigma^2} & -\left(\sigma^2\right)^{-2}\sum_{i=1}^{N}\left(Y_i-\mu\right) \\ -\left(\sigma^2\right)^{-2}\sum_{i=1}^{N}\left(Y_i-\mu\right) & \frac{N}{2}\left(\sigma^2\right)^{-2} - \left(\sigma^2\right)^{-3}\sum_{i=1}^{N}\left(Y_i-\mu\right)^2 \end{pmatrix} \tag{2.16}$$

Notice that the diagonal elements contain the second derivatives from Equations 2.11 and 2.12, and the new addition from Equation 2.15 appears in the off-diagonal elements. The subscript on $\mathbf{H}_{\mathrm{O}}$ indicates that the derivative equations depend on the observed data (an alternate approach described below replaces data values with the expectations or averages), and $\boldsymbol{\theta}$ denotes the parameter values. Substituting the maximum likelihood estimates into the expressions gives $\mathbf{H}_{\mathrm{O}}(\hat{\boldsymbol{\theta}})$.

Computing standard errors involves the same three steps as before. First, multiplying the matrix of second derivatives by –1 gives the **observed information matrix**.

$$\mathbf{I}_{\mathrm{O}}\left(\hat{\boldsymbol{\theta}}\right) = -\mathbf{H}_{\mathrm{O}}\left(\hat{\boldsymbol{\theta}}\right) \tag{2.17}$$

As before, this step rescales the derivatives so that large positive values reflect greater precision or confidence in the estimates. Second, taking the inverse of the information matrix (the matrix analogue of a reciprocal) gives the **variance–covariance matrix of the parameter estimates**.

$$\hat{\boldsymbol{\Sigma}}_{\hat{\theta}} = \mathbf{I}_{\mathrm{O}}^{-1}\left(\hat{\boldsymbol{\theta}}\right) \tag{2.18}$$

The parameter covariance matrix for the univariate analysis has sampling variances on the diagonal and the covariance between the two estimates in the off-diagonal elements.

$$\hat{\boldsymbol{\Sigma}}_{\hat{\theta}} = \begin{pmatrix} \mathrm{var}\left(\hat{\mu}\right) & \mathrm{cov}\left(\hat{\mu},\hat{\sigma}^2\right) \\ \mathrm{cov}\left(\hat{\sigma}^2,\hat{\mu}\right) & \mathrm{var}\left(\hat{\sigma}^2\right) \end{pmatrix} = \begin{pmatrix} 0.35 & 0 \\ 0 & 61.55 \end{pmatrix} \tag{2.19}$$

You can see that the covariance between the mean and variance is 0, because the deviation scores in the Hessian's off-diagonal sum to 0. The independence of the mean and variance (or more generally, a model's mean parameters and its variance–covariance parameters) is a well-known feature of maximum likelihood estimation. As you will see in the next chapter, this independence doesn't necessarily hold with missing data (Kenward & Molenberghs, 1998; Savalei, 2010). Finally, taking the square root of the sampling variances on the diagonal of the variance–covariance matrix gives the standard errors (e.g., $SE_{\hat{\mu}} = \sqrt{0.35} = .59$ and $SE_{\hat{\sigma}^2} = \sqrt{61.55} = 7.85$).

Standard Errors Based on Expected Information

The observed information matrix is so named, because individual elements of the Hessian matrix include deviation scores that rely on observed data values. Although this is usually the preferable way to compute standard errors, an alternative method based on the expected information matrix warrants brief discussion. With complete data, the

observed and expected information are often equivalent and produce identical standard errors. However, the two approaches are not always the same with missing data (Kenward & Molenberghs, 1998; Savalei, 2010).

Revisiting the Hessian matrix in Equation 2.16, the second derivatives reflect summations across the *N* scores. To see how expected information works, it is useful to look at a single observation's contribution to these sums.

$$\mathbf{H}_{\mathrm{O}}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \begin{pmatrix} \frac{-1}{\sigma^2} & -\sigma^2(Y_i - \mu) \\ -\sigma^2(Y_i - \mu) & \frac{1}{2}\left(\sigma^2\right)^{-2} - \left(\sigma^2\right)^{-3}(Y_i - \mu)^2 \end{pmatrix} \tag{2.20}$$

The expected information matrix invokes a computational shortcut that replaces $(Y_i - \mu)$ and $(Y_i - \mu)^2$ with their expectations or long-run averages.

$$\begin{aligned} E(Y_i - \mu) &= 0 \\ E(Y_i - \mu)^2 &= \sigma^2 \end{aligned} \tag{2.21}$$

Substituting the expectations simplifies the Hessian as follows:

$$\mathbf{H}_{\mathrm{E}}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \begin{pmatrix} -\frac{1}{\sigma^2} & 0 \\ 0 & -\frac{1}{2}\left(\sigma^2\right)^{-2} \end{pmatrix} \tag{2.22}$$

Substituting the maximum likelihood estimates into the Hessian and multiplying the matrix by –1 gives the **expected information matrix**.

$$\mathbf{I}_E(\hat{\boldsymbol{\theta}}) = -\mathbf{H}_{\mathrm{E}}(\hat{\boldsymbol{\theta}}) = -\begin{pmatrix} -\frac{N}{\hat{\sigma}^2} & 0 \\ 0 & -\frac{N}{2}\left(\hat{\sigma}^2\right)^{-2} \end{pmatrix} \tag{2.23}$$

Finally, taking the inverse of the information matrix gives the variance–covariance matrix of the estimates, the diagonal of which contains squared standard errors.

As you can see, the expected information is simpler to compute, because it does not rely on the raw data. With complete data, standard errors based on the observed and expected information are often indistinguishable, as they are in this example. This equality doesn't necessarily hold with missing data, as the expectations in Equation 2.21 require an MCAR process where missingness is unrelated to the data. In contrast, standard errors based on the observed information assume a less stringent MAR mechanism where missingness depends on the observed data. Simulation results favor standard errors based on observed information (Kenward & Molenberghs, 1998; Savalei, 2010), so I strictly rely on this approach.

2.8 ALTERNATIVE APPROACHES TO ESTIMATING STANDARD ERRORS

The normal curve plays an integral role in every phase of maximum likelihood estimation, as its log-likelihood function provides a basis for identifying the optimal estimates for the data and computing standard errors. Of course, non-normal data are exceedingly common, and some authors argue that normality is the exception rather than the rule (Micceri, 1989). Depending on the analysis model, maximum likelihood estimates may still be consistent when normality is violated, meaning that they converge to their true population values as the sample size increases (Yuan, 2009b; Yuan & Bentler, 2010). However, standard errors and significance tests are almost certainly compromised.

This section describes two alternate (and very different) strategies for estimating sampling variation when normality is violated: so-called "robust" or sandwich estimator standard errors (Freedman, 2006; Greene, 2017; White, 1980) and bootstrap resampling (Efron, 1987; Efron & Gong, 1983; Efron & Tibshirani, 1993). These methods have a long history in the literature and a substantial body of literature that generally supports their use (Arminger & Sobel, 1990; Enders, 2001; Finch, West, & MacKinnon, 1997; Gold & Bentler, 2000; Hancock & Liu, 2012; Rhemtulla, Brosseau-Liard, & Savalei, 2012; Savalei & Falk, 2014; Yuan, 2009b; Yuan & Bentler, 2000, 2010; Yuan, Bentler, & Zhang, 2005; Yuan, Yang-Wallentin, & Bentler, 2012). I discuss analogous corrective procedures for significance tests later in the chapter.

Robust Standard Errors

The previous standard error formulation assumes that the model—including the assumed population distribution—is correctly specified. We can and often do apply maximum likelihood to non-normal or heteroscedastic data, in which case the estimation procedure is known as **quasi-maximum likelihood** or **pseudo maximum likelihood** estimation (Gourieroux, Monfort, & Trognon, 1984; Greene, 2017; White, 1996). Depending on the analysis model, pseudo maximum likelihood estimation may still provide consistent estimates that converge to the true population values as the sample size gets larger (Yuan, 2009b; Yuan & Bentler, 2010), but the usual expressions for standard errors are invalid. Alternative standard error expressions for misspecified models are widely referred to as **robust standard errors** or **sandwich estimator standard errors**.

Robust or sandwich estimator standard errors are a family of procedures that attempt to adjust for different types of model misspecification. For example, the standard errors I outline below are designed for distributional misspecifications but do not address independence violations resulting from clustered data (e.g., repeated measurements nested in persons, students nested within schools); different types of misspecifications require different corrective procedures. I give a brief description of robust standard errors for non-normal data in this section, and several good tutorial papers are available to readers who want additional details (Freedman, 2006; Hayes & Cai, 2007; Savalei, 2014).

The term *sandwich estimator* stems from fact that the "robustified" parameter covariance matrix has a three-part structure that resembles a sandwich. The normal-theory

covariance matrix from Equation 2.18 forms the outer pieces of "bread," and the "meat" in the middle of the sandwich is a new matrix that captures deviations between the data and the assumed normal distribution. The sandwich estimator covariance matrix is

$$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\theta}}} = \text{bread} \times \text{meat} \times \text{bread} = \mathbf{I}_{\text{O}}^{-1}\left(\hat{\boldsymbol{\theta}}\right)\hat{\boldsymbol{\Sigma}}_{\text{S}\left(\hat{\boldsymbol{\theta}}\right)}\mathbf{I}_{\text{O}}^{-1}\left(\hat{\boldsymbol{\theta}}\right) \tag{2.24}$$

where $\mathbf{I}_{\text{O}}(\boldsymbol{\theta})$ is the information matrix from Equation 2.17, and the meat in the middle term is a new covariance matrix based on first derivatives (described below).

Revisiting Equations 2.7 and 2.9, the first derivative or slope expressions reflect summations across the *N* scores. To illustrate the composition of the meat term, we need to look at a single observation's contribution to these equations. Arranging the terms in an array gives the so-called **score vector** for a single observation.

$$\text{S}_i\left(\boldsymbol{\theta}\right) = \begin{pmatrix} \left(\sigma^2\right)^{-1}\left(Y_i - \mu\right) \\ -\frac{1}{2}\left(\sigma^2\right)^{-1} + \frac{1}{2}\left(\sigma^2\right)^{-2}\left(Y_i - \mu\right)^2 \end{pmatrix} \tag{2.25}$$

The meat of the sandwich is the variance–covariance matrix of these score vectors evaluated at the maximum likelihood estimates (i.e., the $\hat{\boldsymbol{\Sigma}}_{\text{S}\left(\hat{\boldsymbol{\theta}}\right)}$ term in Equation 2.24).

To understand how the formula works, you need know that $\mathbf{I}_{\text{O}}(\hat{\boldsymbol{\theta}})$ and $\hat{\boldsymbol{\Sigma}}_{\text{S}\left(\hat{\boldsymbol{\theta}}\right)}$ both estimate the information matrix, albeit in different ways. When the data are normal, the two matrices are equivalent and effectively cancel out when multiplying one by the inverse of the other (the resulting product is an inert identity matrix), leaving only the normal-theory covariance matrix from Equation 2.18. In contrast, when the data are non-normal, the product of the two matrices has diagonal elements that reflect the relative magnitude of the two information matrices, and this array serves to rescale the parameter covariance matrix in a way that compensates for kurtosis. Returning to the score vector in Equation 2.25, notice that the first derivative expressions include deviation scores. When the data are leptokurtic, the thicker tails produce a higher proportion of large deviation scores than a normal curve, and multiplying the first piece of bread by the meat returns a matrix containing large diagonal values that inflate the parameter covariance matrix (the rightmost piece of bread). In contrast, when the data are platykurtic, the distribution has fewer extreme scores than a normal curve, and the bread × meat product returns a matrix with fractional values that attenuate the covariance matrix elements.

Recall from the earlier example that the normal-theory standard errors for the mean and variance were $SE_{\hat{\mu}} = \sqrt{0.35} = 0.59$ and $SE_{\hat{\sigma}^2} = \sqrt{61.55} = 7.85$, respectively. The sandwich estimator covariance matrix for the same data is as follows:

$$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\theta}}} = \begin{pmatrix} \text{var}\left(\hat{\mu}\right) & \text{cov}\left(\hat{\mu}, \hat{\sigma}^2\right) \\ \text{cov}\left(\hat{\sigma}^2, \hat{\mu}\right) & \text{var}\left(\hat{\sigma}^2\right) \end{pmatrix} = \begin{pmatrix} 0.35 & 0.17 \\ 0.17 & 60.30 \end{pmatrix} \tag{2.26}$$

Taking the square root of the diagonal elements gives $SE_{\hat{\mu}} = \sqrt{0.35} = .59$ and $SE_{\hat{\sigma}^2} = \sqrt{60.31} = 7.77$. This example highlights two points. First, the standard error of the mean is the same in both cases, because this parameter is unaffected by the robustification pro-

cess (White, 1982; Yuan et al., 2005). Second, the standard error of the variance barely changes, because the data are essentially normal (as noted previously, the sandwich estimator simplifies to the conventional covariance matrix in this case). More generally, a divergence between the two covariance matrices would likely signal a model misspecification (e.g., the normal distribution is a poor approximation for the data; King & Roberts, 2015; White, 1982).

Bootstrap Resampling

Bootstrap resampling (Efron, 1987; Efron & Gong, 1983; Efron & Tibshirani, 1993) is a second approach to generating standard errors that are robust to normality violations. The bootstrap uses Monte Carlo computer simulation to generate an empirical sampling distribution of each parameter estimate, the standard deviation of which is the standard error. This section describes a so-called "naive bootstrap" that generates standard errors, and modifications to the basic procedure can also generate sampling distributions of test statistics (Beran & Srivastava, 1985; Bollen & Stine, 1992; Enders, 2002; Hancock & Liu, 2012; Savalei & Yuan, 2009).

The basic idea behind the bootstrap is to treat the observed data as a surrogate for the population and draw B samples of size N *with replacement;* that is, after being selected for a bootstrap sample, each observation returns to the surrogate population and is eligible to be chosen again. The sampling with replacement scheme ensures that some data records appear more than once in each sample, whereas others do not appear at all. To illustrate, Table 2.2 shows five bootstrap samples from a small toy data set with 10 observations. Drawing many bootstrap samples (e.g., $B > 2{,}000$) and fitting a model to each data set gives an empirical sampling distribution of the estimates. The standard deviation of B estimates is the bootstrap standard error

$$SE_{\hat{\theta}} = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b - \bar{\theta}\right)^2 = SD_{\hat{\theta}} \tag{2.27}$$

TABLE 2.2. Five Bootstrap Samples Drawn with Replacement

Y	Sample 1	Sample 3	Sample 3	Sample 4	Sample 5
63	71	49	61	71	71
53	71	55	61	38	54
71	57	63	49	55	71
57	55	49	63	63	38
55	61	38	57	63	49
54	51	53	57	61	61
49	71	71	57	53	55
61	71	61	51	57	54
51	61	55	54	57	38
38	38	51	61	53	53

where $\hat{\theta}_b$ is the maximum likelihood estimate from sample b, and $\overline{\theta}$ is the average estimate across the B samples. Finally, the 2.5 and 97.5% quantiles of the empirical distribution (i.e., the estimates that separate the most extreme 2.5% of the lower and upper tails of the distribution) define a 95% confidence interval. Unlike their theoretical counterparts, bootstrap confidence intervals need not be symmetric around the average point estimate.

2.9 ITERATIVE OPTIMIZATION ALGORITHMS

This chapter focuses primarily on analyses with analytic solutions for the maximum likelihood estimates. Beyond the univariate example, this includes linear regression models and multivariate analyses involving a mean vector and covariance matrix. Many, if not most, applications of maximum likelihood do not have analytic solutions, and even the tidy problems from this chapter become messy later with missing data. I describe two such algorithms in this chapter, gradient ascent and Newton's method, and in Chapter 3, I describe the expectation maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977; Rubin, 1991).

Returning to the log-likelihood function in Figure 2.7, an optimization algorithm tasked with finding the maximum likelihood estimates is like a hiker trying to reach the summit of a mountain. The hiker could start the trek at different trailheads, and that starting point would dictate the direction of travel and rate of ascent. Similarly, optimization algorithms need initial guesses about the parameter values, and software defaults could generate starting values on either side of hill. The first derivative is like a compass in the sense that its sign tells the algorithm the direction it needs to travel to reach the curve's maximum elevation. For example, starting the climb at $\mu = 45$ requires positive adjustments to the parameter, whereas starting at $\mu = 65$ requires negative adjustments. The starting coordinates also dictate the size of the hiker's steps. If the trek begins far from the peak, the hiker can take big steps without worrying about missing the summit. In contrast, the surface flattens near the top where very tiny steps are needed to find the exact location of the peak. The size of each step links to the magnitude of the derivatives in Figure 2.9, with larger slopes inducing bigger steps, and slopes closer to 0 requiring very small steps.

Gradient Ascent

Gradient ascent (or equivalently, gradient *descent,* if you invert the log-likelihood function) is a good starting point for exploring iterative optimization, because it parallels the hiking analogy. Starting with an initial guess about the parameter, the algorithm takes repeated steps in the direction of maximum until it finds the optimal estimate for the data. The iterative recipe for gradient ascent is straightforward: At each iteration, compute an updated estimate that equals the previous estimate plus some adjustment, the size of which depends on the first derivative or slope. More formally, the updating step is

$$\text{new estimate} = \text{current estimate} + \text{step size} \tag{2.28}$$

$$\theta^{(t+1)} = \theta^{(t)} + \left(\frac{\partial LL}{\partial \theta} \times \text{constant} \right)$$

where θ denotes the parameter of interest, *t* indexes the iterations, and the step size term in parentheses is the first derivative (evaluated at the current estimate) times a small constant, sometimes referred to as the *learning rate*.

To illustrate iterative optimization, I applied gradient ascent to the mean (to keep the illustration simple, I held the variance at its maximum likelihood estimate). A custom R program is available on the companion website for readers interested in coding the algorithm by hand. To begin, I initiated the process with a starting value of $\mu^{(0)} = 0$ and a constant learning rate of .25 (the constant is usually some small value between 0 and 1). Substituting the initial parameter value into the first derivative expression from Equation 2.7 (i.e., evaluating the function at $\mu = 0$) gives a slope equal to 161.86. The huge positive slope implies a correspondingly large positive adjustment to the parameter. Multiplying the derivative by the learning rate gives a step size equal to 161.86 × .25 = 40.47 and an updated parameter value equal to $\mu^{(1)} = 40.47$. The new estimate is closer to the peak, so the slope coefficient decreases in magnitude to 46.53. Repeating the process gives a step size equal to 46.53 × .25 = 11.63 and an updated estimate equal to $\mu^{(2)} = 52.10$.

Table 2.3 gives the parameter updates, first derivatives, and log-likelihood values from 17 iterations. As you can see, the first few cycles produced steep slope coefficients

TABLE 2.3. Iterative Updates from a Gradient Ascent Algorithm

Iteration	μ	Slope	Log-likelihood
0	0.0000000	161.8619279	–5510.230027773
1	40.4654820	46.5319355	–1293.850964489
2	52.0984659	13.3769630	–945.391338786
3	55.4427066	3.8455985	–916.593142769
4	56.4041062	1.1055295	–914.213136500
5	56.6804886	0.3178167	–914.016442587
6	56.7599428	0.0913657	–914.000186960
7	56.7827842	0.0262657	–913.998843525
8	56.7893507	0.0075509	–913.998732498
9	56.7912384	0.0021707	–913.998723322
10	56.7917810	0.0006240	–913.998722564
11	56.7919371	0.0001794	–913.998722501
12	56.7919819	0.0000516	–913.998722496
13	56.7919948	0.0000148	–913.998722496
14	56.7919985	0.0000043	–913.998722496
15	56.7919996	0.0000012	–913.998722496
16	56.7919999	0.0000004	–913.998722496

and large adjustments to the parameter. The vertical elevation of the log-likelihood also increased rapidly as the algorithm took large strides toward the peak. In contrast, the final few iterations induced *very* small adjustments to the mean, and changes to the log-likelihood were in the 10th decimal. Continuing to iterate until the derivative equals 0 is inefficient and unnecessary, because any additional improvement to the estimate would be infinitesimally small (e.g., after 17 iterations, the estimate is changing in the seventh decimal place). Instead, I terminated the iterations when the estimates from consecutive steps differed by less than .000001, as changes of this magnitude effectively signal that the algorithm has reached the summit.

Newton's Algorithm

Gradient ascent is useful for establishing some intuition about iterative optimization, but the simple variant I describe here can be slow to converge and may not converge at all when variables have different scales. **Newton's algorithm** (also known as the **Newton–Raphson algorithm**) similarly parallels the hiking analogy, but it uses a more complex formulation for the step size that requires first *and* second derivatives. The upside of this additional complexity is that the updating step naturally provides the building blocks for computing standard errors after the final iteration. To illustrate the basic ideas, reconsider the log-likelihood function with respect to the variance in Figure 2.8. Although the log-likelihood is a complex curve with multiple bends, magnifying a graph of the function at its maximum would reveal a simpler curved line that resembles a quadratic function (i.e., an inverted U, or a parabola). Leveraging this idea, Newton's algorithm uses the first and second derivative values (i.e., the linear slope and curvature at a specific point on the function) to construct a parabolic curve that extends from the current parameter value toward the log-likelihood's peak. The apex of each quadratic function represents the algorithm's best guess about the maximum likelihood estimate at a particular iteration, and this temporary peak becomes the updated parameter value for the next iteration.

Figure 2.12 shows the log-likelihood function, with black dots denoting four consecutive parameter values. The three dashed lines are quadratic curves assembled from the first and second derivative formulas. To illustrate the iterative updates, suppose that the optimizer begins its ascent from a starting value of $\sigma^{2(0)} = 50$. A black dot appears on the log-likelihood function at this coordinate, and the leftmost dashed curve (the smallest of the three) is the parabolic function that projects from the starting value. The dashed curve is trying to approximate what the log-likelihood function looks like near its summit, and the apex of the curve represents the parabola's best guess about the maximum likelihood estimate at the initial iteration. The peak of the quadratic curve, located at $\sigma^{2(1)} = 65.03$, becomes the new estimate for the next iteration. Repeating the process, the algorithm substitutes the updated estimate into the first and second derivative expressions and uses the resulting quantities to project another quadratic function from the new coordinate. The middle of the three dashed curves shows the parabola for this step, the peak of which is located at $\sigma^{2(2)} = 78.40$. Similarly, the rightmost dashed curve shows the quadratic approximation at the third iteration, the maximum of which corresponds to $\sigma^{2(3)} = 85.93$. You can see that the dashed curves become wider and flat-

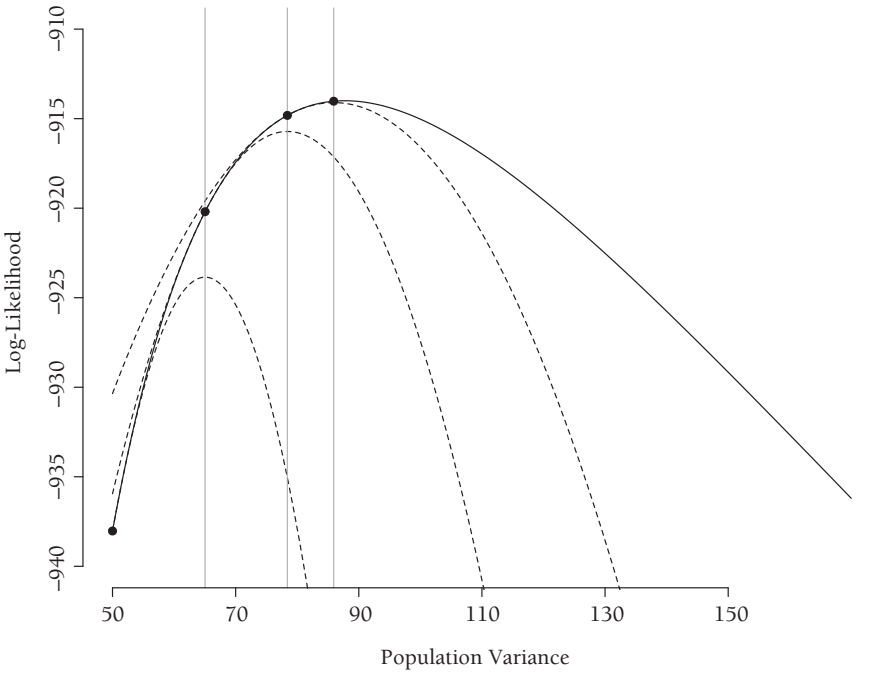

FIGURE 2.12. The likelihood function with respect to the variance, holding the mean constant at its maximum likelihood estimate. The black dots represent four consecutive updates to the variance beginning at the starting value $\sigma^{2(0)} = 50$. The three dashed lines are quadratic curves assembled from the first and second derivative formulas, and the peak of each parabola identifies the updated parameter value at the next iteration.

ter as elevation increases, such that each successive update does an increasingly better job at approximating the shape of the log-likelihood function near its peak. After a few more iterations, the algorithm locates the summit.

More formally, the jump from the current to the updated parameter value is as follows:

$$\text{new estimate} = \text{current estimate} + \text{step size} \tag{2.29}$$

$$\theta^{(t+1)} = \theta^{(t)} - \left(\frac{\partial^2 LL}{\partial \theta^2}\right)^{-1} \left(\frac{\partial LL}{\partial \theta}\right)$$

The step size, computed as the ratio of the first and second derivatives at the current parameter value θ^t, corresponds to the horizontal distance between the current estimate and the peak of the projected quadratic curve. In effect, Newton's algorithm is breaking the total vertical elevation into several smaller hikes, and the derivative terms function as a wayfinder that plots the route to each intermediate peak. The updating step readily extends to more complex models with multiple parameters. In this case, the multivariate updating equation is

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \mathbf{H}^{-1}(\boldsymbol{\theta}^{(t)})\mathbf{S}(\boldsymbol{\theta}^{(t)}) \tag{2.30}$$

where $\boldsymbol{\theta}$ is a vector of parameter values, t indexes the iterations, $\mathbf{S}(\boldsymbol{\theta}^{(t)})$ is the vector of first derivatives (i.e., the score vector, computed as the sum of Equation 2.25 across all N observations), and the rightmost term is the Hessian matrix of second derivatives.

To illustrate a multivariate optimization scheme, I used Newton's algorithm to estimate the mean and variance of the math posttest scores. A custom R program is available on the companion website for readers interested in coding the algorithm by hand. In this example, $\mathbf{S}(\boldsymbol{\theta})$ is a vector containing the slope expressions from Equations 2.7 and 2.9, and $\mathbf{H}(\boldsymbol{\theta})$ is the second derivative matrix from Equation 2.16. The multivariate updating scheme is virtually identical to the univariate scheme depicted in Figure 2.12, except that each parameter's parabolic approximation now accounts for the associations in the Hessian's off-diagonal. Table 2.4 shows the iterative updates from a climb initiated at (terrible) starting values of $\mu^{(0)} = 0$ and $\sigma^{2(0)} = 1$. Notice that the algorithm immediately locates the optimal estimate of the mean after the first update. Returning to Figure 2.7, the log-likelihood with respect to the mean is itself a parabolic function, so the optimizer can immediately predict the peak of the function from any starting value. In contrast, the algorithm requires 17 iterations to locate the optimal value of the variance. Consistent with gradient ascent, you can see that the optimizer makes large adjustments at first and very small alterations as it approaches the peak.

TABLE 2.4. Iterative Updates from Newton's Algorithm

Iteration	μ	σ^2	Log-likelihood
0	0	1	–414360.734633301
1	56.79200000	1.49992453	–7590.507280671990
2	56.79200000	2.24341946	–5218.181032416420
3	56.79200000	3.35059911	–3653.304333842760
4	56.79200000	4.99327916	–2626.616256992080
5	56.79200000	7.41677626	–1958.552805664100
6	56.79200000	10.96146467	–1529.318174000000
7	56.79200000	16.07692601	–1258.915758159560
8	56.79200000	23.30441653	–1093.809160106400
9	56.79200000	33.17164090	–997.987688507726
10	56.79200000	45.89009316	–946.947360485552
11	56.79200000	60.70697190	–923.606989421236
12	56.79200000	74.99905776	–915.615473776683
13	56.79200000	84.49594080	–914.087289046441
14	56.79200000	87.48858990	–913.999146772176
15	56.79200000	87.71555229	–913.998722506935
16	56.79200000	87.71673597	–913.998722495553
17	56.79200000	87.71673600	–913.998722495553

2.10 LINEAR REGRESSION

This section extends maximum likelihood estimation to a multiple regression analysis. As you will see, the previous concepts readily generalize to this analysis with virtually no modifications, because estimation still relies on the univariate normal curve. A single-predictor model is a useful starting point, because the log-likelihood function for the coefficients can be visualized in a three-dimensional graph. The simple regression model is

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i = E\left(Y_i \mid X_i\right) + \varepsilon_i \tag{2.31}$$
$$Y_i \sim N_1\left(E\left(Y_i \mid X_i\right), \sigma_\varepsilon^2\right)$$

where $E(Y|X)$ is a predicted value (i.e., the expected value or mean of Y given a particular X score), the tilde means "distributed as," N_1 denotes the univariate normal distribution function (i.e., the probability distribution in Equation 2.3), and the conditional mean and residual variance inside the parentheses are the distribution's two parameters. The bottom row of the expression is simply stating our usual assumption that outcome scores are normally distributed around a regression line with constant residual variation.

Switching gears to a different substantive context, I use the smoking data from the companion website to illustrate multiple regression. The data set includes several sociodemographic correlates of smoking intensity from a survey of N = 2,000 young adults (e.g., age, whether a parent smoked, gender, income). To facilitate graphing, I start with a simple regression model where the parental smoking indicator (0 = *parents did not smoke,* 1 = *one or both parents smoked*) predicts smoking intensity (higher scores reflect more cigarettes smoked per day):

$$INTENSITY_i = \beta_0 + \beta_1\left(PARSMOKE_i\right) + \varepsilon_i \tag{2.32}$$

The intercept represents the expected smoking intensity score for a respondent whose parents did not smoke, and the slope is the group mean difference. The analysis example later in this section expands the model to include additional explanatory variables.

Probability Distribution and Log-Likelihood

Linear regression leverages the univariate normal distribution function from Equation 2.3, and the only difference is that a predicted value and residual variance replace μ and σ^2, respectively. Using generic notation, the probability distribution (normal curve equation) for the simple regression is as follows:

$$f\left(Y_i \mid \boldsymbol{\beta}, \sigma_\varepsilon^2, X_i\right) = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left(-\frac{1}{2}\frac{\left(Y_i - E\left(Y_i \mid X_i\right)\right)^2}{\sigma_\varepsilon^2}\right) \tag{2.33}$$

To reiterate recurring notation, the function on the left side of the equation can be read as "the relative probability of a score given assumed values for the model parameters."

Visually, "*f* of *Y*" is the height of the *conditional* normal curve that describes the spread of scores around a particular point on the regression line (e.g., the normal distribution of smoking intensity scores for participants who share the same value of the parental smoking indicator). The main component in the kernel is still a squared *z*-score, but that quantity now represents the standardized distance between a score and its predicted value. As before, the fraction to the left of the exponential function is a scaling term that ensures the area under the probability distribution sums or integrates to 1. Finally, note that explanatory variables function as fixed constants like the parameters. This feature will change in Chapter 3, where incomplete predictors appear as variables in a probability distribution.

As you know, maximum likelihood estimation reverses the probability distribution to get the likelihood of different combinations of population parameters given the observed data. Taking the natural logarithm of each observation's likelihood and summing the transformed probabilities gives a log-likelihood function that summarizes the data's evidence about the coefficients and residual variance.

$$\begin{aligned} LL\left(\boldsymbol{\beta}, \sigma_\varepsilon^2 \mid \text{data}\right) &= \sum_{i=1}^{N} \ln\left(\frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left(-\frac{1}{2}\frac{\left(Y_i - \left(\beta_0 + \beta_1 X_i\right)\right)^2}{\sigma_\varepsilon^2}\right)\right) \\ &= -\frac{N}{2}\ln\left(2\pi\right) - \frac{N}{2}\ln\left(\sigma_\varepsilon^2\right) - \frac{1}{2}\left(\sigma_\varepsilon^2\right)^{-1}\sum_{i=1}^{N}\left(Y_i - \left(\beta_0 + \beta_1 X_i\right)\right)^2 \qquad (2.34) \\ &= -\frac{N}{2}\ln\left(2\pi\right) - \frac{N}{2}\ln\left(\sigma_\varepsilon^2\right) - \frac{1}{2}\left(\sigma_\varepsilon^2\right)^{-1}\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right)'\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right) \end{aligned}$$

The compact matrix expression in the bottom row stacks the *N* outcome scores into a vector **Y**, and it uses **X** to denote a corresponding matrix that contains predictor variables and a column of ones for the intercept.

With only two coefficients, we can visualize the log-likelihood surface of β_0 and β_1 in three dimensions. Figure 2.13 is a contour plot conveying the perspective of a drone hovering over the peak of the log-likelihood surface, with smaller contours denoting higher elevation (and vice versa). The data's support for the parameters increases as the contours get smaller, and the maximum likelihood estimates of β_0 and β_1 are located at the peak of the surface, shown as a black dot. The angle of the ellipses owes to the fact that the intercept and slope coefficients are negatively correlated (i.e., the data's support for a larger mean difference requires concurrent support for lower comparison group average). Identifying the optimal parameters for the data is again analogous to a hiker climbing a mountain peak. Following the univariate example, we can derive an exact solution or use an iterative optimization approach such as Newton's algorithm.

Maximum Likelihood Estimates and Standard Errors

As before, the process of deriving maximum likelihood estimates and standard errors requires the first and second derivatives of the log-likelihood function. Applying differential calculus rules to Equation 2.34 leads to the following first derivative expressions:

$$\frac{\partial LL}{\partial \boldsymbol{\beta}} = -\frac{1}{\sigma_\varepsilon^2}\left(-\mathbf{X}'\mathbf{Y} + \mathbf{X}'\mathbf{X}\boldsymbol{\beta}\right) \tag{2.35}$$

$$\frac{\partial LL}{\partial \sigma_\varepsilon^2} = -\frac{N}{2}\left(\sigma_\varepsilon^2\right)^{-1} + \frac{1}{2}\left(\sigma_\varepsilon^2\right)^{-2}\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right)'\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right) \tag{2.36}$$

Setting these slope equations to 0 and solving for the unknown parameters at the peak of the log-likelihood surface gives the maximize likelihood estimates below.

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'Y = \hat{\boldsymbol{\beta}}_{\text{OLS}} \tag{2.37}$$

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{N}\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) \tag{2.38}$$

Notice that the coefficients are identical to those of ordinary least squares, but the residual variance differs, because the sample size is not adjusted for the number of estimates in $\hat{\boldsymbol{\beta}}$. This matches the earlier result for the mean and variance.

From Section 2.6, you know that second derivatives quantify the curvature or steepness of the log-likelihood function near its peak (i.e., the rate at which the first-

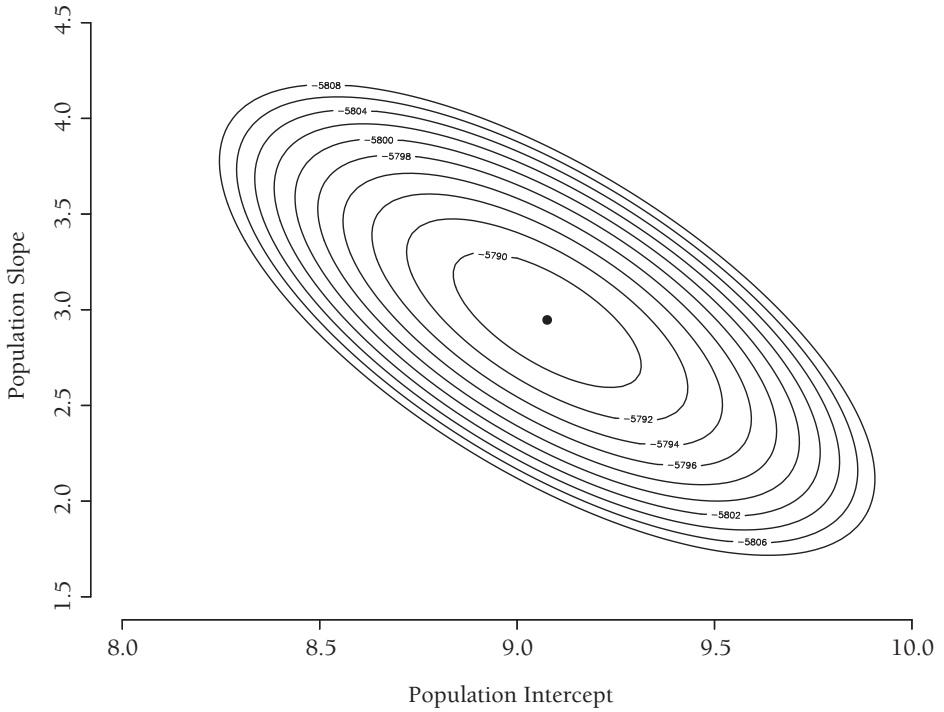

FIGURE 2.13. Contour plot that conveys the perspective of a drone hovering over the peak of the log-likelihood surface for a simple regression model, with smaller contours denoting higher elevation (and vice versa). The maximum likelihood estimates of β_0 and β_1 are located at the peak of surface (shown as a black dot).

order slopes change across the range of parameter values). These second derivatives are obtained by applying differential calculus rules to Equations 2.35 and 2.36, and the Hessian collects these equations in a matrix.

$$\mathbf{H}_{\mathrm{O}}(\boldsymbol{\theta}) = \begin{pmatrix} -\left(\sigma_{\varepsilon}^{2}\right)^{-1}\mathbf{X}'\mathbf{X} & -\left(\sigma_{\varepsilon}^{2}\right)^{-2}\mathbf{X}'\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\right)' \\ -\left(\sigma_{\varepsilon}^{2}\right)^{-2}\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\right)'\mathbf{X} & \frac{N}{2}\left(\sigma_{\varepsilon}^{2}\right)^{-2} - \left(\sigma_{\varepsilon}^{2}\right)^{-3}\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\right)'\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\right) \end{pmatrix} \quad (2.39)$$

Substituting the maximum likelihood estimates into the expression and multiplying $\mathbf{H}_{\mathrm{O}}(\hat{\boldsymbol{\theta}})$ by –1 gives the observed information matrix, then taking its inverse (the matrix analogue of a reciprocal) gives the variance–covariance matrix of the parameter estimates. Equations 2.17 and 2.18 depict these steps. The parameter covariance matrix for the simple regression analysis is symmetric with three rows and columns, one per parameter.

$$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\theta}}} = \begin{pmatrix} \mathrm{var}\left(\hat{\beta}_0\right) & \mathrm{cov}\left(\hat{\beta}_0,\hat{\beta}_1\right) & \mathrm{cov}\left(\hat{\beta}_0,\hat{\sigma}_{\varepsilon}^{2}\right) \\ \mathrm{cov}\left(\hat{\beta}_1,\hat{\beta}_0\right) & \mathrm{var}\left(\hat{\beta}_1\right) & \mathrm{cov}\left(\hat{\beta}_1,\hat{\sigma}_{\varepsilon}^{2}\right) \\ \mathrm{cov}\left(\hat{\sigma}_{\varepsilon}^{2},\hat{\beta}_0\right) & \mathrm{cov}\left(\hat{\sigma}_{\varepsilon}^{2},\hat{\beta}_1\right) & \mathrm{var}\left(\hat{\sigma}_{\varepsilon}^{2}\right) \end{pmatrix} = \begin{pmatrix} 0.018 & -0.018 & 0 \\ -0.018 & 0.039 & 0 \\ 0 & 0 & 0.365 \end{pmatrix} \quad (2.40)$$

Finally, taking the square root of the diagonal elements gives the standard errors (e.g., $SE_{\hat{\beta}_1} = \sqrt{0.039} = 0.20$). To establish further linkages to ordinary least squares, the expression in the upper left block of Equation 2.39 is a 2 × 2 matrix that contains derivatives with respect to the two coefficients. Multiplying this submatrix by –1 and taking its inverse gives an expression that is identical to a parameter covariance matrix from ordinary least squares regression.

Analysis Example

To illustrate maximum likelihood estimation for multiple regression, I expanded the previous analysis model to include age and income as predictors. I centered the additional variables at their grand means to maintain the intercept's interpretation as the expected smoking intensity score for a respondent whose parents did not smoke.

$$INTENSITY_i = \beta_0 + \beta_1\left(PARSMOKE_i\right) + \beta_2\left(AGE_i - \mu_2\right) + \beta_3\left(INCOME_i - \mu_3\right) + \varepsilon_i \quad (2.41)$$

Importantly, the smoking intensity distribution has substantial positive skewness and kurtosis, so I used robust (sandwich estimator) standard errors and the bootstrap to illustrate different corrective procedures. Analysis scripts are available on the companion website, including a custom R program for readers interested in coding Newton's algorithm by hand.

Table 2.5 shows the maximum likelihood estimates, along with ordinary least squares results as a comparison. As expected, the two estimators produced identical

TABLE 2.5. Maximum Likelihood and Ordinary Least Squares Estimates

	Maximum likelihood				OLS	
Parameter	Est.	*SE*	*RSE*	*BSE*	Est.	*SE*
β_0	9.09	0.126	0.120	0.119	9.09	0.126
β_1 (*PARSMOKE*)	2.91	0.187	0.186	0.183	2.91	0.187
β_2 (*AGE*)	0.59	0.040	0.040	0.040	0.59	0.040
β_3 (*INCOME*)	–0.10	0.027	0.032	0.032	–0.10	0.027
σ_ε^2	17.15	0.542	1.673	1.685	17.18	—
R^2	0.19	0.016	0.026	0.025	0.19	—

Note. RSE, robust standard errors; *BSE,* bootstrap standard errors.

coefficients, but the maximum likelihood residual variance is very slightly smaller, because it does not subtract the four degrees of freedom spent estimating the coefficients. This slight difference aside, the estimates themselves have the same meaning. For example, the intercept ($\hat{\beta}_0 = 9.09$, $SE = .12$) is the expected number of cigarettes smoked per day for a respondent whose parents didn't smoke, and the parental smoking indicator slope ($\hat{\beta}_1 = 2.91$, $SE = .19$) is the mean difference, controlling for age and income. The corrective procedures induced relatively minor changes to the coefficients' standard errors, but they had a dramatic impact on the standard error of the residual variance. As is often the case with a reasonably large sample size, sandwich estimator and bootstrap standard errors were effectively equivalent.

2.11 SIGNIFICANCE TESTS

Maximum likelihood estimation offers three significance testing options: the Wald test (Wald, 1943), likelihood ratio statistic (Wilks, 1938), and the score test or Lagrange multiplier (Rao, 1948). The latter is commonly referred to as the modification index in structural equation modeling applications (Saris, Satorra, & Sörbom, 1987; Sörbom, 1989). I describe the first two approaches, because they are widely available in general-purpose software packages, and Buse (1982) provides a nice tutorial on this "trilogy of tests" for readers who are interested in additional details.

The Wald test and likelihood ratio statistic can evaluate the same hypotheses, but they do so in different ways. The Wald test compares the discrepancy between the estimates and hypothesized parameter values (usually zeros) to sampling variation. The simplest incarnation of the test statistic is just a *z*-score or chi-square. In contrast, the likelihood ratio statistic compares log-likelihood values from two competing models, the simpler of which aligns with the null hypothesis. The two tests are equivalent in very large samples but can give markedly different answers in small to moderate samples (Buse, 1982; Fears, Benichou, & Gail, 1996; Greene, 2017; Pawitan, 2000). These differences are sometimes attributable to the fact that the Wald test inappropriately assumes that sampling distributions follow a normal curve, but discrepancies can arise for other

reasons that are more difficult to predict. Statistical issues aside, the likelihood ratio test is somewhat less convenient to implement, because it requires two analyses, but this is not a compelling disadvantage.

Wald Test

The simplest incarnation of the Wald test is the familiar *z*-statistic that compares the difference between an estimate and hypothesized parameter value (e.g., $\theta_0 = 0$) to the estimate's standard error.

$$z = \frac{\hat{\theta} - \theta_0}{SE_{\hat{\theta}}} \tag{2.42}$$

Leveraging the large-sample normality of maximum likelihood estimates, a standard normal distribution generates a probability value for the test, and symmetrical confidence interval limits are computed by multiplying the standard error by the appropriate *z* critical value, then adding and subtracting that product (i.e., the margin of error or half-width) to the estimate.

$$\text{CI} = \hat{\theta} \pm z_{\text{CV}} \times SE_{\hat{\theta}} \tag{2.43}$$

The *z* critical values for different alpha levels are available in textbooks and online (e.g., $z_{\text{CV}} = \pm 1.96$ for $\alpha = .05$).

Squaring the *z*-score gives an alternative expression for the Wald statistic that instead follows a chi-square distribution with a single degree of freedom.

$$T_{\text{W}} = \left(\frac{\hat{\theta} - \theta_0}{SE_{\hat{\theta}}} \right)^2 = \frac{\left(\hat{\theta} - \theta_0\right)\left(\hat{\theta} - \theta_0\right)}{\text{var}\left(\hat{\theta}\right)} \tag{2.44}$$

The chi-square formulation readily generalizes to multiple parameters:

$$T_{\text{W}} = \left(\hat{\boldsymbol{\theta}}_{\text{Q}} - \boldsymbol{\theta}_0\right)' \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\theta}}_{\text{Q}}}^{-1} \left(\hat{\boldsymbol{\theta}}_{\text{Q}} - \boldsymbol{\theta}_0\right) \tag{2.45}$$

where $\hat{\boldsymbol{\theta}}_{\text{Q}}$ is a vector of Q estimates, $\boldsymbol{\theta}_0$ is the corresponding vector of hypothesized values (typically zeros), and $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\theta}}_{\text{Q}}}$ is a variance–covariance matrix that contains Q rows and columns from full parameter covariance matrix (or its robustified counterpart). The numerical value of T_{W} is the sum of squared standardized differences between the estimates and their hypothesized values. If the null hypothesis is true, the test statistic follows a central chi-square distribution with Q degrees of freedom, and statistical significance implies that one or more of the estimates in $\hat{\boldsymbol{\theta}}_{\text{Q}}$ are different from their hypothesized values.

Likelihood Ratio Test

The likelihood ratio statistic evaluates the relative fit of two **nested models**. Nested models can take a variety of forms, but a common application compares the substantive

analysis to a more restrictive version of the model that fixes a subset of parameters to 0. Returning to the earlier regression analysis, we could use the likelihood ratio statistic to evaluate the null hypothesis that $R^2 = 0$ by comparing the fit of the analysis model from Equation 2.41 to that of an empty model that constrains the slope coefficients to 0. A slightly different application of the likelihood ratio test occurs in structural equation modeling analyses in which a researcher compares the fit of a saturated model (i.e., a model that places no restrictions on the mean vector and covariance matrix) to that of a more parsimonious analysis model that imposes a structure on the data (e.g., a confirmatory factor analysis model). In either scenario, the simpler model with Q fewer parameters aligns with the null hypothesis, so I denote the restricted model's parameters as $\boldsymbol{\theta}_0$ and the full model's parameters as $\boldsymbol{\theta}$.

The likelihood ratio statistic is

$$T_{\text{LR}} = -2\left(LL\left(\hat{\boldsymbol{\theta}}_0 \mid \text{data}\right) - LL\left(\hat{\boldsymbol{\theta}} \mid \text{data}\right)\right) \tag{2.46}$$

where $LL(\hat{\boldsymbol{\theta}}_0|\text{data})$ is the sample log-likelihood value for the restricted model (e.g., an empty regression model with only an intercept), and $LL(\hat{\boldsymbol{\theta}}|\text{data})$ is the log-likelihood for the more complex model (e.g., the full regression model). The more complex model with additional parameters will always achieve better fit and a higher log-likelihood, but that improvement should be very small when the null hypothesis is true. If the two models are equivalent in the population, the likelihood ratio statistic follows a central chi-square distribution with Q degrees of freedom, which in this case is the difference between the number of parameters in the two models. A significant test statistic indicates that the data provide more support for the full model than the restricted model (e.g., one or more parameters are significantly different from zero).

Robust Test Statistics

As discussed in Section 2.8, non-normal data may or may not compromise point estimates, but they certainly distort standard errors. The same is true for significance tests, as the Wald and likelihood ratio statistics no longer follow the optimal chi-square distribution. The Wald test is easily robustified by substituting a sandwich estimator covariance matrix into Equation 2.45 (or a robust standard error into Equation 2.42). The likelihood ratio statistic can be rescaled to more closely approximate the correct chi-square distribution (Satorra & Bentler, 1988; Satorra & Bentler, 1994; Yuan & Bentler, 2000), or a p-value can be obtained by referencing the biased test statistic against a bootstrap sampling distribution that honors the distribution of the data (Beran & Srivastava, 1985; Bollen & Stine, 1992; Enders, 2002; Savalei & Yuan, 2009). I describe these two approaches below and illustrate their application in one of the later analysis examples.

Readers familiar with structural equation models are undoubtedly familiar with the rescaled likelihood ratio statistic, which is commonly known as the **Satorra–Bentler chi-square** (Satorra & Bentler, 1988; Satorra & Bentler, 1994). The general procedure for comparing two nested models involves dividing the likelihood ratio statistic by a constant scaling term that largely depends on the kurtosis of the data (Satorra & Bentler, 2001; Yuan et al., 2005). The rescaled test statistic is

$$T_{SB} = \frac{T_{LR}}{c_{LR}} \tag{2.47}$$

$$c_{LR} = \frac{P_0 c_0 - P_F c_F}{P_0 - P_F}$$

where T_{LR} is the likelihood ratio statistic from Equation 2.46, and c_{LR} is a scaling constant that combines the number of parameters in the full and restricted models, P_F and P_0, respectively, and model-specific scaling terms, c_F and c_0.

The scaling term can be understood by revisiting the sandwich estimator covariance matrix in Equation 2.24. As explained previously, the "bread × meat" product yields a matrix with diagonal elements that reflect the relative magnitude of two information matrices, one of which is sensitive to outlier scores. When the data are normal, the two matrices are equivalent and cancel out when multiplying one by the inverse of the other (the resulting product is an inert identity matrix). In contrast, when the data are non-normal, the resulting product contains fractional diagonal terms that can be smaller or larger than 1, depending on the kurtosis of the data. Multiplying this matrix by the rightmost piece of "bread" inflates or deflates elements in parameter covariance matrix accordingly.

The rescaling terms for the likelihood ratio test also leverage discrepancies between the two information matrices. In the simplest possible univariate application (e.g., the analysis from Section 2.8), the scaling term is a fraction that compares a single diagonal value from each information matrix (Yuan et al., 2005). More generally, c_F and c_0 pool the elements of the "bread × meat" product into a single scalar value that rescales the test statistic to have the same expected value or mean as its optitmal central chi-square distribution (Satorra & Bentler, 1988, 1994, 2001). As such, referencing T_{SB} to a chi-square distribution with Q degrees of freedom gives an approximate p-value, and a significant test statistic indicates that the data provide more support for the full model than the restricted model (e.g., one or more parameters are significantly different from 0).

A second option for getting a robust significance test is to use the original T_{LR} from Equation 2.46 but reference the test statistic against a simulation-based bootstrap sampling distribution that honors the data's shape. This is essentially the opposite tack of rescaling, which fixes up the test statistic and leaves the theoretical sampling distribution intact. As explained in Section 2.8, the bootstrap procedure treats the observed data as a surrogate for the population and draws many samples of size N with replacement. Fitting the analysis model to each data set produces a collection of estimates that form empirical sampling distributions, the standard deviations of which are robust standard errors. A slight modification is needed to apply the bootstrap to test statistics. As you know, a probability value reflects the likelihood that the observed test statistic originated from a hypothetical population where the null hypothesis is exactly true. To achieve this interpretation from the bootstrap, you need to first transform the observed data to match the null hypothesis. Returning to the multiple regression model from Equation 2.41, a null hypothesis that $R^2 = 0$ implies that all regression slopes equal 0. The estimated slopes will never be exactly 0, yet the sample data must be exactly consistent with this condition for the bootstrap to work properly.

Beran and Srivastava (1985) and Bollen and Stine (1992) modified the bootstrap procedure by first applying an algebraic transformation that aligns the mean and covariance structure of the data to the null hypothesis (the procedure is sometimes referred to as the **model-based bootstrap**). Importantly, this transformation does not modify distribution shapes, so drawing bootstrap samples from the rescaled data gives an empirical sampling distribution that reflects the natural variation of the test statistic with non-normal data. A robust *p*-value is then obtained by computing the proportion of bootstrap samples that give a test statistic larger than T_{LR} from the original analysis. The transformation expression is

$$\tilde{\mathbf{Y}}_i = \left(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}\right)' \hat{\boldsymbol{\Sigma}}^{-.5} \hat{\boldsymbol{\Sigma}}_0^{-.5} + \hat{\boldsymbol{\mu}}_0' \tag{2.48}$$

where $\tilde{\mathbf{Y}}_i$ is the transformed data for observation i, $\mathbf{Y}_i$ is the corresponding vector of observed scores, $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ are the mean vector and covariance matrix of the sample data, and $\hat{\boldsymbol{\mu}}_0$ and $\hat{\boldsymbol{\Sigma}}_0$ are model-implied mean vector and covariance matrix from the restricted model (i.e., the model that aligns with the null hypothesis). I use Y as a generic symbol for the analysis variables, but this vector could include predictors and outcomes. The equation essentially applies two transformations: The $(\mathbf{Y}_i - \hat{\boldsymbol{\mu}})'\hat{\boldsymbol{\Sigma}}^{-.5}$ part of the expression "erases" the mean and the covariance structure from the data by converting the variables to uncorrelated z-scores, and $\hat{\boldsymbol{\Sigma}}_0^{-.5} + \hat{\boldsymbol{\mu}}_0'$ rescales the z-scores to match the associations implied by the null hypothesis. Returning to the math achievement regression model from Equation 2.41, a null hypothesis that $R^2 = 0$ would induce a transformation where explanatory variables are correlated with each other but mutually uncorrelated with the outcome. Applying the bootstrap procedure to the rescaled data and collecting the B test statistics creates an empirical sampling distribution, and the robust probability value is then the proportion of these statistics that exceed T_{LR}, the likelihood ratio statistic from the raw data.

Analysis Example

Returning to the multiple regression model from Equation 2.41, I use the Wald test and likelihood ratio statistic to evaluate the null hypothesis that $R^2 = 0$. Both tests function like the omnibus F test from ordinary least squares in this context. To begin, the Wald test standardizes discrepancies between the estimates and null values against the parameter covariance matrix. The full covariance matrix is a 5 × 5 matrix, but the test uses only the elements related to the slope coefficients. The composition of the test statistic for this example is as follows:

$$T_W = \begin{pmatrix} \hat{\beta}_1 & 0 \\ \hat{\beta}_2 & -0 \\ \hat{\beta}_3 & 0 \end{pmatrix}' \begin{pmatrix} \text{var}\left(\hat{\beta}_1\right) & \text{cov}\left(\hat{\beta}_1, \hat{\beta}_2\right) & \text{cov}\left(\hat{\beta}_1, \hat{\beta}_3\right) \\ \text{cov}\left(\hat{\beta}_2, \hat{\beta}_1\right) & \text{var}\left(\hat{\beta}_2\right) & \text{cov}\left(\hat{\beta}_2, \hat{\beta}_3\right) \\ \text{cov}\left(\hat{\beta}_3, \hat{\beta}_1\right) & \text{cov}\left(\hat{\beta}_3, \hat{\beta}_2\right) & \text{var}\left(\hat{\beta}_3\right) \end{pmatrix}^{-1} \begin{pmatrix} \hat{\beta}_1 & 0 \\ \hat{\beta}_2 & -0 \\ \hat{\beta}_3 & 0 \end{pmatrix} \tag{2.49}$$

The diagonal elements of the middle matrix are the sampling variances (i.e., squared

standard errors), and the off-diagonal elements capture the degree to which the estimates covary across repeated samples. Substituting the appropriate estimates into the previous expression gives $T_W = 481.19$, the value of which represents the sum of squared standardized differences from zero. Referencing the test statistic to a chi-square distribution with $Q = 3$ degrees of freedom gives $p < .001$; consistent with an analogous F test, we can conclude that at least one of the slopes is nonzero. The sandwich estimator (robust) test statistic was markedly lower at $T_W = 423.15$ but gave the same conclusion.

The likelihood ratio statistic evaluates the same hypothesis but requires a nested or restricted model that aligns with the null. This secondary model is an empty regression that fixes the three slope coefficients to zero. With complete data, you can get the restricted model log-likelihood by constraining the slope coefficients to zero during estimation or by excluding the explanatory variables from the analysis. Although it makes no difference here, explicitly constraining the slopes to zero as follows is preferable, because it generalizes to missing data analyses.

$$INTENSITY_i = \beta_0 + (0)(PARSMOKE_i) + 0(AGE_i - \mu_2) + 0(INCOME_i - \mu_3) + \varepsilon_i \quad (2.50)$$

Fitting the two models and substituting the resulting log-likelihood values into Equation 2.46 gives the following test statistic:

$$T_{LR} = -2\left(LL\left(\hat{\theta}_0 \mid \text{data}\right) - LL\left(\hat{\theta} \mid \text{data}\right)\right) = 2\left((-5{,}895.145) - (-5{,}679.545)\right) = 431.20 \quad (2.51)$$

As you can see, fixing the slopes to zero substantially decreased the log-likelihood from –5,679.545 to –5,895.145, indicating that the restricted model's parameters are located at a much lower vertical elevation on the log-likelihood surface. Referencing the test statistic to a chi-square distribution with $Q = 3$ degrees of freedom returns a probability value of $p < .001$, which, again, indicates that one or more of the slopes' coefficients are nonzero. The corresponding rescaled test statistic from Equation 2.47 was markedly lower at $T_{SB} = 173.97$ ($c_{LR} = 2.48$) but gave the same conclusion. Although T_W and T_{LR} produced the same substantive conclusion, their numerical values aren't particularly well calibrated. This is not unusual, as the tests often require a much larger sample size to achieve equivalence.

2.12 MULTIVARIATE NORMAL DATA

The multivariate normal distribution plays an important role throughout the book, and it appears prominently in Chapter 3, where it provides a flexible framework for missing data handling. To set the stage for missing data, this section uses the distribution as a backdrop for estimating a mean vector and variance–covariance matrix. As you will see, the concepts we've already established readily generalize to multivariate data with virtually no modifications (although some of the equations are messier). I use the employee data from the companion website to provide a substantive context. The data set includes several workplace-related variables (e.g., work satisfaction, turnover inten-

tion, employee–supervisor relationship quality) for a sample of $N = 630$ employees. The illustration uses a 7-point work satisfaction rating (1 = *extremely dissatisfied* to 7 = *extremely satisfied*) and two composite scores that measure employee empowerment and a construct known as leader–member exchange scale (the quality of an employee's relationship with his or her supervisor). I treat work satisfaction as a normally distributed variable, because it has a sufficient number of response options and a symmetric distribution (Rhemtulla et al., 2012). The Appendix gives a description of the data set and variable definitions.

Probability Distribution and Log-Likelihood

To tie the multivariate normal distribution back to earlier material, it is useful to cast the analysis as three empty regression models. Using generic notation, the models are as follows:

$$\mathbf{Y}_i = \begin{pmatrix} WORKSAT_i \\ EMPOWER_i \\ LMX_i \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \varepsilon_{3i} \end{pmatrix} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} \tag{2.52}$$

$$\mathbf{Y}_i \sim N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

The bottom equation is shorthand notation to reference data that follow a multivariate normal distribution; N_3 denotes a three-dimensional normal distribution, and the first and second terms in parentheses are the mean vector and variance–covariance matrix (the multivariate distribution's parameters).

The multivariate normal distribution function generalizes the normal curve to multiple variables. In addition to a mean and variance for each variable, the distribution also incorporates covariances among the variables (or alternatively, correlated residuals). To illustrate, Figure 2.14 shows an idealized bivariate normal distribution for the pain interference and depression composite variables. The distribution retains its familiar shape and looks like a bell-shaped surface in three-dimensional space. The probability distribution function that describes the shape of the surface has the same basic structure as its univariate sibling in Equation 2.3, with vectors and matrices replacing scalar quantities.

$$f(\mathbf{Y}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{(-V \times .5)} |\boldsymbol{\Sigma}|^{-.5} \exp\left(-\frac{1}{2} (\mathbf{Y}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}) \right) \tag{2.53}$$

The column vector $\mathbf{Y}_i$ now contains V observations for a participant i, $\boldsymbol{\mu}$ is the corresponding vector of population means, and $\boldsymbol{\Sigma}$ is a variance–covariance matrix of the V variables. As before, the function on the left side of the expression can be read as "the relative probability of the V observations given assumed values for the model parameters." Visually, the equation describes the height of the surface in Figure 2.14 at the intersection of score values along the horizontal and depth axes. The term in the exponential function, $(\mathbf{Y}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{Y}_i - \boldsymbol{\mu})$, is a key component that equals the sum of squared

standardized differences between the scores and the distribution's center (a quantity known as **Mahalanobis distance**). Finally, the terms to the left of the exponential function scale the distribution so the area under the surface sums or integrates to 1.

As you know, a probability distribution treats scores as variable and the parameters as known constants. To illustrate the distribution function's output, assume that the true population parameters are as follows (these happen to be the maximum likelihood estimates for the employee empowerment and leader–member exchange variables):

$$\boldsymbol{\mu} = \begin{pmatrix} 28.61 \\ 9.59 \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} 20.38 & 5.37 \\ 5.37 & 9.10 \end{pmatrix} \tag{2.54}$$

The contour plot in Figure 2.15 shows the perspective of a drone hovering over the peak of the bivariate normal distribution in Figure 2.14, with smaller contours denoting higher elevation and larger relative probabilities (and vice versa). The overhead perspective better reveals the positive correlation between pain interference and depression. The black diamond corresponds to interference and depression scores of $\mathbf{Y}_1$ = (32.00, 13.18)′, and the black circle corresponds to $\mathbf{Y}_2$ = (33.25, 9)′. Substituting everything into

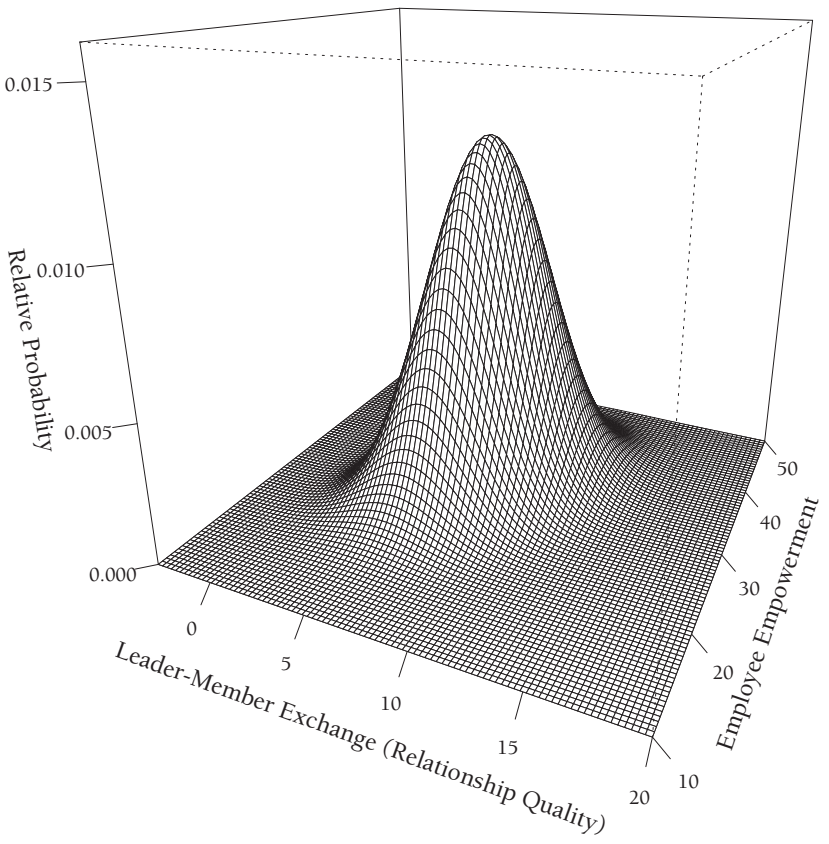

FIGURE 2.14. An idealized bivariate normal probability distribution for the employee empowerment and leader–member exchange variables.

FIGURE 2.15. The contour plot shows the perspective of a drone hovering over the peak of the bivariate normal distribution in Figure 2.14, with smaller contours denoting higher elevation and larger relative probabilities (and vice versa). The overhead perspective better reveals the positive correlation between pain interference and depression. The black circle and diamond are two pairs of scores located at the same vertical elevation.

Equation 2.53 returns relative probability values of $f(\mathbf{Y}_1|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = 0.006$ and $f(\mathbf{Y}_2|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = 0.006$. The two pairs of scores have the same relative probability (i.e., are located at the same vertical elevation), despite the fact that the straight line connecting Y_1 to the center of the distribution is noticeably shorter than the line connecting Y_2 to the peak. This result happens, because the positive correlation rotates the contours in such a way that elevation drops rapidly directly above and below the distribution's peak. This feature is also apparent in Equation 2.53, where scaling the squared deviation scores relative to the variance–covariance matrix standardizes the distances in a way that accounts for the correlations among the variables.

Following established concepts, estimation "reverses" the probability distribution's arguments to get the likelihood of different combinations of population parameters given the observed data. Taking the natural logarithm gives the log-likelihood contribution for a single observation:

$$LL_i\left(\boldsymbol{\mu},\boldsymbol{\Sigma}\,|\,\text{data}\right) = -\frac{V}{2}\ln\left(2\pi\right) - \frac{1}{2}\ln\left|\boldsymbol{\Sigma}\right| - \frac{1}{2}\left(\mathbf{Y}_i - \boldsymbol{\mu}\right)' \boldsymbol{\Sigma}^{-1}\left(\mathbf{Y}_i - \boldsymbol{\mu}\right) \tag{2.55}$$

and summing across the N observations gives the sample log-likelihood.

$$LL(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \text{data}) = -N\frac{V}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{N}(\mathbf{Y}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i - \boldsymbol{\mu}) \tag{2.56}$$

Numerically, the log-likelihood is a large negative value that summarizes the data's evidence for a specific combination of parameter values in $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, with higher or less negative numbers reflecting better fit (and vice versa). Visually, the log-likelihood corresponds to the height of a multidimensional surface at specific values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. As always, the goal of estimation is to identify the parameter values that maximize fit to the observed data (or equivalently, minimize the sum of the squared z-scores in the rightmost term).

Maximum Likelihood Estimates and Standard Errors

Consistent with the previous examples, we can derive an exact solution for the mean vector and covariance matrix or use an iterative optimization approach such as Newton's algorithm. An exact solution requires first and second derivatives of the log-likelihood function. The underlying logic is the same as before—solve for the parameters after setting the derivative expressions to 0—but getting the derivative expressions is more complex and requires matrix calculus (Magnus & Neudecker, 1999). Although most of the equations are not intuitive, I include them as a resource for interested readers. Equations aside, you can still follow the gist of estimation, because all quantities retain their previous meaning (e.g., a first derivative gives the slope at a particular point on the log-likelihood surface; a second derivative captures curvature or steepness at the peak).

The first derivatives with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are as follows:

$$\frac{\partial LL}{\partial \boldsymbol{\mu}} = -N\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\Sigma}^{-1}\sum_{i=1}^{N}\mathbf{Y}_i \tag{2.57}$$

$$\frac{\partial LL}{\partial \boldsymbol{\Sigma}} = -\frac{1}{2}\sum_{i=1}^{N}\left(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i - \boldsymbol{\mu})(\mathbf{Y}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}\right) \tag{2.58}$$

Setting these equations to 0 and solving for the parameters gives the following analytic solutions for the maximum likelihood estimates:

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{Y}_i \tag{2.59}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{Y}_i - \hat{\boldsymbol{\mu}})(\mathbf{Y}_i - \hat{\boldsymbol{\mu}})' \tag{2.60}$$

The analytic solutions highlight a recurring theme, which is that maximum likelihood estimates of variances and covariances do not adjust for the degrees of freedom spent estimating the means; as such, variance–covariance estimates are biased in small samples but approach their true population values as sample size increases (i.e., the estimates are said to be consistent).

Second derivatives quantify the curvature or steepness of the log-likelihood function near its peak (i.e., the rate at which the first-order slopes change across the range of parameter values). Second derivatives are obtained by applying matrix calculus rules to Equations 2.57 and 2.58, and the Hessian collects these equations in a symmetric matrix with P rows and columns, where P is the number of unique parameters in $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

$$\mathbf{H}_{\mathrm{O}}(\boldsymbol{\theta}) = \begin{pmatrix} \dfrac{\partial^2 LL}{\partial \boldsymbol{\mu}^2} & \dfrac{\partial^2 LL}{\partial \boldsymbol{\mu} \partial \boldsymbol{\Sigma}} \\ \dfrac{\partial^2 LL}{\partial \boldsymbol{\Sigma} \partial \boldsymbol{\mu}} & \dfrac{\partial^2 LL}{\partial \boldsymbol{\Sigma}^2} \end{pmatrix} \tag{2.61}$$

The second derivative equations below are the building blocks for the observed information matrix, and analogous expressions for the expected information are available in the literature (Savalei, 2010; Savalei & Bentler, 2009; Yuan & Hayashi, 2006) and in Chapter 3.

$$\frac{\partial^2 LL}{\partial \boldsymbol{\mu}^2} = -N\boldsymbol{\Sigma}^{-1} \tag{2.62}$$

$$\frac{\partial^2 LL}{\partial \boldsymbol{\Sigma}^2} = -\sum_{i=1}^{N} \mathbf{D}_V' \left(\boldsymbol{\Sigma}^{-1} \otimes \left\{ \boldsymbol{\Sigma}^{-1} \left(\mathbf{Y}_i - \boldsymbol{\mu} \right) \left(\mathbf{Y}_i - \boldsymbol{\mu} \right)' \boldsymbol{\Sigma}^{-1} - .5\boldsymbol{\Sigma}^{-1} \right\} \right) \mathbf{D}_V$$

$$\frac{\partial^2 LL}{\partial \boldsymbol{\mu} \partial \boldsymbol{\Sigma}} = -\sum_{i=1}^{N} \left(\boldsymbol{\Sigma}^{-1} \otimes \left(\mathbf{Y}_i - \boldsymbol{\mu} \right)' \boldsymbol{\Sigma}^{-1} \right) \mathbf{D}_V$$

The $\otimes$ symbol is a Kronecker product that multiplies one matrix by each element of another matrix, and $\mathbf{D}_V$ is the so-called "duplication matrix" (Magnus & Neudecker, 1999). Each covariance term appears twice in the first derivative matrix from Equation 2.58 but only once in the Hessian (and similarly, only once in the parameter covariance matrix). The duplication matrix combines these redundant terms into a single value. Substituting the maximum likelihood estimates into the derivative expressions, multiplying $\mathbf{H}_{\mathrm{O}}(\hat{\boldsymbol{\theta}})$ by –1, then taking its inverse gives the variance–covariance matrix of the estimates.

Analysis Example

Returning to the empty regression models in Equation 2.52, I use work satisfaction, employee empowerment, and leader–member exchange scales to illustrate maximum likelihood estimation. Analysis scripts are available on the companion website, including a custom R program for readers interested in coding Newton's algorithm by hand. Table 2.6 gives the maximum likelihood estimates of the means, standard deviations, variances and covariances, and correlations (in bold typeface above the diagonal). I computed the standard deviations and correlations by transforming the maximum likelihood estimates of the variances and covariances (e.g., a correlation is a covariance divided by square root of the product of two variances). As a comparison, Table 2.6 also gives results from the usual unbiased estimator of the variance–covariance matrix. The

TABLE 2.6. Maximum Likelihood Descriptive Statistics

Variable	1	2	3
	Maximum likelihood		
1. *WORKSAT*	1.58	**.29**	**.42**
2. *EMPOWER*	1.64	20.38	**.39**
3. *LMX*	1.61	5.37	9.10
Means	3.99	28.61	9.59
SD	1.26	4.52	3.02
	Unbiased sample estimates		
1. *WORKSAT*	1.59	**.29**	**.42**
2. *EMPOWER*	1.64	20.42	**.39**
3. *LMX*	1.61	5.37	9.11
Means	3.99	28.61	9.59
SD	1.26	4.52	3.02

Note. **Bold** typeface denotes correlations.

maximum likelihood estimates of these parameters are consistently lower (albeit by a trivial amount), because the estimator from Equation 2.60 has N rather than $N - 1$ in the denominator.

2.13 CATEGORICAL OUTCOMES: LOGISTIC AND PROBIT REGRESSION

Looking ahead to missing data analyses, we now have flexible estimators that accommodate mixtures of categorical and continuous incomplete variables. To set the stage for later examples, I illustrate complete-data maximum likelihood estimation for a binary outcome variable. Continuing with the employee data set, I use a dichotomous measure of turnover intention that equals 0 if an employee has no plan to leave his or her position and 1 if the employee has the intention of quitting. The bar graph in Figure 2.16 shows the distribution of the discrete responses.

Latent Response Variable Formulation

Logit and probit regression envision binary scores originating from an underlying latent response variable that represents one's underlying proclivity or propensity to endorse the highest category (Agresti, 2012; Johnson & Albert, 1999). Applied to the turnover intention measure, this latent variable represents an unobserved, continuous dimension of quitting intentions. To illustrate, Figure 2.17 shows the latent variable distribution for the bar graph in Figure 2.16. The vertical line represents the precise cutoff point or threshold in the latent distribution where discrete scores switch from 0 to 1 (or more generally, from the lowest code to the highest code). The areas under the curve above and

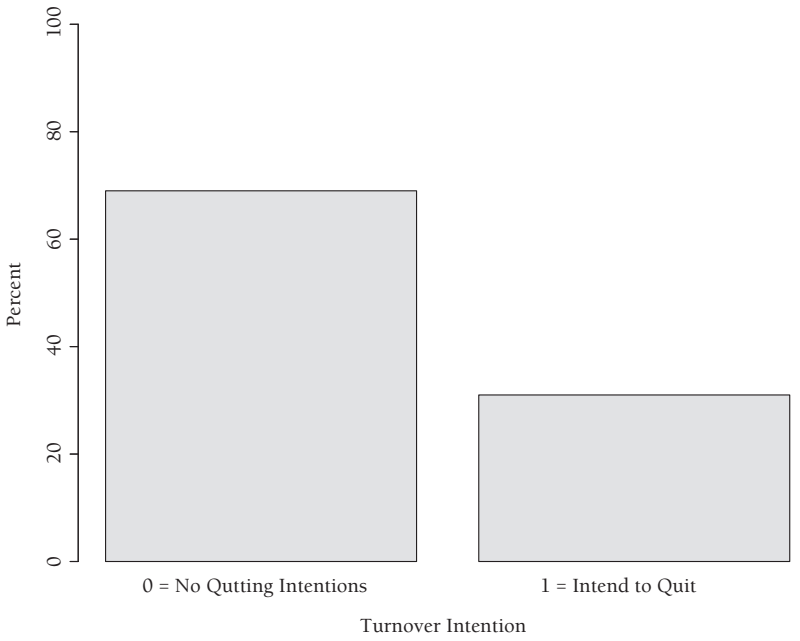

FIGURE 2.16. Bar graph of the dichotomous measure of turnover intention. *TURNOVER* = 0 if an employee has no plan to leave his or her position, and *TURNOVER* = 1 if the employee has intentions of quitting.

below this threshold correspond to the category proportions in the bar chart: 69% of the area under the curve falls below the threshold, and 31% falls above in the shaded region. Using generic notation, the link between the latent scores and categorical responses is

$$Y_i = \begin{cases} 0 \text{ if } Y_i^* \leq \tau \\ 1 \text{ if } Y_i^* > \tau \end{cases} \tag{2.63}$$

where Y_i is the binary outcome for individual i, Y_i^* is the corresponding latent response score, and τ is the threshold parameter (the vertical line in Figure 2.17). Fixing the latent response variable's mean or its threshold parameter to 0 provides a necessary identification constraint, and I always adopt the latter strategy.

Adding an explanatory variable to the latent response model is a relatively small step forward. To illustrate, consider a simple regression with leader–member exchange (employee–supervisor relationship quality) predicting turnover intention, the latent variable model for which is as follows:

$$TURNOVER_i^* = \beta_0 + \beta_1(LMX_i) + \epsilon_i \tag{2.64}$$

The key difference between logistic and probit regression is the distribution of the residual term—the probit model defines ϵ_i as a standard normal variable, whereas logis-

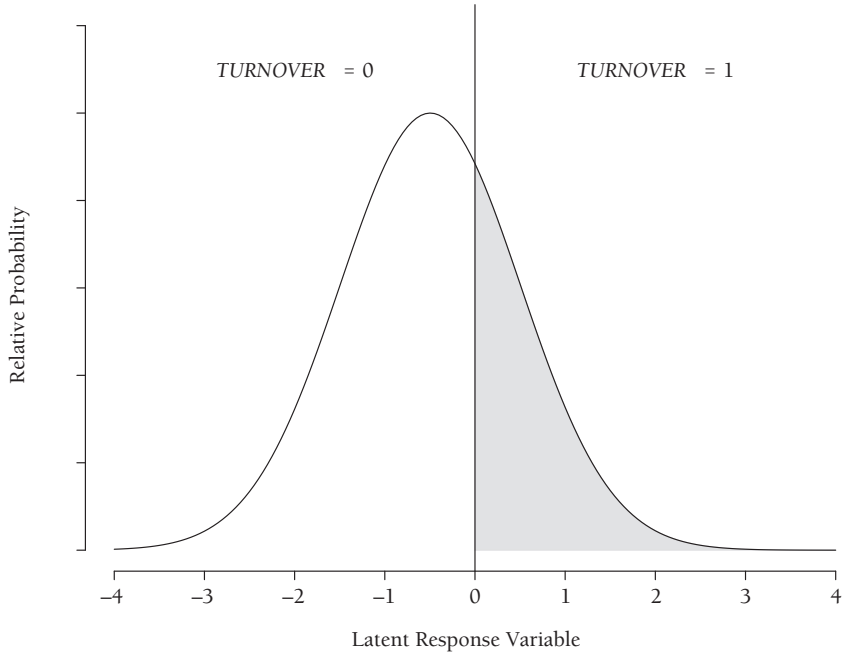

FIGURE 2.17. Latent response distribution for a binary variable. The vertical line at 0 is a threshold parameter τ that divides the latent distribution into two regions. Employees with no quitting intentions have latent scores below the threshold, and employees who intend to quit have scores above the threshold. The area under the shaded region of the curve is the probability of quitting (the proportion of 1's in the data).

tic regression defines the residual as a standard logistic variable. To illustrate a probit regression model, Figure 2.18 shows the latent variable distributions at three values of the explanatory variable, with the area above the threshold parameter (the predicted probabilities) shaded in gray. The black dots represent predicted values, and the contour rings convey the perspective of a drone hovering over the peak of a bivariate normal distribution, with smaller contours denoting higher elevation (and vice versa). The graph for a logistic regression is similar, but standard logistic distributions have thicker tails than the normal curves in the figure.

Going forward, I use the following notation for probit regression models to emphasize the normally distributed latent response variable, which later functions as an incomplete variable to be imputed:

$$Y_i^* = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad (2.65)$$
$$\varepsilon_i \sim N_1(0,1)$$

The second term in the normal distribution function indicates that the latent response variable's variance is fixed at 1 to provide a metric. I write the logistic model in its more usual format as

$$\ln\left(\frac{\Pr(Y_i=1)}{1-\Pr(Y_i=1)}\right)=\beta_0+\beta_1 X_i \tag{2.66}$$

where the term on the left side of the equation is the log odds or logit. The logistic model also has a fixed variance, which I omit from the expression.

Both modeling frameworks provide a conversion to the probability metric, albeit using different functions. The predicted probability of endorsing the highest category (e.g., the probability of quitting) from the probit model is

$$\Pr(Y_i=1\,|\,\boldsymbol{\beta},\text{data})=1-\Phi\left(\frac{\tau-\mathbf{X}_i\boldsymbol{\beta}}{\sigma_\varepsilon^2}\right)=1-\Phi(-\mathbf{X}_i\boldsymbol{\beta})=\Phi(-\mathbf{X}_i\boldsymbol{\beta})=\pi_i \tag{2.67}$$

where $\mathbf{X}_i$ is the predictor vector for individual i (including a column of 1's for the intercept), $\boldsymbol{\beta}$ contains the coefficients, $\mathbf{X}_i\boldsymbol{\beta}$ is the predicted latent response, and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal curve. The subtraction inside the parentheses expresses the threshold as a z-score (recall that $\tau = 0$ and $\sigma_\varepsilon^2 = 1$), and

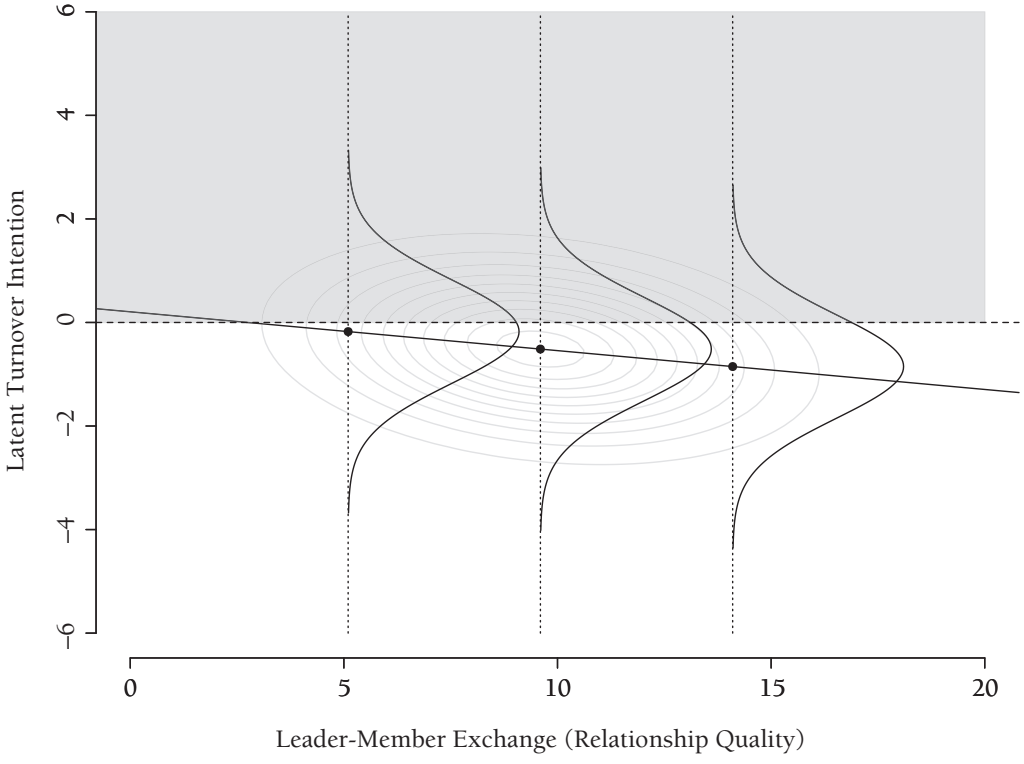

FIGURE 2.18. Latent response distribution for a binary variable. The vertical line at 0 is a threshold parameter τ that divides the latent distribution into two regions. Employees with no quitting intentions have latent scores below the threshold, and employees who intend to quit have scores above the threshold. The area under the shaded region of the curve is the probability of quitting (the proportion of 1's in the data).

the function returns the area *below* this value in a standard normal curve. Subtracting that result from 1 gives the area above the threshold (e.g., the shaded regions of the normal curves in Figure 2.18). Similarly, the logit link function translates predicted latent response scores to the probability metric as follows:

$$\Pr(Y_i = 1 \mid \boldsymbol{\beta}, \text{data}) = \frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i\boldsymbol{\beta})} = \pi_i \tag{2.68}$$

Probability Distribution and Log-Likelihood

Probit regression is appealing, because it leverages a normal distribution for the underlying response variable. In later chapters, I adopt a likelihood expression that features the latent response scores in the normal curve expression from Equation 2.3, but for now, I use an alternative equation that represents an individual's likelihood contribution as a predicted probability (area under the standard normal distribution).

$$L_i(\boldsymbol{\beta} \mid \text{data}) = \Phi(-\mathbf{X}_i\boldsymbol{\beta})^{Y_i} \times \left(1 - \Phi(-\mathbf{X}_i\boldsymbol{\beta})\right)^{1-Y_i} = \pi_i^{Y_i}(1-\pi_i)^{(1-Y_i)} \tag{2.69}$$

In the context of the employee turnover example, the likelihood features the product of the predicted probability of quitting (left term) and not quitting (right term). The scores in the exponents act like on–off switches that activate the left term (the predicted probability that $Y = 1$) if $Y = 1$ and trigger the right term (the predicted probability that $Y = 0$) if $Y = 0$. Taking the natural logarithm and summing across the N cases gives the following sample log-likelihood expression:

$$LL(\boldsymbol{\beta} \mid \text{data}) = \sum_{i=1}^{N}\left(Y_i \times \ln\left(\Phi(-\mathbf{X}_i\boldsymbol{\beta})\right) + (1 - Y_i) \times \ln\left(1 - \Phi(-\mathbf{X}_i\boldsymbol{\beta})\right)\right) \tag{2.70}$$

Numerically, the log-likelihood is a large negative number that equals the sum of logarithmically transformed probability values. Conceptually, this value represents the data's support for a particular combination of population regression coefficients in $\boldsymbol{\beta}$.

The log-likelihood for logistic regression has the same form as Equation 2.70 but uses the Bernoulli distribution probability distribution from Equation 2.1. Reversing the probability distribution's arguments by taking data values as given and varying the parameters gives the likelihood expression for a single observation.

$$L_i(\boldsymbol{\beta} \mid \text{data}) = \left(\frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i\boldsymbol{\beta})}\right)^{Y_i} \times \left(1 - \frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i\boldsymbol{\beta})}\right)^{1-Y_i} = \pi_i^{Y_i}(1-\pi_i)^{(1-Y_i)} \tag{2.71}$$

Consistent with Equation 2.70, the likelihood features the product of the predicted probability of quitting (left term) and not quitting (right term), and the scores in the exponent activate the probability that corresponds to one's binary response. Taking the natural logarithm and summing across the N cases gives the following sample log-likelihood expression, which again represents the data's support for a particular combination of regression parameters:

$$LL(\boldsymbol{\beta} \mid \text{data}) = \sum_{i=1}^{N} \left(Y_i \times \ln\left(\frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i\boldsymbol{\beta})} \right) + (1-Y_i) \times \ln\left(\frac{1}{1+\exp(\mathbf{X}_i\boldsymbol{\beta})} \right) \right) \quad (2.72)$$

Unlike the other models in this chapter, there is no analytic solution for the probit and logistic regression coefficients, and iterative optimizers such as Newton's algorithm are a must. Iterative optimization works the same as it did with normally distributed data, so I point readers to the literature for additional technical details (Agresti, 2012; Greene, 2017). Putting aside the technicalities, the process of computing standard errors follows the same procedure described earlier in the chapter; manipulating the matrix of second derivatives that quantifies the curvature of the log-likelihood function gives the variance–covariance matrix of the estimates, the diagonal of which contains squared standard errors. Similarly, the significance testing options described in Section 2.11 are no different with categorical variable models.

Analysis Example

Expanding on the employee turnover example, I used maximum likelihood estimation to fit probit and logistic regression models that use leader–member exchange, employee empowerment, and a male dummy code (0 = *female,* 1 = *male*) to predict a binary measure of turnover intention (*TURNOVER* = 0 if an employee has no plan to leave his or her position, and *TURNOVER* = 1 if the employee has intentions of quitting).

$$TURNOVER_i^* = \beta_0 + \beta_1(LMX_i) + \beta_2(EMPOWER_i) + \beta_3(MALE_i) + \varepsilon_i \quad (2.73)$$

$$\ln\left(\frac{\Pr(TURNOVER_i = 1)}{1-\Pr(TURNOVER_i = 1)} \right) = \beta_0 + \beta_1(LMX_i) + \beta_2(EMPOWER_i) + \beta_3(MALE_i)$$

The probit model's residual variance is fixed at 1 for identification, and the model additionally incorporates a fixed threshold parameter that divides the latent response variable distribution into two segments. The logistic regression can also be viewed as a latent response model, but it is typical to write the equation without a residual. Note that I use β's to represent focal model parameters, but the estimated coefficients will not be the same (logit coefficients are approximately 1.7 times larger than probit coefficients; Birnbaum, 1968). As always, analysis scripts are available on the companion website.

Table 2.7 shows the maximum likelihood analysis results for both models. Starting with the probit regression results, the Wald test of the full model was statistically significant, $T_W(3) = 20.00$, $p < .001$, meaning that the estimates are at odds with the null hypothesis that all three population slopes equal zero. Each slope coefficient reflects the expected *z*-score change in the latent response variable for a one unit increase in the predictor, controlling for other regressors. For example, the leader–member exchange coefficient indicates that a one-unit increase in relationship quality is expected to *decrease* the latent proclivity to quit by 0.06 *z*-score units ($\hat{\beta}_1 = -0.06$, $SE = .02$), holding other predictors constant.

Turning to the logistic regression results, the Wald test of the full model was again statistically significant, and the test statistic's numerical value was comparable to that

TABLE 2.7. Probit and Logistic Regression Estimates

Parameter	Est.	*RSE*	*z*	*p*	OR
		Probit regression			
β_0	0.80	0.35	2.25	.03	—
β_1 (*LMX*)	–0.06	0.02	–2.99	.00	—
β_2 (*EMPOWER*)	–0.03	0.01	–1.83	.07	—
β_3 (*MALE*)	–0.03	0.11	–0.30	.77	—
R^2	.06	.03	2.36	.02	—
		Logistic regression			
β_0	1.37	0.60	2.30	.02	—
β_1 (*LMX*)	–0.10	0.04	–2.96	.00	0.90
β_2 (*EMPOWER*)	–0.04	0.02	–1.81	.07	0.96
β_3 (*MALE*)	–0.06	0.18	–0.31	.75	0.95
R^2	.05	.02	2.30	.02	—

Note. RSE, robust standard error; OR, odds ratio.

of the probit model, $T_W(3) = 19.35$, $p < .001$. Each slope coefficient now reflects the expected change in the log odds of quitting for a one-unit increase in the predictor, holding all other covariates constant. For example, the leader–member exchange slope indicates that a one-unit increase in relationship quality decreases the log odds of quitting by .10 ($\hat{\beta}_1 = -0.10$, $SE = .04$), controlling for employee empowerment and gender. Notice that the logistic coefficients are approximately 1.7 times larger than the probit slopes, as expected (Birnbaum, 1968). Exponentiating each slope gives an odds ratio that reflects the multiplicative change in the odds (the probability ratio on the left side of Equation 2.66) for a one-unit increase in a predictor (e.g., a one-point increase on the leader–member exchange scale multiplies the odds of quitting by 0.90).

The analysis results highlight that probit and logistic models are effectively equivalent and almost always lead to the same conclusions. Some researchers favor the logistic framework, because it yields odds ratios, but there is otherwise little reason to prefer one approach to the other. As you will see, probit regression plays a more central role with Bayesian estimation and multiple imputation.

2.14 SUMMARY AND RECOMMENDED READINGS

Maximum likelihood is the go-to estimator for many common statistical models, and it is one of the three major pillars of this book. As its name implies, the estimator identifies the population parameters that are most likely responsible for a particular sample of data. Much of this chapter has unpacked this definition in the context of linear regression models and multivariate analyses based on the normal distribution, and the last section has outlined logistic and probit models for categorical outcomes. Having estab-

lished all the major details behind estimation and inference, Chapter 3 applies maximum likelihood to missing data problems. As you will see, everything from this chapter carries over to missing data applications, where the goal remains to identify parameter values that maximize fit to the data—the only difference is that some participants have more of it than others. Finally, I recommend the following articles for readers who want additional details on topics from this chapter:

Buse, A. (1982). The likelihood ratio, Wald, and Lagrange multiplier tests: An expository note. *American Statistician, 36,* 153–157.

Eliason, S. R. (1993). *Maximum likelihood estimation: Logic and practice.* Newbury Park, CA: Sage.

Greene, W. H. (2017). *Econometric analysis* (8th ed.). Boston: Prentice Hall.

Savalei, V. (2014). Understanding robust corrections in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 21,* 149–160.