3

An Introduction to Maximum Likelihood Estimation

3.1 CHAPTER OVERVIEW

Many modern statistical procedures in widespread use today rely on maximum likelihood estimation. Maximum likelihood also plays a central role in missing data analyses and is one of two approaches that methodologists currently regard as state of the art (Schafer & Graham, 2002). This chapter introduces the mechanics of maximum likelihood estimation in the context of a complete-data analysis. Although the basic estimation process is largely the same with missing data, understanding the basic estimation principles is made easier without this additional complication.

The starting point for a maximum likelihood analysis is to specify a distribution for the population data. Researchers in the social and the behavioral sciences routinely assume that their variables are normally distributed in the population, so I describe maximum likelihood in the context of multivariate normal data. The normal distribution provides a familiar platform for illustrating estimation principles, but it also offers the basis for the missing data handling procedure that I outline in Chapters 4 and 5. Although the normal distribution plays an integral role throughout the entire estimation process, the basic mechanics of estimation are largely the same with other population distributions. For example, Chapter 6 describes a maximum likelihood analysis that uses the binomial distribution for a binary outcome, and many of the key ideas from this chapter resurface in that example.

3.2 THE UNIVARIATE NORMAL DISTRIBUTION

Most applications of maximum likelihood estimation rely on the multivariate normal distribution. However, a univariate example is a useful starting point for illustrating basic estimation principles. As you will see, the estimation process is largely the same with multivariate data. The mathematical machinery behind maximum likelihood relies heavily on a **probability**

density function that describes the distribution of the population data. The density function for the univariate normal distribution is

$$L_i = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-.5(y_i-\mu)^2}{\sigma^2}} \tag{3.1}$$

where y_i is a score value, μ is the population mean, σ^2 is the population variance, and L_i is a likelihood value that describes the height of the normal curve at a particular score value. In words, the density function describes the relative probability of obtaining a score value from a normally distributed population with a particular mean and variance. Although the density function is complex, the driving force behind the equation is simply a squared z score, $(y_i-\mu)^2/\sigma^2$. This **Mahalanobis distance** term quantifies the standardized distance between a score and the mean and largely determines the result of the equation. Density functions typically contain a collection of scaling terms that make the area under the distribution sum (i.e., integrate) to one, and the portion of the equation to the left of the exponent symbol serves this purpose for the normal curve. These terms are not vital for understanding the estimation process.

To illustrate the probability density function, consider the IQ scores in Table 3.1. I designed this small data set to mimic an employee selection scenario in which prospective employees complete an IQ test during their interview and a supervisor subsequently rates their job performance following a 6-month probationary period. Ultimately, maximum likelihood uses the density function in Equation 3.1 to estimate the population parameters, but

TABLE 3.1. IQ and Job Performance Data

IQ	Job performance
78	9
84	13
84	10
85	8
87	7
91	7
92	9
94	9
94	11
96	7
99	7
105	10
105	11
106	15
108	10
112	10
113	12
115	14
118	16
134	12

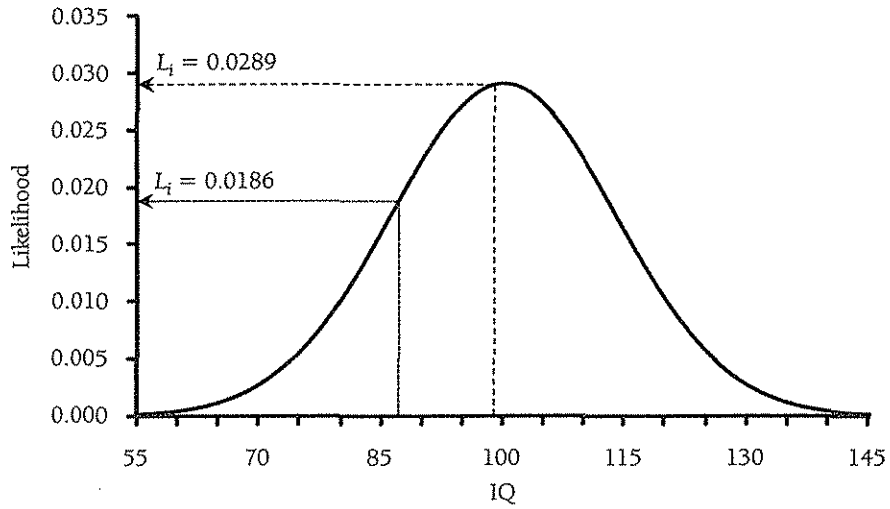

FIGURE 3.1. Univariate normal distribution with $\mu = 100$ and $\sigma^2 = 189.60$. The likelihood values represent the height of the distribution at score values of 99 and 87.

understanding the basic estimation principles is easier when the parameter values are known. Consequently, I temporarily assume that the population mean is $\mu = 100$ and the population variance is $\sigma^2 = 189.60$.

The density function in Equation 3.1 describes the relative probability of obtaining a score value from a normally distributed population with a particular mean and variance. For example, consider two IQ scores, 99 and 87. Substituting $y_i = 99$, $\mu = 100$, and $\sigma^2 = 189.60$ into the density function yields a **likelihood** value of $L_i = .0289$. Similarly, substituting an IQ score of 87 into Equation 3.1 returns a likelihood of $L_i = .0186$. Although they resemble probabilities, it is more accurate to think of a likelihood value as the relative probability of drawing a particular IQ score from a normal distribution with a mean of 100 and a variance of 189.60. Consequently, it is incorrect to say that an IQ score of 99 has a probability of .0289, but it is true that an IQ score of 99 is more probable than a score of 87. (With a continuous score distribution, there are an infinite number of y_i values, so the probability of any single score is effectively zero.) Visually, the likelihood represents the height of the normal curve at a particular score value. To illustrate, Figure 3.1 presents a graphical depiction of the previous likelihood values. Notice that the elevation of the normal curve is higher at an IQ score of 99, which is consistent with the relative magnitude of the two likelihood values.

It is also useful to view the likelihood as a measure of "fit" between a score and the population parameters. In Figure 3.1, the largest possible likelihood value (i.e., the highest point on the distribution) corresponds to the score that is exactly equal to the population mean, and the likelihood values decrease in magnitude as the distance from the mean increases. Returning to Equation 3.1, this implies that smaller Mahalanobis distance values (i.e., smaller squared z scores) produce larger likelihood values, whereas larger Mahalanobis distance values yield smaller likelihoods. Consequently, a score that yields a high likelihood value also has a good fit because it falls close to the population mean. As you will see, interpreting the likelihood as a measure of fit becomes useful when the population parameters are unknown.

3.3 THE SAMPLE LIKELIHOOD

The goal of maximum likelihood estimation is to identify the population parameter values that have the highest probability of producing a particular *sample* of data. Identifying the most likely parameter values requires a summary fit measure for the entire sample, not just a single score. In probability theory, the joint probability for a set of independent events is the product of individual probabilities. For example, the probability of flipping a fair coin twice and getting two heads is $.50 \times .50 = .25$. Although they are not exactly probabilities, the same rule applies to likelihood values. Consequently, the likelihood for a sample of cases is the product of N individual likelihood values.

More formally, the sample likelihood is

$$L = \prod_{i=1}^{N} \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-.5(y_i-\mu)^2}{\sigma^2}} \right\} \tag{3.2}$$

where the braces contain the likelihood of a single score (i.e., Equation 3.1), and $\prod$ is the multiplication operator. In words, Equation 3.2 says to compute the likelihood for each member of a sample and multiply the resulting values. For example, Table 3.2 shows the likelihood values for the IQ scores in Table 3.1. Multiplying the 20 values gives the likelihood for the entire sample, $L = 7.89E{-}36$ (in scientific notation, E –36 means to move the decimal to the left by 36 places). The sample likelihood quantifies the joint probability of drawing this

TABLE 3.2. Individual Likelihood and Log-Likelihood Values

IQ	L_i	$\log L_i$
78	.008	–4.818
84	.015	–4.217
84	.015	–4.217
85	.016	–4.135
87	.019	–3.987
91	.023	–3.755
92	.024	–3.710
94	.026	–3.636
94	.026	–3.636
96	.028	–3.584
99	.029	–3.544
105	.027	–3.607
105	.027	–3.607
106	.026	–3.636
108	.024	–3.710
112	.020	–3.921
113	.019	–3.987
115	.016	–4.135
118	.012	–4.396
134	.001	–6.590

collection of 20 scores from a normal distribution with a mean of 100 and a variance of 189.60.

Because a number of factors influence the value of the sample likelihood (e.g., the sample size, the number of variables), there is no cutoff that determines good or bad fit. Consistent with the interpretation of the individual likelihood values, it is best to view the sample likelihood as a measure of relative fit. Ultimately, the likelihood (or more accurately, the log-likelihood) will provide a basis for choosing among a set of plausible population parameter values.

3.4 THE LOG-LIKELIHOOD

Because the sample likelihood is such a small number, it is difficult to work with and is prone to rounding error. Computing the natural logarithm of the individual likelihood values solves this problem and converts the likelihood to a more tractable metric. To illustrate, the rightmost column of Table 3.2 shows the **log-likelihood** value for each IQ score. Taking the natural logarithm of a number between zero and one yields a negative number, but the log-likelihood values serve the same role and have the same meaning as the individual likelihoods. For example, reconsider the IQ scores of 99 and 87, the likelihood values for which are .0289 and .0186, respectively. The corresponding log-likelihood values are –3.544 versus –3.987, respectively. Again, the IQ score of 99 has a higher likelihood than a score of 87 because it is closer to the mean. An IQ score of 99 also has a higher (i.e., "less negative") log-likelihood value than a score of 87. The log-likelihood values still quantify relative probability, but they simply do so using a different metric. Consequently, values that are closer to zero reflect a higher relative probability and a closer proximity to the population mean.

Working with logarithms simplifies the computation of the sample log-likelihood. One of the basic logarithm rules states that log(AB) is equal to log(A) + log(B). Consequently, the **sample log-likelihood** is the *sum* of the individual log-likelihood values, as follows:

$$\log L = \sum_{i=1}^{N} \log\left\{\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{.5(y_i-\mu)^2}{\sigma^2}}\right\} \tag{3.3}$$

Returning to the data in Table 3.2, note that summing the log-likelihood values yields log$L$ = –80.828. Consistent with the sample likelihood, the sample log-likelihood is a summary measure that quantifies the joint probability of drawing the sample of 20 scores from a normal distribution with a mean of 100 and a variance of 189.60.

3.5 ESTIMATING UNKNOWN PARAMETERS

Thus far, I have assumed that the population parameters (i.e., μ and σ^2) are known. These parameters typically need to be estimated from the data. Fortunately, switching to a situation

where the parameter values are unknown does change the previous computations. Conceptually, the estimation procedure is an iterative process that repeatedly "auditions" different values for μ and σ^2 until it finds the estimates that are most likely to have produced the data. It does this by repeating the log-likelihood computations many times, each time with different values of the population parameters. The sample log-likelihood gauges the relative fit of the prospective estimates and provides a basis for choosing among a set of plausible parameter values. The ultimate goal of estimation is to identify the unique combination of estimates that maximize the log-likelihood and thus produce the best fit to the data (i.e., the estimates that minimize the standardized distances between the scores and the mean).

To illustrate the estimation process, reconsider the IQ data in Table 3.1. Suppose that the company wants to use maximum likelihood to estimate the IQ mean. One way to identify the most likely value of the population mean is to substitute different values of μ into Equation 3.3 and compute the sample log-likelihood for each estimate. Table 3.3 gives the log-likelihood values for five different estimates of the population mean. (Substituting any nonzero value of the variance into Equation 3.3 leads to the same estimate of the mean, so I continue to fix σ^2 at 189.60.) To begin, notice that each mean estimate yields a different set of individual log-likelihood values. For example, when $\mu = 98$, an IQ score of 96 is close to the mean and has a higher log-likelihood (i.e., better fit) than a score of 105. In contrast,

TABLE 3.3. Individual and Sample Log-Likelihood Values for Five Different Estimates of the Population Mean

	Population mean				
IQ	$\mu = 98$	$\mu = 99$	$\mu = 100$	$\mu = 101$	$\mu = 102$
78	–4.596	–4.704	–4.818	–4.936	–5.060
84	–4.058	–4.135	–4.217	–4.304	–4.396
84	–4.058	–4.135	–4.217	–4.304	–4.396
85	–3.987	–4.058	–4.135	–4.217	–4.304
87	–3.860	–3.921	–3.987	–4.058	–4.135
91	–3.671	–3.710	–3.755	–3.805	–3.860
92	–3.636	–3.671	–3.710	–3.755	–3.805
94	–3.584	–3.607	–3.636	–3.671	–3.710
94	–3.584	–3.607	–3.636	–3.671	–3.710
96	–3.552	–3.565	–3.584	–3.607	–3.636
99	–3.544	–3.541	–3.544	–3.552	–3.565
105	–3.671	–3.636	–3.607	–3.584	–3.565
105	–3.671	–3.636	–3.607	–3.584	–3.565
106	–3.710	–3.671	–3.636	–3.607	–3.584
108	–3.805	–3.755	–3.710	–3.671	–3.636
112	–4.058	–3.987	–3.921	–3.860	–3.805
113	–4.135	–4.058	–3.987	–3.921	–3.860
115	–4.304	–4.217	–4.135	–4.058	–3.987
118	–4.596	–4.493	–4.396	–4.304	–4.217
134	–6.959	–6.772	–6.590	–6.413	–6.242
logL =	–81.039	–80.881	–80.828	–80.881	–81.039

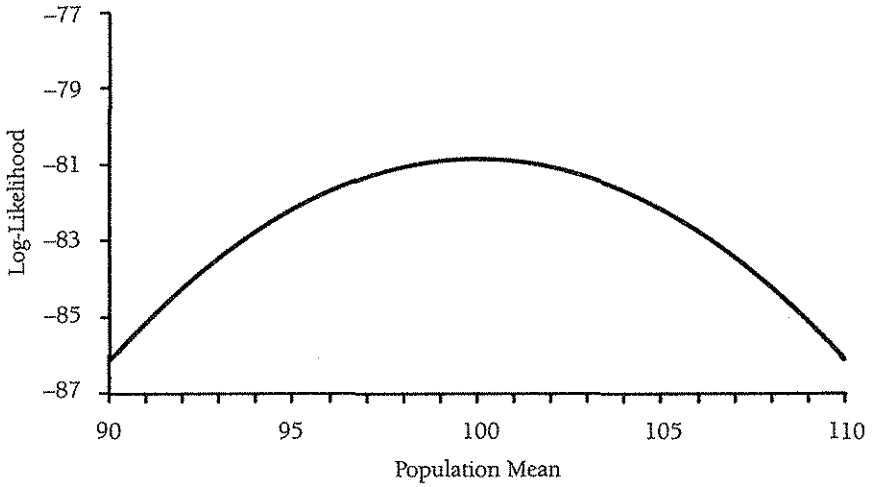

FIGURE 3.2. The log-likelihood function for the mean. The figure shows how the sample log-likelihood values vary across a range of plausible values for the population mean. The maximum of the function occurs at $\mu = 100$.

substituting a value of $\mu = 102$ into the equation reverses the relative fit of these two data points because the IQ score of 105 is closer to the mean. The sample log-likelihood is the sum of the individual log-likelihood values, so changing the population mean affects its value as well. Comparing the relative fit of the five mean estimates, $\mu = 100$ yields the highest log-likelihood and thus provides the best fit to the data.

The sample log-likelihood values in the bottom row of Table 3.3 suggest that $\mu = 100$ is the best estimate of the mean, but thus far I have only considered five possible values. I conducted a more comprehensive search by computing the sample log-likelihood for mean values between 90 and 110. Figure 3.2 is a **log-likelihood function** that plots the resulting log-likelihood values against the corresponding estimates of the mean on the horizontal axis. The log-likelihood function resembles a hill, with the most likely parameter value located at its peak. Conceptually, the estimation process is akin to hiking to the top of the hill. Consistent with Table 3.3, the peak of the log-likelihood function is located at $\mu = 100$, and the sample log-likelihood values decrease as μ gets farther away from 100 in either direction. After thoroughly auditioning a range of plausible parameter values, the data provide the most evidence in support of $\mu = 100$. Consequently, $\hat{\mu} = 100$ is the **maximum likelihood estimate** of the mean, or the population parameter with the highest probability of producing this sample of IQ scores.

Next, I applied the same iterative search procedure to the population variance. Specifically, I fixed the value of μ at 100 in Equation 3.3 and computed the sample log-likelihood for variance values between 50 and 450. Figure 3.3 shows a log-likelihood function that plots the resulting log-likelihood values against the corresponding estimates of σ^2 on the horizontal axis. The log-likelihood function of the variance looks very different from that of the mean, but it works in exactly the same way. Although it is difficult to determine graphically, the peak of the log-likelihood function is located at $\sigma^2 = 189.60$. Consequently, $\hat{\sigma}^2 = 189.60$ is the maximum likelihood estimate of the variance (i.e., the population variance that has the highest probability of producing the sample of IQ scores in Table 3.1).

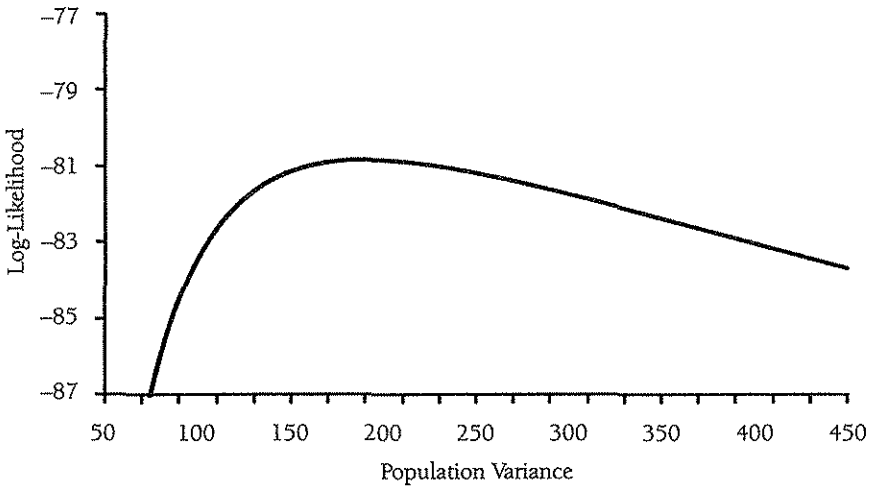

FIGURE 3.3. The log-likelihood function for the variance. The figure shows how the sample log-likelihood values vary across a range of plausible values for the population variance. The maximum of the function occurs at $\sigma^2 = 189.60$.

3.6 THE ROLE OF FIRST DERIVATIVES

The random search process in the previous examples would become exceedingly tedious in most real-world estimation problems. In practice, software packages use calculus derivatives to identify the maximum of the log-likelihood function (i.e., the peak of the hill). Returning to Figure 3.2, the **first derivative** is the slope of the log-likelihood function at a particular value of the population mean (or more accurately, the slope of a line that is tangent to a certain point on the function). To illustrate, imagine using a magnifying glass to zoom in on a very small section of the log-likelihood function located directly above $\mu = 95$. Although the entire function has substantial curvature, the log-likelihood would begin to resemble a positively sloping straight line as the magnifying glass comes into sharper focus. The slope of this minute section of the log-likelihood function is the first derivative (or more accurately, the first derivative of the log-likelihood function with respect to the mean). Now imagine focusing the magnifying glass on the highest point of the log-likelihood function, directly above $\mu = 100$. Again, with a sharp enough focus, the log-likelihood would appear as a straight line, this time with a slope of zero. Figure 3.4 shows a tangent line at the maximum of the log-likelihood function. The slope of this line is the first derivative.

Obtaining the first derivatives of the log-likelihood equation is tedious and involves a process known as differentiation. Illustrating the mechanics of differential calculus is beyond the scope of this chapter, but most introductory calculus texts contain the differentiation rules. The important point is that first derivatives are equations that give the slope of each parameter's log-likelihood at any given point along the function. More importantly, Figure 3.4 suggests that substituting the maximum likelihood estimate into the derivative equation returns a slope of zero. This implies a relatively straightforward strategy: set the result of the derivative formula to zero and solve for the unknown parameter value.

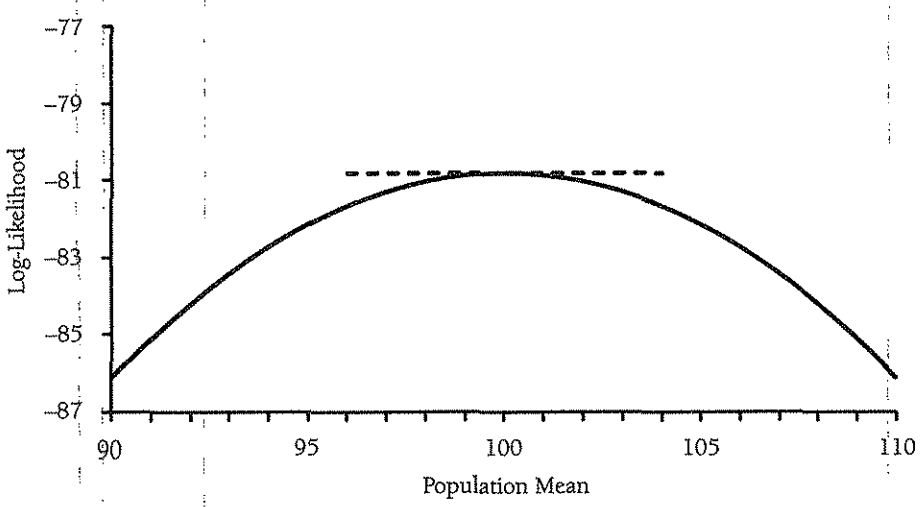

FIGURE 3.4. The log-likelihood function with a tangent line imposed at its maximum. The slope of this line is the first derivative of the log-likelihood function at $\mu = 100$.

To illustrate how derivatives simplify the estimation process, I used differential calculus to obtain the first derivative of the log-likelihood function with respect to μ. The first derivative equation for the population mean is as follows:

$$\frac{\partial \log L}{\partial \mu} = \frac{1}{\sigma^2}\left(-N\mu + \sum_{i=1}^{N} y_i\right) \tag{3.4}$$

In words, the terms to the left of the equal sign read "the first derivative of the log-likelihood function with respect to the population mean" (the ∂ symbol denotes a derivative), and the equation to the right of the equal sign defines the slope of the log-likelihood function at a particular value of μ. Substituting the maximum likelihood estimate of the mean into the equation returns a slope of zero, so the first step is to set the slope equation equal zero. Next, multiplying both sides of the resulting equation by σ^2 eliminates the variance from the formula and leaves the collection of terms in parentheses equal to zero. Finally, using algebra to solve for μ gives the maximum likelihood estimate of the mean.

$$\hat{\mu} = \sum_{i=1}^{N} y_i / N \tag{3.5}$$

Notice that Equation 3.5 is the usual formula for the sample mean.

The same differentiation process applies to the population variance. Applying differential calculus rules to the log-likelihood equation gives the derivative equation for the variance.

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \sum_{i=1}^{N} (y_i - \mu)^2 / 2\sigma^4 \tag{3.6}$$

Setting the right side of the equal to zero and solving for σ^2 gives the maximum likelihood estimate of the variance, as follows:

$$\hat{\sigma}^2 = \sum_{i=1}^{N} (y_i - \mu)^2 / N \tag{3.7}$$

Notice that Equation 3.7 has N rather than $N - 1$ in the denominator, so it is identical to the usual formula for the population variance. The use of N in the denominator of the variance formula implies that maximum likelihood estimation yields negatively biased estimates of variances (and covariances). This is a well-known property of maximum likelihood that extends to more complex analyses (e.g., structural equation models, multilevel models). However, this bias is only a concern in small samples because it quickly becomes negligible as the sample size increases.

The previous examples are straightforward because familiar equations define the maximum likelihood estimates. This is true in a limited number of situations (e.g., means, variances, covariances, regression coefficients), but more complex applications of maximum likelihood estimation (e.g., structural equation models, multilevel models, missing data estimation) generally require iterative optimization algorithms to identify the most likely set of parameter values. The expectation maximization (EM) algorithm is one such method that I discuss in the next chapter. Nevertheless, estimating the mean and the variance is a useful exercise because it provides a familiar platform from which to explore maximum likelihood.

3.7 ESTIMATING STANDARD ERRORS

The primary goal of a statistical analysis is to estimate a set of unknown model parameters, but obtaining standard errors for the resulting point estimates is an important secondary goal. The log-likelihood function provides a mechanism for estimating standard errors, and this too relies heavily on calculus derivatives. To illustrate, Figure 3.5 shows the log-likelihood functions for two data sets, both of which have a mean of 100. I used the data in Table 3.1 to generate the top function, and the bottom function corresponds to a set of IQ scores with a variance that is exactly two and a half times larger than that of the data in Table 3.1.

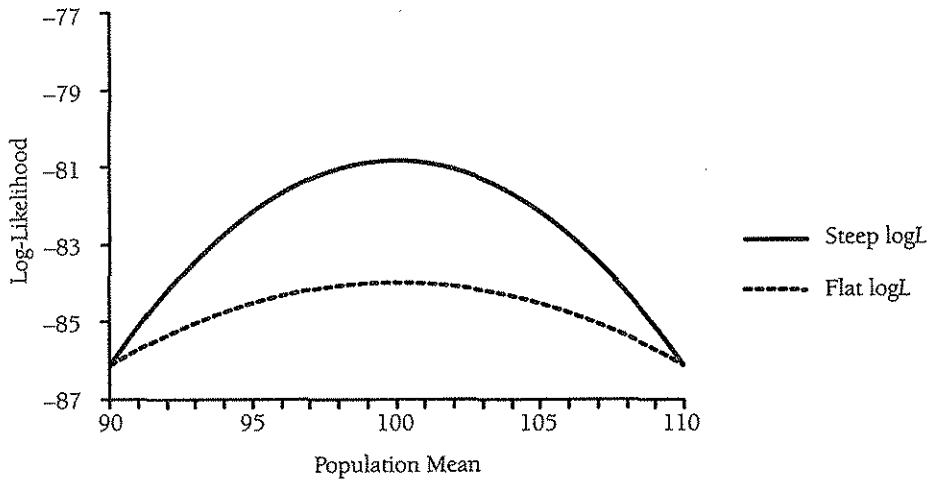

FIGURE 3.5. Two log-likelihood functions for the mean. The steep function is from a sample of 20 IQ scores with $\mu = 100$ and $\sigma^2 = 189.60$, and the flat function corresponds to a data set with $\mu = 100$ and $\sigma^2 = 474.00$. The two functions produce the same estimate of the mean (i.e., the maxima are located at $\mu = 100$), but they have very different curvatures. The steep function has a larger second derivative (i.e., is more peaked) and a smaller standard error.

Throughout this section, I refer to the top function as the "steep" log-likelihood and to the bottom function as the "flat" log-likelihood. Although the two log-likelihood functions produce the same estimate of the mean (i.e., the maxima are located at $\mu = 100$), they have a very different curvature. As you will see, the magnitude of this curvature largely determines the maximum likelihood standard error.

At an intuitive level, the curvature of the log-likelihood function provides important information about the uncertainty of an estimate. A flat function makes it difficult to discriminate among competing estimates because the log-likelihood values are relatively similar across a range of plausible parameter estimates. In contrast, a steep log-likelihood function more clearly differentiates the maximum likelihood estimate from other possible parameter values. To illustrate, consider the log-likelihood functions in Figure 3.5. The flat function yields log-likelihood values of –84.518 and –83.991 at $\mu = 95$ and $\mu = 100$, respectively, which is a difference of 0.527. In contrast, the corresponding log-likelihood values for the steep function are –82.147 and –80.828, which is a difference of 1.319. Both functions yield the same estimate of the population mean, but the log-likelihood values from the steep function decrease more rapidly as μ gets farther away from 100. Consequently, the steep function better differentiates $\mu = 100$ from other plausible parameter estimates. Not coincidentally, the steep function decreases at a rate that is two and a half times larger than that of the flat function (i.e., 1.319 / 0.527 = 2.5). Recall that this is the same factor by which the variances differed.

The Role of Second Derivatives

Mathematically, the **second derivative** quantifies the curvature of the log-likelihood function. Technically, a second derivative measures the rate at which the first derivatives change across a function. For example, a steep log-likelihood function has rapidly changing first derivatives (i.e., slopes), so its second derivative is large. In contrast, a flat log-likelihood function has a small second derivative because its first derivatives change slowly (i.e., the slopes are relatively flat across the entire range of the function). To make this idea more concrete, Table 3.4 shows the first derivatives of the two functions in Figure 3.5 (I obtained the derivatives by substitut-

TABLE 3.4. First Derivative Values for the Steep and Flat Log-Likelihood Functions

μ	Steep function	Flat Function
95	0.527	0.211
96	0.422	0.169
97	0.316	0.127
98	0.211	0.084
99	0.105	0.042
100	0.000	0.000
101	–0.105	–0.042
102	–0.211	–0.084
103	–0.316	–0.127
104	–0.422	–0.169
105	–0.527	–0.211

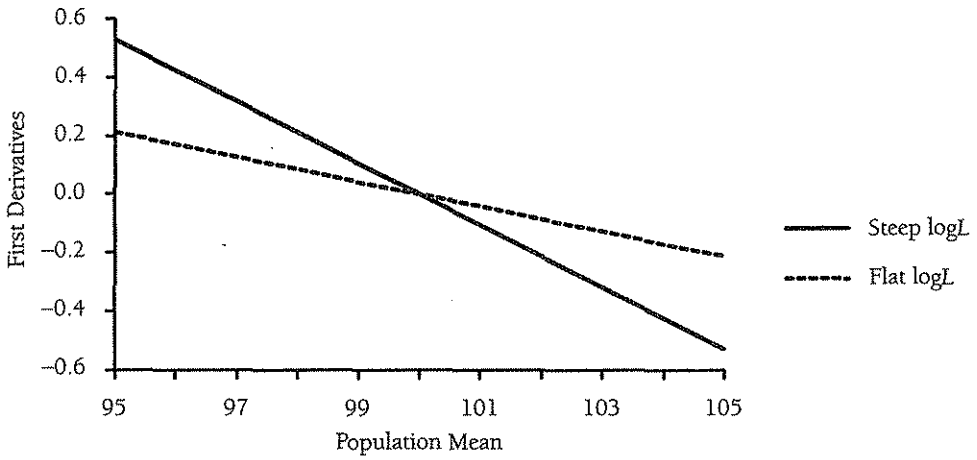

FIGURE 3.6. First from two log-likelihood functions. The solid line plots the first derivatives of the steep log-likelihood function in Figure 3.4, and the dashed line depicts the first derivatives for the flat log-likelihood function in Figure 3.4.

ing the appropriate values of μ and σ^2 into Equation 3.4). Beginning at $\mu = 95$, the first derivatives are positive (i.e., the slope of the log-likelihood function is positive) and decrease in magnitude until $\mu = 100$, after which they become increasingly negative (i.e., the log-likelihood has a negative slope when μ is greater than 100). This trend is true for both functions, but the steep function's derivatives change at a faster rate. Figure 3.6 shows these first derivatives plotted against the values of the population mean on the horizontal axis. The two lines depict the rate of change in the first derivatives, and the slopes of these lines are the second derivatives. Again, the values of the second derivatives determine the magnitude of standard errors, such that larger second derivatives (i.e., more peaked functions) translate into smaller standard errors.

An Example: The Standard Error of the Mean

Having established some important background information, I now show how the second derivatives translate into standard errors. Computing a maximum likelihood standard involves four steps: (1) calculate the value of the second derivative, (2) multiply the second derivative by negative 1, (3) compute the inverse (i.e., reciprocal) of the previous product, and (4) take the square root of the resulting inverse. To keep things simple, I outline the computational steps for the standard error of the mean, but the process is identical for other parameters.

The first step of the standard error computations requires the second derivative equations. Applying differential calculus rules to the first derivative equations (e.g., Equation 3.4) produces the necessary formulas. As an example, differentiating Equation 3.4 yields the second derivative equation for the mean, which is simply $-N^2/\sigma$. As I explain later, the second derivative should always be a negative number, which it is in this case. The next step is to multiply the second derivative by negative 1. This operation yields a quantity called **information**. Information quantifies the curvature of the log-likelihood function, such that steeper functions produce larger (i.e., more positive) information values. The third step is to compute the inverse (i.e., the reciprocal) of the information. Taking the inverse of the information

gives the **sampling variance** (i.e., squared standard error) of the mean. Equation 3.8 summarizes the first three steps

$$\text{var}(\hat{\mu}) = -\left[\frac{\partial^2 \text{Log}L}{\partial^2 \mu}\right]^{-1} = -\left[\frac{-N}{\sigma^2}\right]^{-1} = \frac{\sigma^2}{N} \tag{3.8}$$

where var($\hat{\mu}$) denotes the sampling variance, and ∂^2 symbolizes a second derivative. The rightmost term in Equation 3.8 may look familiar because it is the square of the standard error of the mean. Researchers typically report sampling error on the standard deviation metric rather than the variance metric, so the final step is to take the square root of the sampling variance. Doing so yields $\sigma/\sqrt{N}$, which is the usual formula for the standard error of the mean.

To illustrate the computation of maximum likelihood standard errors, reconsider the log-likelihood functions in Figure 3.5. The steep function is from a sample of 20 IQ scores with $\mu = 100$ and $\sigma^2 = 189.60$, and the flat function corresponds to a data set with $\mu = 100$ and $\sigma^2 = 474.00$. Substituting the sample size and the variance into the second derivative formula yields derivative values of –0.105 and –0.042 for the steep and flat functions, respectively. Visually, these values are the slopes of the two lines in Figure 3.6. Multiplying the second derivative values by negative 1 gives the information. Again, peaked log-likelihood functions produce larger information values, so the relative magnitude of the two information values (0.105 versus 0.042) reflects the fact that the two functions have different curvature. Computing the inverse of the information yields the sampling variance of the mean (i.e., squared standard error), the values of which are 9.48 and 23.70 for the steep and flat functions, respectively. Notice that the sampling variances differ by a ratio of 2.5, which is the same factor that differentiates the second derivatives and the score variances. Finally, taking the square root of the sampling variance yields the standard error. Not surprisingly, the steep function has a smaller standard error than the flat function (3.08 versus 4.87, respectively), owing to the fact that its second derivative value is larger in absolute value.

Why Is the Second Derivative Value Negative?

It may not be immediately obvious, but the fact that the second derivative takes on a negative value is important. To understand why this is the case, imagine a U-shaped log-likelihood function that is a mirror image of the function in Figure 3.2. With a U-shaped log-likelihood, the first derivative takes on a value of zero at the *lowest* point on the function (i.e., the bottom of the valley). Consequently, setting the first derivative formula to zero and solving for the unknown parameter value yields an estimate with the lowest possible log-likelihood value. The fact that the peak and the valley of a function both have first derivative values of zero is problematic because there is no way to differentiate the "best" and "worst" parameter values based on first derivatives alone. From the perspective of the first derivative formula, the top of the hill and the bottom of the valley look identical because both points on the function have a zero slope.

Fortunately, the sign of the second derivative provides a mechanism for differentiating the minimum and the maximum of a function. To illustrate, imagine climbing to the top of the log-likelihood function in Figure 3.2 beginning at a value of $\mu = 95$. The first derivatives

are positive during the ascent to the top of the function and become negative on the descent past $\mu = 100$. This sequence of positive to negative derivatives produces the negative sloping line (i.e., negative second derivative) in Figure 3.6. In contrast, imagine traversing a U-shaped function beginning at a value of $\mu = 95$. In this case, the first derivatives are negative during the descent to the minimum of the function and turn positive during the ascent back up the hill. Unlike Figure 3.6, this sequence of negative to positive values yields a line with a positive slope (i.e., a positive second derivative). Consequently, a negative second derivative indicates that the parameter estimate is located at the maximum, rather than the minimum, of the log-likelihood function.

3.8 MAXIMUM LIKELIHOOD ESTIMATION WITH MULTIVARIATE NORMAL DATA

A univariate example is useful for illustrating the mathematics behind maximum likelihood estimation, but most realistic applications of maximum likelihood (including maximum likelihood missing data handling) rely on the multivariate normal distribution. Applying maximum likelihood to multivariate data is typically more complex because the search process involves several parameters. In the subsequent sections, I use the IQ and job performance scores from Table 3.1 to extend the previous estimation principles to two variables. A bivariate analysis is still relatively straightforward, but the underlying logic generalizes to data sets with any number of variables.

As its name implies, the multivariate normal distribution generalizes the normal curve to multiple variables. For example, Figure 3.7 shows a multivariate normal distribution with two variables. This **bivariate normal distribution** retains the familiar shape of the normal curve and looks like a bell-shaped mound in three-dimensional space. Consistent with the univariate normal curve, a probability density function defines the shape of the multivariate normal distribution:

$$L_i = \frac{1}{(2\pi)^{k/2}\left|\boldsymbol{\Sigma}\right|^{1/2}} e^{-.5(\mathbf{Y}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i-\boldsymbol{\mu})} \tag{3.9}$$

The univariate density function has three primary components: a score value, the population mean, and the population variance. These quantities now appear as matrices in Equation 3.9. Specifically, each individual now has a set of k scores that are contained in the score vector $\mathbf{Y}_i$. Similarly, the equation replaces the mean and the variance with a mean vector and a covariance matrix (i.e., $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively). The key portion of the formula is the Mahalanobis distance value to the right of the exponent, $(\mathbf{Y}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i - \boldsymbol{\mu})$. Despite the shift to matrices, this portion of the formula is still a squared z score that quantifies the standardized distance between an individual's data points and the center of the multivariate normal distribution. Consistent with the univariate normal density, small deviations between the score vector and the mean vector produce large likelihood (i.e., relative probability) values, whereas large deviations yield small likelihoods. Finally, the collection of terms to the left of the exponent symbol is a scaling factor that makes the area under the distribution sum (i.e., integrate) to 1.

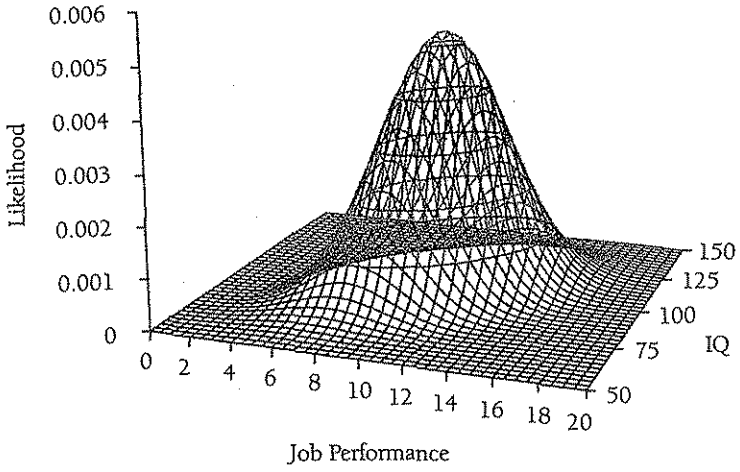

FIGURE 3.7. A bivariate normal distribution. The population mean and variance of the IQ variable are 100 and 189.60, respectively, and the mean and variance of the job performance variable are 10.35 and 6.83, respectively. The correlation between the variables is .55.

Computing Individual Likelihoods

The multivariate normal density describes the relative probability of drawing a *set* of scores from a multivariate normal distribution with a particular mean vector and covariance matrix. To illustrate the computations, consider the IQ and job performance scores in Table 3.1. For the sake of demonstration, assume that the population parameter values are as follows:

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_{IQ} \\ \mu_{JP} \end{bmatrix} = \begin{bmatrix} 100.00 \\ 10.35 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2_{IQ} & \sigma_{IQ,JP} \\ \sigma_{JP,IQ} & \sigma^2_{JP} \end{bmatrix} = \begin{bmatrix} 189.60 & 19.50 \\ 19.50 & 6.83 \end{bmatrix}$$

To begin, consider the individual who has an IQ score of 99 and a job performance rating of 7. Substituting these scores into Equation 3.9 yields a likelihood value of .0018, as follows:

$$L_i = \frac{1}{(2\pi)^{2/2} \begin{vmatrix} 189.60 & 19.50 \\ 19.50 & 6.83 \end{vmatrix}^{1/2}} e^{-.5\left(\begin{bmatrix} 99 \\ 7 \end{bmatrix} - \begin{bmatrix} 100.00 \\ 10.35 \end{bmatrix}\right)^T \begin{bmatrix} 189.60 & 19.50 \\ 19.50 & 6.83 \end{bmatrix}^{-1} \left(\begin{bmatrix} 99 \\ 7 \end{bmatrix} - \begin{bmatrix} 100.00 \\ 10.35 \end{bmatrix}\right)} = .0018$$

In the context of a bivariate analysis, the likelihood is the relative probability of drawing scores of 99 and 7 from a bivariate normal distribution with the previous mean vector and covariance matrix. Visually, the likelihood is the height of the bivariate normal distribution at the point where scores of 99 and 7 intersect. Next, consider the case with IQ and job performance scores of 87 and 7, respectively. Substituting these scores into the density function returns a likelihood value of 0.0022.

At first glance, the previous likelihood values might seem counterintuitive because the pair of scores with the largest deviations from the mean (i.e., 87 and 7) produces the higher

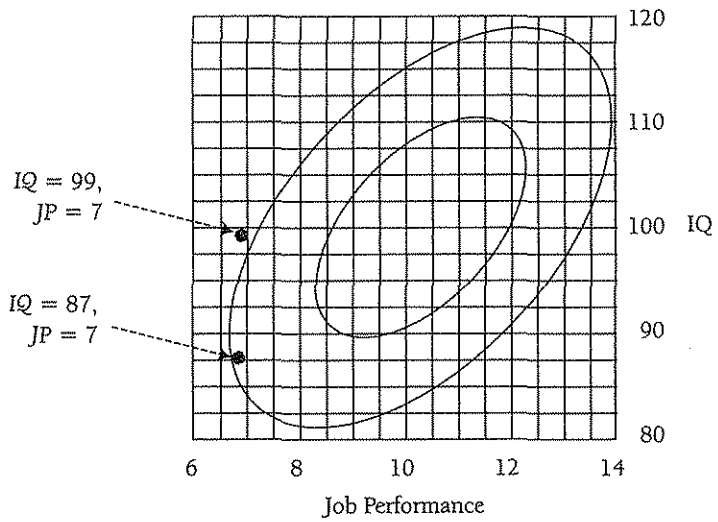

FIGURE 3.8. The bivariate normal distribution shown from an overhead perspective. The center of the distribution ($\mu = 100, 10.35$) is located in the middle of the ellipse. The location of two pairs of scores is marked by a •. The angle of the ellipse indicates a positive correlation between IQ and job performance. Because of the positive correlation, the intersection of 87 and 7 is actually at a slightly higher elevation (i.e., closer to the center of the distribution) than the intersection of 99 and 7.

likelihood value (i.e., better fit). To illustrate why this is the case, Figure 3.8 shows the bivariate normal distribution from an overhead perspective with contour rings that denote the elevation of the surface. The diagonal orientation of the contour rings follows from the fact that the two variables are positively correlated. This, in turn, puts the intersection of 87 and 7 at a slightly higher elevation (i.e., closer to the center of the distribution) than the intersection of 99 and 7. The Mahalanobis distance measure that quantifies the standardized distance between the score vector and the mean vector accounts for the positive correlation, so the seemingly counterintuitive likelihood values are accurate. Interested readers can consult any number of multivariate statistics textbooks for additional details on Mahalanobis distance (e.g., Johnson & Wichern, 2007; Tabachnick & Fidell, 2007).

The Multivariate Normal Log-Likelihood

As I explained earlier in the chapter, computing the natural logarithm of the individual likelihood values simplifies the mathematics of maximum likelihood. The individual log-likelihood for multivariate normal data is

$$\log L_i = \log\left\{\frac{1}{(2\pi)^{k/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-.5(Y_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(Y_i-\boldsymbol{\mu})}\right\} \tag{3.10}$$

where the terms in the braces produce the likelihood value for case i. After distributing the logarithm, the individual log-likelihood becomes

$$\log L_i = -\frac{k}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(Y_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(Y_i-\boldsymbol{\mu}) \tag{3.11}$$

Although Equations 3.10 and 3.11 are equivalent, the missing data literature often uses Equation 3.11 to express an individual's contribution to the sample log-likelihood. This formula will resurface in the next chapter, so it is worth mentioning at this point.

The log-likelihood values serve the same role and have the same meaning as the individual likelihoods. For example, reconsider the individual with an IQ score of 99 and a job performance rating of 7. The likelihood for this case is 0.0018, and the corresponding log-likelihood is –6.343. Next, the case with IQ and job performance scores of 87 and 7, respectively, has a likelihood value of 0.0022 and a log-likelihood of –6.113. Notice that the case with the highest likelihood value also has the highest (i.e., least negative) log-likelihood. Again, the log-likelihood values still quantify relative probability, but they do so using a different metric. Consequently, the score values of 87 and 7 have a better relative fit to the parameter values because they are closer to the center of the distribution.

Consistent with the univariate context, the sample log-likelihood is the sum of the individual log-likelihood values, as follows:

$$\log L = \sum_{i=1}^{N} \log L_i \tag{3.12}$$

As before, the sample log-likelihood is a summary measure that quantifies the fit of the sample data to the parameter estimates, such that higher values (i.e., values closer to zero) are indicative of better fit. Again, the sample log-likelihood provides a basis for choosing among a set of plausible parameter values.

Identifying the Maximum Likelihood Estimates

Estimating the parameters of the multivariate normal distribution (i.e., the mean vector and the covariance matrix) follows the same logic as univariate estimation. Conceptually, the estimation routine repeats the log-likelihood computations many times, each time with different estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Each unique combination of parameter estimates yields a different log-likelihood value, and the goal of estimation is to identify the particular constellation of estimates that produce the highest log-likelihood and thus the best fit to the data. Again, model fitting programs tend to use calculus derivatives to facilitate the estimation process.

Although the logic of estimation does not change much with multivariate data, identifying the maximum likelihood estimates is more complex because the search process involves multiple parameters. As an illustration, consider a simple bivariate analysis where the goal is to estimate the IQ and job performance means from the data in Table 3.1. The log-likelihood equation now depends on five parameters (i.e., two means and three unique covariance matrix elements), but fixing the covariance matrix elements to their sample estimates simplifies the illustration and has no impact on the mean estimates. Fixing the covariance matrix elements leaves the variable means as the only unknown quantities in Equation 3.11. I conducted a comprehensive search by computing the sample log-likelihood for many different combinations of the IQ and job performance means. Figure 3.9 shows the resulting log-likelihood values plotted against the values of μ_{IQ} and μ_{JP}. The log-likelihood function is now a three-dimensional surface with the pair of maximum likelihood estimates located at its peak. The orientation of the graph makes it difficult to precisely determine the parameter values,

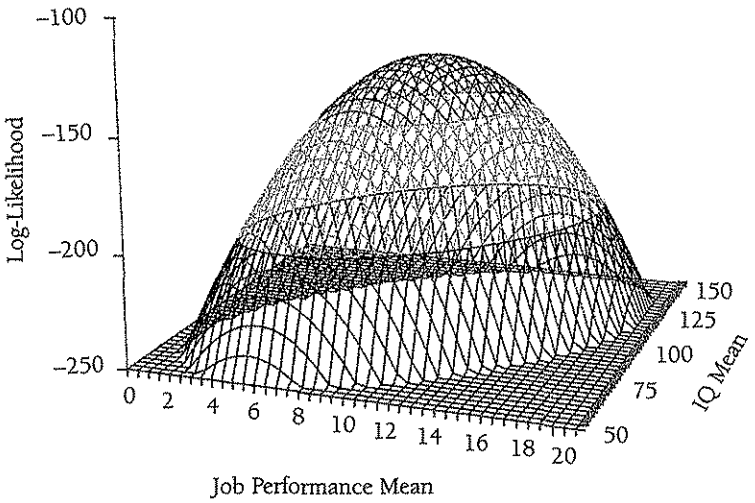

FIGURE 3.9. The log-likelihood function for a bivariate estimation problem. The figure shows the log-likelihood values plotted against different estimates of the IQ and job performance means. The maximum likelihood estimates are located at the peak of the function, which is roughly located at the intersection of $\mu_{IQ} = 100$ and $\mu_{JP} = 10$.

but the maximum of the function is approximately located at the intersection of $\mu_{IQ} = 100$ and $\mu_{JP} = 10$.

3.9 A BIVARIATE ANALYSIS EXAMPLE

Figure 3.9 provides a rough estimate of the variable means, but a more precise solution requires differential calculus. I described the role of first derivatives earlier in the chapter, so there is no need to delve deeper into the calculus details. Instead, I use the analysis results from a statistical software package to illustrate the details of a bivariate analysis. The maximum likelihood estimates of the mean vector and covariance matrix from the data in Table 3.1 are as follows:

$$\hat{\boldsymbol{\mu}} = \begin{bmatrix} \hat{\mu}_{IQ} \\ \hat{\mu}_{JP} \end{bmatrix} = \begin{bmatrix} 100.00 \\ 10.35 \end{bmatrix}$$

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} \hat{\sigma}^2_{IQ} & \hat{\sigma}_{IQ,JP} \\ \hat{\sigma}_{JP,IQ} & \hat{\sigma}^2_{JP} \end{bmatrix} = \begin{bmatrix} 189.60 & 19.50 \\ 19.50 & 6.83 \end{bmatrix}$$

The maximum likelihood means are identical to the usual sample means, but the variances and covariances are somewhat different because they use N in the denominator rather than $N - 1$. For example, the standard formula for the sample variance yields $\hat{\sigma}^2_{IQ} = 199.58$ and $\hat{\sigma}^2_{JP} = 7.19$. The negative bias in the maximum likelihood estimates is particularly evident in this example because of the small sample size. However, the bias quickly becomes negligible as the sample size increases, so it is usually not a major concern.

Recall from a previous section that maximum likelihood standard errors involve four computational steps: (1) calculate the value of the second derivative, (2) multiply the second

TABLE 3.5. Hessian, Information, and Parameter Covariance Matrices from the Bivariate Analysis Example

Parameter	1	2	3	4	5
			Hessian matrix		
1: μ_{IQ}	–0.149358				
2: μ_{JP}	0.426582	–4.147689			
3: μ^2_{IQ}	0	0	–0.000558		
4: $\mu_{IQ,JP}$	0	0	0.003186	–0.040073	
5: μ^2_{JP}	0	0	–0.004549	0.088466	–0.430083
			Information matrix		
1: μ_{IQ}	0.149358				
2: μ_{JP}	–0.426582	4.147689			
3: μ^2_{IQ}	0	0	0.000558		
4: $\mu_{IQ,JP}$	0	0	–0.003186	0.040073	
5: μ^2_{JP}	0	0	0.004549	–0.088466	0.430083
			Parameter covariance matrix		
1: μ_{IQ}	9.480000				
2: μ_{JP}	0.975000	0.341375			
3: μ^2_{IQ}	0	0	3594.816100		
4: $\mu_{IQ,JP}$	0	0	369.719890	83.737176	
5: μ^2_{JP}	0	0	38.024977	13.313618	4.661474

Note. Bold typeface denotes the sampling variance (i.e., squared standard error) of each parameter estimate.

derivative by negative one, (3) compute the inverse (i.e., reciprocal) of the previous product, and (4) take the square root of the resulting inverse. With multivariate analyses, the basic steps remain the same, but the computations involve matrices. Because each parameter has a unique derivative formula, the standard error computations start with a matrix of second derivatives. This so-called **Hessian matrix** is a symmetric matrix that has the same number of rows and columns as the number of parameters. The top panel of Table 3.5 shows the Hessian matrix for the bivariate analysis example. As seen in the table, the Hessian is a 5 by 5 symmetric matrix where each row and column corresponds to one of the estimated parameters. The diagonal elements contain the second derivatives, and the off-diagonal elements quantify the extent to which the log-likelihood functions for two parameters share similar curvature. Notice that the diagonal elements of the matrix are negative, which verifies that the parameter estimates are located at the maximum of the log-likelihood function.

The elements of the Hessian have a visual interpretation that is similar to that of the previous univariate example. To illustrate, consider the block of derivative values that correspond to the variable means (i.e., the elements in the upper left corner of the matrix). Returning to Figure 3.9, imagine standing midway along the IQ axis at the base of the log-likelihood surface. From this perspective, the log-likelihood would appear as a two-dimensional hill, and

the derivative value of –0.149 quantifies the curvature of that hill. Similarly, imagine viewing the log-likelihood surface from the midway point of the job performance axis. The derivative value of –4.148 quantifies the curvature of the two-dimensional hill from this angle. Finally, the off-diagonal element of 0.427 essentially quantifies the extent to which the two estimates have similar curvature (i.e., whether their first derivatives are changing at a similar rate across the function).

The second computational step multiplies the second derivatives by negative 1. In the univariate example, this operation produced a quantity known as information. In a multivariate analysis, multiplying the Hessian matrix by negative 1 yields the so-called **information matrix** (also known as **Fisher's information matrix**). As seen in the middle panel of Table 3.5, this step simply reverses the sign of each element in the Hessian. The main diagonal of the information matrix contains the information for each parameter estimate. These values quantify the curvature of each parameter's log-likelihood function, holding the other parameters constant.

With a single parameter, taking the reciprocal of information gives the sampling variance (i.e., squared standard error). There is no division in matrix algebra, but the inverse of a matrix is analogous to the reciprocal of a single number. Illustrating how to compute the inverse of a matrix is beyond the scope of this book, and there is typically no need to perform these computations by hand. The important point is that the inverse of the information matrix is another symmetric matrix known as the **parameter covariance matrix**. The bottom panel of Table 3.5 shows the parameter covariance matrix for the bivariate analysis example. The diagonal elements of the parameter covariance matrix contain sampling variances (i.e., squared standard errors), and the off-diagonals contain covariances between pairs of estimates. These covariances quantify the extent to which two estimates are dependent on one another. The diagonal elements of the parameter covariance matrix are particularly important because the square roots of these values are the maximum likelihood standard errors. For example, the standard error of the IQ mean is $\sqrt{9.480} = 3.079$, and the standard error of the covariance between IQ and job performance is $\sqrt{83.737} = 9.151$. As an aside, the block of zeros in the parameter covariance matrix follow from the fact that the mean and the covariance structure of the data are independent (e.g., recall from the earlier univariate example that I was able to estimate the mean without worrying about the variance). This is a well-established characteristic of maximum likelihood estimation with complete data.

3.10 ITERATIVE OPTIMIZATION ALGORITHMS

Estimating a mean vector and a covariance matrix is relatively straightforward because the first derivatives of the log-likelihood function produce familiar equations that define the maximum likelihood estimates. Maximum likelihood estimation is actually far more flexible than my previous examples imply because the mean vector and the covariance matrix can be functions of other model parameters. For example, a multiple regression analysis expresses the mean vector and the covariance matrix as a function of the regression coefficients and a residual variance estimate. Similarly, a confirmatory factor analysis model defines $\boldsymbol{\Sigma}$ as a model-

implied covariance matrix that depends on factor loadings, residual variances, and the latent variable covariance matrix, and it defines $\boldsymbol{\mu}$ as a model-implied mean vector, the values of which depend on factor means, factor loadings, and measurement intercepts (Bollen, 1989). Estimating one of these more complex models typically involves a collection of equations, each of which contains one or more unknown parameter values. Because solving for the unknown parameter values in a set of equations can be complex, advanced applications of maximum likelihood estimation generally require iterative optimization algorithms. A detailed overview of optimization algorithms could easily fill an entire chapter, so I give a brief conceptual explanation of the process. Eliason (1993) provides an accessible overview of a few common algorithms.

To understand how iterative algorithms work, imagine climbing to the top of the log-likelihood surface in Figure 3.9. The first step is to choose the starting coordinates for the hike. Starting the climb from a position that is close to the peak is advantageous because it reduces the number of steps required to get to the top. Iterative algorithms also require some initial coordinates, and these coordinates take the form of a set of **starting values** that provide an initial guess about the parameter estimates. Model fitting programs generally default to a set of starting values that do not closely resemble the true parameter values (e.g., correlation values of zero). However, many programs allow the user to specify starting values, and there are good reasons for doing so. For one, good starting values can reduce the number of steps to the peak of the log-likelihood function. In addition, some log-likelihood surfaces are difficult to climb because they are comprised of a number of smaller peaks and valleys. A good set of initial coordinates can improve the chances of locating the maximum of the function as opposed to the top of one of the smaller peaks (i.e., a local maximum).

After determining the initial coordinates, the rest of the climb consists of a series of steps toward the peak of the log-likelihood surface. Each step corresponds to a single iteration of the optimization process. Getting to the top requires a positioning device that keeps the climb going in a vertical direction, and the sample log-likelihood essentially serves as the algorithm's altimeter. At the first step, the algorithm substitutes the starting values into the density function and computes the log-likelihood. The goal of each subsequent step is to adjust the parameter values in a direction that increases the log-likelihood value. Algorithms differ in the numerical methods that they use to make these sequential improvements. For example, the EM algorithm I described in Chapter 4 uses a regression-based procedure, whereas other optimization routines (e.g., the scoring algorithm) use derivatives to adjust the parameters and improve the log-likelihood.

The log-likelihood keeps the algorithm climbing in a vertical direction, but it also determines when the climb is finished. The first few steps toward the peak often produce the largest changes in the log-likelihood (and thus the parameters), whereas the latter steps yield much smaller changes. In effect, the optimization algorithm traverses the steepest portion of the ascent at the beginning of the hike, and the climb becomes more gradual near the plateau. As the algorithm nears the peak of the function, each additional step produces a very small improvement in the log-likelihood value (i.e., a small change in altitude). Near the end of the climb, the adjustments to the parameter estimates are so small that the log-likelihood effectively remains the same between successive steps. At this point, the climb is over, and the algorithm has converged on the maximum likelihood estimates.

3.11 SIGNIFICANCE TESTING USING THE WALD STATISTIC

Testing whether a parameter estimate is within sampling error of some hypothesized value is an important part of a statistical analysis. Maximum likelihood estimation provides two significance testing mechanisms: the Wald statistic and the likelihood ratio test. This section outlines univariate and multivariate versions of the Wald statistic, and the next section describes the likelihood ratio test. The univariate Wald test is analogous to the t statistic from an ordinary least squares analysis, and its multivariate counterpart is akin to an omnibus F statistic.

The Univariate Wald Test

The univariate Wald statistic compares the difference between a point estimate and a hypothesized value to the standard error, as follows:

$$\omega = \frac{\hat{\theta} - \theta_0}{SE} \tag{3.13}$$

where $\hat{\theta}$ is a maximum likelihood parameter estimate, and θ_0 is some hypothesized value. Researchers typically want to determine whether a parameter is significantly different from zero, in which case the Wald test reduces to the ratio of the point estimate to its standard error. Maximum likelihood estimates are asymptotically (i.e., in very large samples) normally distributed, so the standard normal distribution generates a probability value for the Wald test. For this reason, the methodology literature sometimes refers to the test as the Wald z statistic.

Squaring Equation 3.13 gives an alternate formulation of the Wald test. This version of the test is

$$\omega = \frac{(\hat{\theta} - \theta_0)^2}{\text{var}(\hat{\theta})} \tag{3.14}$$

where $\text{var}(\hat{\theta})$ is the sampling variance (i.e., squared standard error) of the parameter. Squaring a standard normal z score yields a chi-square variable, so a central chi-square distribution with one degree of freedom generates a probability value for this version of the test. The chi-square formulation of the Wald test is arguably more flexible because it can accommodate multiple parameters.

To illustrate the Wald test, consider the covariance between IQ scores and job performance ratings. The previous bivariate analysis produced a parameter estimate of $\hat{\sigma}_{JPIQ} = 19.50$ and a standard error of $SE = 9.15$. Using the Wald z test to determine whether the estimate is significantly different from zero gives $\omega = (19.50—0) / 9.15 = 2.13$, and referencing the test statistic to a unit normal table returns a two-tailed probability value of $p = .03$. Alternatively, Equation 3.14 yields a Wald test of $\omega = (19.50—0)^2 / 9.15^2 = 4.54$. Referencing this value against a chi-square distribution with one degree of freedom also yields $p = .03$, so the choice of test statistic makes no difference.

The Multivariate Wald Test

In many situations it is of interest to determine whether a set of parameters is significantly different from zero. For example, in a multiple regression analysis, researchers are often interested in testing whether two or more regression slopes are mutually different from zero. In an ordinary least squares analysis with complete data, it is standard practice to use an omnibus *F* test for this purpose. In the context of maximum likelihood estimation, the multivariate Wald test is an analogous procedure.

The multivariate Wald test is

$$\omega = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \text{var}(\hat{\boldsymbol{\theta}})^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \tag{3.15}$$

where $\hat{\boldsymbol{\theta}}$ is a vector of parameter estimates, $\boldsymbol{\theta}_0$ is a vector of hypothesized values (typically zeros), and $\text{var}(\hat{\boldsymbol{\theta}})$ contains the elements from the parameter covariance matrix that correspond to the estimates in $\hat{\boldsymbol{\theta}}$. Equation 3.15 is fundamentally the same as its univariate counterpart, but it replaces each term in Equation 3.14 with a matrix (with a single parameter, Equation 3.15 reduces to Equation 3.14). If the null hypothesis is true, the multivariate Wald test follows a central chi-square distribution with degrees of freedom equal to the number of parameters in $\hat{\boldsymbol{\theta}}$. I illustrate this test in one of the data analysis examples later in the chapter.

3.12 THE LIKELIHOOD RATIO TEST STATISTIC

The likelihood ratio test is a common alternative to the Wald statistic. Like the Wald statistic, the likelihood ratio test is flexible and can accommodate a single estimate or multiple estimates. However, the likelihood ratio test takes the very different tack of comparing the relative fit of two **nested models**. Nested models can take on a variety of different forms, but a common example occurs when the parameters from one model are a subset of the parameters from a second model. For example, consider a multiple regression analysis in which a researcher is interested in testing whether two regression slopes are significantly different from zero. In this situation, the regression analysis that includes both predictor variables serves as the **full model**, and a second regression analysis that constrains the regression slopes to zero during estimation is the **restricted model**. The difference between the log-likelihood values from the two analyses provides the basis for a significance test. The restricted model can also differ from the full model by a set of complex parameter constraints. For example, in a confirmatory factor analysis, the full model is a saturated model (e.g., a model that estimates the sample covariance matrix), and the restricted model is the factor model that expresses the population covariance matrix as a function of the factor model parameters. The so-called chi-square test of model fit is a likelihood ratio test that compares the relative fit of these two models.

The likelihood ratio test is

$$\text{LR} = -2(\log L_{\text{Restricted}} - \log L_{\text{Full}}) \tag{3.16}$$

where $\log L_{\text{Restricted}}$ and $\log L_{\text{Full}}$ are the log-likelihood values from the restricted and the full models, respectively. The restricted model always has fewer parameters than the full model, so its log-likelihood must be less than or equal to that of the full model (i.e., because it uses fewer parameters to explain the data, the restricted model must have worse fit than the full model). The question is whether the log-likelihood difference exceeds random chance. If the null hypothesis is true (i.e., the full and restricted models have the same fit), the likelihood ratio follows a central chi-square distribution with degrees of freedom equal to the difference in the number of estimated parameters between the two models. A significant likelihood ratio test indicates that the restricted model does not fit the data as well as the full model (e.g., the estimates in question are significantly different from zero).

To illustrate the likelihood ratio test, reconsider the covariance between IQ scores and job performance ratings. To begin, I estimated the mean vector and the covariance matrix from the data in Table 3.1. This initial analysis estimated five parameters (two means and three unique covariance matrix elements) and served as the full model for the likelihood ratio test. Next, I estimated a restricted model by constraining the covariance to a value of zero during estimation (statistical software packages routinely allow users to specify parameter constraints such as this). The two models produced log-likelihood values of $\log L_{\text{Full}} = -124.939$ and $\log L_{\text{Restricted}} = -128.416$. Notice that the log-likelihood for the restricted model is somewhat lower than that of the full model, which suggests that the restricted model has worse fit to the data. Substituting the log-likelihood values into Equation 3.16 gives a likelihood ratio statistic of LR = 6.96. The two models differ by a single parameter, so a chi-square distribution with one degree of freedom generates a probability value for the test, $p = .008$. The fact that the restricted model is significantly worse than that of the full model suggests that the covariance between IQ and job performance is statistically different from zero (i.e., constraining the covariance to zero during estimation significantly degrades model fit).

3.13 SHOULD I USE THE WALD TEST OR THE LIKELIHOOD RATIO STATISTIC?

The Wald test and the likelihood ratio statistic can address identical hypotheses, so the natural question is, "Which test should I use?" The answer to this question largely depends on the sample size and the parameters that you are testing. The two tests are asymptotically (i.e., in very large samples) equivalent but can give markedly different answers in small to moderate samples (Buse, 1982). The potential for different test results stems from the fact that some parameter estimates (e.g., variances, covariances, correlations) have skewed sampling distributions. These sampling distributions eventually normalize as the sample size gets very large, but they can be markedly nonnormal in small and moderate samples. This is a problem for the Wald test because it uses the normal distribution to generate probability values (Fears, Benichou, & Gail, 1996; Pawitan, 2000). The likelihood ratio test makes no assumptions about the shape of the sampling distribution, so it is generally superior to the Wald test in small samples.

Statistical issues aside, there are practical considerations to examine when choosing between the Wald and likelihood ratio tests. First, Wald tests are easy to implement because most software packages produce these tests as part of their standard output. The likelihood ratio test is somewhat less convenient because it requires two analyses. In addition, it is often necessary to compute the likelihood ratio test by hand, although this is not a compelling disadvantage. Second, the Wald test is not invariant to changes in model parameterization (Fears, Benichou, & Gail, 1996). For example, researchers frequently estimate confirmatory factor analysis models by fixing either the factor variance or a factor loading to 1. These parameterizations are statistically equivalent (i.e., have the same degrees of freedom and produce the same model fit) but are likely to produce different Wald statistics (Gonzalez & Griffin, 2001). In contrast, the likelihood ratio statistic is invariant to model parameterization, so its value is unaffected by the choice of model specification.

As a final word of caution, non-normal data (particularly excessive kurtosis) can distort the values of the Wald test and the likelihood ratio statistic (e.g., Finney & DiStefano, 2006; West, Finch, & Curran, 1995). Methodological studies have repeatedly demonstrated that non-normal data can inflate type I error rates, so you should interpret these tests with some caution. Fortunately, methodologists have developed corrective procedures for non-normal data, so it is relatively easy to obtain accurate inferences. I outline some of these corrective procedures in Chapter 5.

3.14 DATA ANALYSIS EXAMPLE 1

In the remainder of the chapter, I use two data analysis examples to illustrate maximum likelihood estimation. The first analysis example uses maximum likelihood to estimate a mean vector, covariance matrix, and a correlation matrix.* The data for this analysis are comprised of scores from 480 employees on eight work-related variables: gender, age, job tenure, IQ, psychological well-being, job satisfaction, job performance, and turnover intentions. I generated these data to mimic the correlation structure of published research articles in the management and psychology literature (e.g., Wright & Bonett, 2007; Wright, Cropanzano, & Bonett, 2007).

Table 3.6 shows the maximum likelihood estimates along with the estimates from the usual sample formulas. Notice that the two sets of means are identical, but the maximum likelihood estimates of variances and covariances are slightly smaller in value. I previously explained that maximum likelihood estimates of variances and covariances are negatively biased because they use N rather than $N - 1$ in the denominator. However, with a sample size of 480, the difference in the two sets of estimates is essentially trivial. As an aside, some software packages implement a **restricted maximum likelihood** estimator that effectively uses $N - 1$ to compute variance components (e.g., see Raudenbush & Bryk, 2002, pp. 52–53).

*Analysis syntax and data are available on the companion website, *www.appliedmissingdata.com*.

TABLE 3.6. Mean, Covariance, and Correlation Estimates from Data Analysis Example 1

Variable	1.	2.	3.	4.	5.	6.	7.	8.
			Maximum likelihood					
1. Age	28.908	0.504	–0.010	0.182	0.111	–0.049	–0.150	0.015
2. Tenure	8.459	9.735	–0.034	0.173	0.157	0.016	0.011	0.001
3. Female	–0.028	–0.052	0.248	0.097	0.038	–0.015	0.005	0.068
4. Well-being	1.208	0.667	0.060	1.518	0.348	0.447	–0.296	0.306
5. Satisfaction	0.697	0.576	0.022	0.503	1.377	0.176	–0.222	0.378
6. Performance	–0.330	0.061	–0.009	0.690	0.259	1.570	–0.346	0.426
7. Turnover	–0.377	0.016	0.001	–0.170	–0.122	–0.203	0.218	–0.180
8. IQ	0.674	0.026	0.284	3.172	3.730	4.496	–0.706	70.892
Means	37.948	10.054	0.542	6.271	5.990	6.021	0.321	100.102
			Sample formulas					
1. Age	28.968	0.504	–0.010	0.182	0.111	–0.049	–0.150	0.015
2. Tenure	8.477	9.755	–0.034	0.173	0.157	0.016	0.011	0.001
3. Female	–0.028	–0.052	0.249	0.097	0.038	–0.015	0.005	0.068
4. Well-being	1.210	0.668	0.060	1.521	0.348	0.447	–0.296	0.306
5. Satisfaction	0.699	0.577	0.022	0.504	1.380	0.176	–0.222	0.378
6. Performance	–0.331	0.062	–0.009	0.692	0.259	1.574	–0.346	0.426
7. Turnover	–0.378	0.016	0.001	–0.171	–0.122	–0.203	0.218	–0.180
8. IQ	0.676	0.026	0.285	3.179	3.738	4.505	–0.707	71.040
Means	37.948	10.054	0.542	6.271	5.990	6.021	0.321	100.102

Note. Correlations are in the upper diagonal in bold typeface.

3.15 DATA ANALYSIS EXAMPLE 2

The second analysis example applies maximum likelihood estimation to a multiple regression model.* The analysis uses the employee data set from the first example to estimate the regression of job performance ratings on psychological well-being and job satisfaction, as follows:

$$JP_i = \beta_0 + \beta_1(WB_i) + \beta_2(SAT_i) + \varepsilon$$

Structural equation modeling software programs are a convenient platform for implementing maximum likelihood estimation, with or without missing data. Figure 3.10 shows the path diagram of the regression model. Path diagrams use single-headed straight arrows to denote regression coefficients and double-headed curved arrows to represent correlations. In addition, the diagrams differentiate manifest variables and latent variables using rectangles and ellipses, respectively (Bollen, 1989; Kline, 2005). In Figure 3.10, the predictor variables and the outcome variable are manifest variables (e.g., scores from a questionnaire), and the

* Analysis syntax and data are available on the companion website, *www.appliedmissingdata.com*.

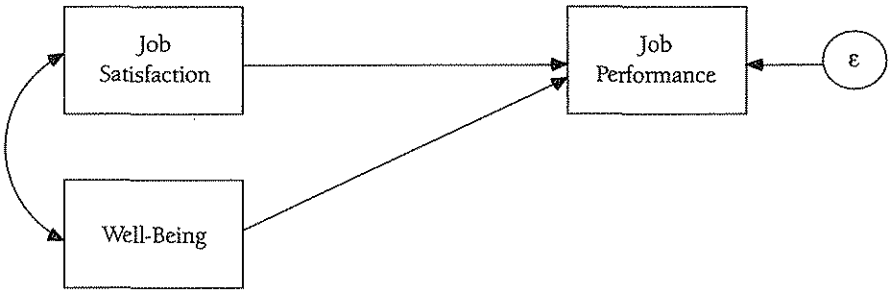

FIGURE 3.10. A path diagram for a multiple regression model. The single-headed straight lines represent regression coefficients, the double-headed curved arrow is a correlation, the rectangles are manifest variables, and the ellipse is a latent variable.

residual term is a latent variable that captures a collection of unobserved influences on the outcome variable.

Researchers typically begin a regression analysis by examining the omnibus F test. As a baseline for comparison, a least squares analysis produced a significant omnibus test, $F(2, 247) = 60.87$, $p < .001$. The likelihood ratio statistic and the multivariate Wald test are analogous procedures in a maximum likelihood analysis. To begin, consider the likelihood ratio test. The full model corresponds to the regression in Figure 3.10, and the restricted model is one that constrains both regression slopes to zero during estimation (the regression intercept is not part of the usual omnibus F test, so it appears in both models). Estimating the two models produced log-likelihood values of $\log L_{\text{Full}} = -1130.977$ and $\log L_{\text{Restricted}} = -1181.065$, respectively. Notice that log-likelihood for the restricted model is quite a bit lower than that of the full model, which suggests that fixing the slopes to zero deteriorates model fit. Substituting the log-likelihood values into Equation 3.16 yields a likelihood ratio statistic of LR = 100.18. The two models differ by two parameters (i.e., the restricted model constrains two coefficients two zero); therefore, referencing the test statistic to a chi-square distribution with two degrees of freedom returns a probability value of $p < .001$. The significant likelihood ratio test indicates that the fit of the restricted model is significantly worse than that of the full model. Consistent with the interpretation of the F statistic, this suggests that at least one of the regression coefficients is significantly different from zero.

For the purpose of illustration, I also used the multivariate Wald statistic to construct an omnibus test. Recall from Equation 3.15 that the Wald test requires elements from the parameter covariance matrix. Software packages that implement maximum likelihood estimation typically offer the option to print this matrix, although it may not be part of the standard output. The regression model has four parameter estimates (i.e., the regression intercept, two slope coefficients, and a residual variance), so the full parameter covariance matrix has four rows and four columns. However, the Wald test only requires the covariance matrix elements for the two slope coefficients (i.e., the 2 by 2 submatrix that contains the sampling variance of each coefficient and the covariance between the two estimates). Substituting the regression coefficients ($\hat{\beta} = 0.025$ and 0.446) and the appropriate elements from the parameter covariance matrix into Equation 3.15 gives a Wald statistic of $\omega = 119.25$, as follows:

$$\omega = \begin{bmatrix} .025 \\ .446 \end{bmatrix}^T \begin{bmatrix} .002175 & -.000720 \\ -.000720 & .001973 \end{bmatrix}^{-1} \begin{bmatrix} .025 \\ .446 \end{bmatrix} = 119.25$$

TABLE 3.7. Regression Model Estimates from Data Analysis Example 2

Parameter	Est.	*SE*	*z*
	Maximum likelihood		
β_0 (Intercept)	6.021	0.051	117.705
β_1 (Well-being)	0.446	0.044	10.083
β_2 (Satisfaction)	0.025	0.046	0.533
σ_ε^2 (Residual)	1.256		
R^2	.200		
	Ordinary least squares		
β_0 (Intercept)	6.021	0.051	117.337
β_1 (Well-being)	0.446	0.044	10.050
β_2 (Satisfaction)	0.025	0.047	0.531
σ_ε^2 (Residual)	1.262		
R^2	.200		

Note. Ordinary least squares uses a *t* statistic.

Comparing the Wald test to a chi-square distribution with two degrees of freedom (i.e., there are two parameters under consideration) returns a probability value of $p < .001$. Consistent with the *F* test and the likelihood ratio statistic, the Wald test suggests that at least one of the regression slopes is different from zero.

Researchers typically follow up a significant omnibus test by examining partial regression coefficients. Table 3.7 gives the maximum likelihood estimates along with those from an ordinary least squares analysis. As seen in the table, psychological well-being was a significant predictor of job performance, $\hat{\beta}_1 = 0.446$, $z = 10.08$, $p < .001$, but job satisfaction was not, $\hat{\beta}_2 = 0.025$, $z = 0.53$, $p = .59$. The interpretation of the estimates is the same for both estimators. For example, holding job satisfaction constant, a one-point increase in psychological well-being yields a .446 increase in job performance ratings, on average. Notice that maximum likelihood and ordinary least squares produced identical regression coefficients but slightly different residual variance estimates. The two estimators share the same equations for the regression coefficients, so it makes sense that these estimates are identical. The slight difference between the residual variances owes to the fact that maximum likelihood variance estimates are negatively biased. Again, the discrepancy in this example is trivial, but the bias would be more apparent in small samples.

3.16 SUMMARY

Many modern statistical procedures that are in widespread use today rely on maximum likelihood estimation. Maximum likelihood also plays a central role in missing data analyses and is one of two approaches that methodologists currently regard as the state of the art (Schafer & Graham, 2002). The purpose of this chapter was to introduce the mechanics of maximum likelihood estimation in the context of a complete-data analysis. Researchers in the social and

the behavioral sciences routinely assume that their variables are normally distributed in the population, so I described maximum likelihood in the context of multivariate normal data. The normal distribution supplies a familiar platform for illustrating estimation principles, but it also provides the basis for the missing data handling procedure outlined in subsequent chapters.

The goal of maximum likelihood estimation is to identify the population parameters that have the highest probability of producing the sample data. The sample log-likelihood is central to this process because it quantifies the relative probability of drawing a sample of scores from a normal distribution with a particular mean vector and covariance matrix. Substituting a score value (or a set of scores) into a probability density function returns the log-likelihood value for a single case, and the sample log-likelihood is the sum of the individual log-likelihood values. The sample log-likelihood quantifies the fit between the data and the parameter estimates and provides a basis for choosing among a set of plausible parameter values.

Conceptually, estimation is an iterative process that repeatedly auditions different parameter values until it finds the estimates that are most likely to have produced the data. The estimation procedure essentially repeats the log-likelihood calculations many times, each time substituting different values of the population parameters into the log-likelihood equation. Each unique combination of parameter estimates yields a different log-likelihood value, and the goal of estimation is to identify the particular constellation of estimates that produce the highest log-likelihood and thus the best fit to the data. In some situations, the first derivatives of the log-likelihood function produce familiar equations that define the maximum likelihood estimates, but more complex applications of maximum likelihood estimation (including missing data handling) require iterative optimization algorithms to identify the most likely parameter values.

The curvature of the log-likelihood function provides important information about the uncertainty of an estimate. A flat log-likelihood function makes it difficult to discriminate among competing estimates because the log-likelihood values are relatively similar across a range of parameter estimates. In contrast, a steep log-likelihood function more clearly differentiates the maximum likelihood estimate from other possible parameter values. Mathematically, the second derivative quantifies the curvature of the log-likelihood function. Second derivatives largely determine the maximum likelihood standard errors, such that larger second derivatives (i.e., more peaked functions) translate into smaller standard errors and smaller second derivatives (i.e., flatter functions) translate into larger standard errors.

Maximum likelihood analyses provide two significance testing mechanisms, the Wald statistic and the likelihood ratio test. The univariate Wald test is the ratio of the point estimate to its standard error. The multivariate Wald test is similar to its univariate counterpart but uses matrices to determine whether a set of estimates is significantly different from zero. The likelihood ratio test is procedurally different from the Wald statistic because it requires two analysis models: a full model that includes the parameters of substantive interest, and a restricted model (i.e., nested model) that constrains one or more of the parameter values to zero during estimation. The difference between the log-likelihood values from the two models provides the basis for a significance test. Like the Wald statistic, the likelihood ratio test is flexible and can accommodate a single estimate or multiple estimates. The two test statis-

tics are asymptotically equivalent, but the likelihood ratio test is generally superior in small samples.

Chapter 4 extends maximum likelihood estimation to missing data analyses. Conceptually, maximum likelihood estimation works the same way with or without missing data. Consistent with a complete-data analysis, the ultimate goal is to identify the parameter estimates that maximize the log-likelihood and produce the best fit to the data. However, the incomplete data records require a slight alteration to the individual log-likelihood equation. Missing data also introduce some nuances to the standard error computations. I describe these changes in detail in the next chapter.

3.17 RECOMMENDED READINGS

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

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

Enders, C. K. (2005). Estimation by maximum likelihood. In B. Everitt & D.C. Howell (Eds.), *Encyclopedia of behavioral statistics* (pp. 1164–1170). West Sussex, UK: Wiley.