4

Maximum Likelihood Missing Data Handling

4.1 CHAPTER OVERVIEW

Having established some basic estimation principles with complete data, this chapter describes maximum likelihood missing data handling (the literature sometimes refers to this procedure as **full information maximum likelihood** and **direct maximum likelihood**). The idea of using maximum likelihood to deal with missing data is an old one that dates back more than 50 years (Anderson, 1957; Edgett, 1956; Hartley, 1958; Lord, 1955). These early maximum likelihood solutions were limited in scope and had relatively few practical applications (e.g., bivariate normal data with a single incomplete variable). Many of the important breakthroughs came in the 1970s when methodologists developed the underpinnings of modern missing data handling techniques (Beale & Little, 1975; Finkbeiner, 1979; Dempster, Laird, & Rubin, 1977; Hartley & Hocking, 1971; Orchard & Woodbury, 1972). However, maximum likelihood routines have only recently become widely available in statistical software packages.

Recall from Chapter 3 that maximum likelihood estimation repeatedly auditions different combinations of population parameter values until it identifies the particular constellation of values that produces the highest log-likelihood value (i.e., the best fit to the data). Conceptually, the estimation process is the same with or without missing data. However, missing data introduce some additional nuances that are not relevant for complete-data analyses. For one, incomplete data records require a slight alteration to the individual log-likelihood computations to accommodate the fact that individuals no longer have the same number of observed data points. Missing data also necessitate a subtle, but important, adjustment to the standard error computations. Finally, with few exceptions, missing data analyses require iterative optimization algorithms, even for very simple estimation problems. This chapter describes one such algorithm that is particularly important for missing data analyses, the expectation maximization (EM) algorithm.

Methodologists currently regard maximum likelihood as a state-of-the-art missing data technique (Schafer & Graham, 2002) because it yields unbiased parameter estimates under a missing at random (MAR) mechanism. From a practical standpoint, this means that maximum likelihood will produce accurate parameter estimates in situations where traditional approaches fail. Even when the data are missing completely at random (MCAR), maximum likelihood will still be superior to traditional techniques (e.g., deletion methods) because it maximizes statistical power by borrowing information from the observed data. Despite these desirable properties, maximum likelihood estimation is not a perfect solution and will yield biased parameter estimates under a missing not at random (MNAR) mechanism. However, this bias tends to be isolated to a subset of the analysis model parameters, whereas traditional techniques are more apt to propagate bias throughout the entire model. Consequently, maximum likelihood estimation is virtually always a better option than the traditional methods from Chapter 2. The fact that maximum likelihood is easy to implement and is widely available in statistical software packages makes it all the more attractive.

I use the small data set in Table 4.1 to illustrate ideas throughout this chapter. I designed these data to mimic an employee selection scenario in which prospective employees complete an IQ test and a psychological well-being questionnaire during their interview. The company subsequently hires the applicants who score in the upper half of the IQ distribution, and a supervisor rates their job performance following a 6-month probationary period. Note that the job performance scores are MAR because they are systematically missing as a function of IQ scores (i.e., individuals in the lower half of the IQ distribution were never

TABLE 4.1. Employee Selection Data Set

IQ	Psychological well-being	Job performance
78	13	—
84	9	—
84	10	—
85	10	—
87	—	—
91	3	—
92	12	—
94	3	—
94	13	—
96	—	—
99	6	7
105	12	10
105	14	11
106	10	15
108	—	10
112	10	10
113	14	12
115	14	14
118	12	16
134	11	12

hired and thus have no performance rating). In addition, I randomly deleted three of the well-being scores in order to mimic an MCAR mechanism (e.g., the human resources department inadvertently loses an applicant's well-being questionnaire). This data set is too small for a serious application of maximum likelihood estimation, but it is useful for illustrating the basic mechanics of the procedure.

4.2 THE MISSING DATA LOG-LIKELIHOOD

Recall from Chapter 3 that the starting point for a maximum likelihood analysis is to specify a distribution for the population data. To be consistent with the previous chapter, I describe maximum likelihood missing data handling in the context of multivariate normal data. The mathematical machinery behind maximum likelihood relies on a probability density function that describes the shape of the multivariate normal distribution. Substituting a score vector and a set of population parameter values into the density function returns a likelihood value that quantifies the relative probability of drawing the scores from a normally distributed population. Because likelihood values tend to be very small numbers that are prone to rounding error, it is more typical to work with the natural logarithm of the likelihood values (i.e., the log-likelihood). Rather than rehash the computational details of the likelihood values, I use the individual log-likelihood as the starting point for this chapter. Readers who are interested in more information on the likelihood can review Chapter 3.

Assuming a multivariate normal distribution for the population, note that the complete-data log-likelihood for a single case is

$$\log L_i = -\frac{k}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathrm{Y}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathrm{Y}_i - \boldsymbol{\mu}) \tag{4.1}$$

where k is the number of variables, Y_i is the score vector for case i, and $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the population mean vector and covariance matrix, respectively. The key portion of the formula is the Mahalanobis distance value, $(\mathrm{Y}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathrm{Y}_i - \boldsymbol{\mu})$. Mahalanobis distance is a squared z score that quantifies the standardized distance between an individual's data points and the center of the multivariate normal distribution. This value largely determines the magnitude of the log-likelihood, such that small deviations between the score vector and the mean vector produce large (i.e., less negative) log-likelihood values, whereas large deviations yield small likelihoods. In simple terms, Equation 4.1 quantifies the relative probability that an individual's scores originate from a multivariate normal population with a particular mean vector and covariance matrix.

With missing data, the log-likelihood for case i is

$$\log L_i = -\frac{k_i}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}(\mathrm{Y}_i - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathrm{Y}_i - \boldsymbol{\mu}_i) \tag{4.2}$$

where k_i is the number of complete data points for that case and the remaining terms have the same meaning as they did in Equation 4.1. At first glance, the two log-likelihood formulas look identical, except for the fact that the missing data log-likelihood has an i subscript next to the parameter matrices. This subscript is important and denotes the possibility that the size and the contents of the matrices can vary across individuals, such that the log-likelihood computations for case i depend only on the variables and the parameters for which that case has complete data.

To illustrate the missing data log-likelihood, suppose that the company wants to use the data in Table 4.1 to estimate the mean vector and the covariance matrix. Estimating these parameters is relatively straightforward with complete data but requires an iterative optimization algorithm when some of the data are missing. For the sake of demonstration, suppose that the population parameters at a particular iteration are as follows:

$$\hat{\boldsymbol{\mu}} = \begin{bmatrix} \hat{\mu}_{IQ} \\ \hat{\mu}_{JP} \\ \hat{\mu}_{WB} \end{bmatrix} = \begin{bmatrix} 100.00 \\ 10.23 \\ 10.27 \end{bmatrix}$$

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} \hat{\sigma}^2_{IQ} & \hat{\sigma}_{IQ,JP} & \hat{\sigma}_{IQ,WB} \\ \hat{\sigma}_{JP,IQ} & \hat{\sigma}^2_{JP} & \hat{\sigma}_{JP,WB} \\ \hat{\sigma}_{WB,IQ} & \hat{\sigma}_{WB,JP} & \hat{\sigma}^2_{WB} \end{bmatrix} = \begin{bmatrix} 189.60 & 22.31 & 12.21 \\ 22.31 & 8.68 & 5.61 \\ 12.21 & 5.60 & 11.04 \end{bmatrix}$$

The log-likelihood computations for each individual depend only on the variables and the parameters for which a case has complete data. This implies that the log-likelihood formula looks slightly different for each missing data pattern. Returning to the data set in Table 4.1, observe four unique missing data patterns: (1) cases with only IQ scores, (2) cases with IQ and well-being scores, (3) cases with IQ and job performance scores, and (4) cases with complete data on all three variables. To begin, consider the employee with an IQ score of 105, a job performance rating of 10, and a well-being score of 12. Because this individual has complete data, the log-likelihood computations involve every element in the mean vector and the covariance matrix, as follows:

$$\log L_i = -\frac{k_i}{2}\log(2\pi) - \frac{1}{2}\log\begin{vmatrix} \hat{\sigma}^2_{IQ} & \hat{\sigma}_{IQ,JP} & \hat{\sigma}_{IQ,WB} \\ \hat{\sigma}_{JP,IQ} & \hat{\sigma}^2_{JP} & \hat{\sigma}_{JP,WB} \\ \hat{\sigma}_{WB,IQ} & \hat{\sigma}_{WB,JP} & \hat{\sigma}^2_{WB} \end{vmatrix}$$

$$-\frac{1}{2}\left(\begin{bmatrix} IQ_i \\ JP_i \\ WB_i \end{bmatrix} - \begin{bmatrix} \hat{\mu}_{IQ} \\ \hat{\mu}_{JP} \\ \hat{\mu}_{WB} \end{bmatrix}\right)^T \begin{bmatrix} \hat{\sigma}^2_{IQ} & \hat{\sigma}_{IQ,JP} & \hat{\sigma}_{IQ,WB} \\ \hat{\sigma}_{JP,IQ} & \hat{\sigma}^2_{JP} & \hat{\sigma}_{JP,WB} \\ \hat{\sigma}_{WB,IQ} & \hat{\sigma}_{WB,JP} & \hat{\sigma}^2_{WB} \end{bmatrix}^{-1} \left(\begin{bmatrix} IQ_i \\ JP_i \\ WB_i \end{bmatrix} - \begin{bmatrix} \hat{\mu}_{IQ} \\ \hat{\mu}_{JP} \\ \hat{\mu}_{WB} \end{bmatrix}\right)$$

$$= -\frac{3}{2}\log(2\pi) - \frac{1}{2}\log\begin{vmatrix} 189.60 & 22.31 & 12.21 \\ 22.31 & 8.68 & 5.61 \\ 12.21 & 5.60 & 11.04 \end{vmatrix}$$

$$-\frac{1}{2}\left(\begin{bmatrix} 105 \\ 10 \\ 12 \end{bmatrix} - \begin{bmatrix} 100.00 \\ 10.23 \\ 10.27 \end{bmatrix}\right)^T \begin{bmatrix} 189.60 & 22.31 & 12.21 \\ 22.31 & 8.68 & 5.61 \\ 12.21 & 5.60 & 11.04 \end{bmatrix}^{-1} \left(\begin{bmatrix} 105 \\ 10 \\ 12 \end{bmatrix} - \begin{bmatrix} 100.00 \\ 10.23 \\ 10.27 \end{bmatrix}\right) = -7.66$$

Consistent with complete-data maximum likelihood estimation, –7.66 is the relative probability of drawing this set of three scores from a multivariate normal distribution with the previous parameter values. The log-likelihood computations for the remaining complete cases follow the same procedure, but use different score values.

Next, consider the subsample of cases with IQ and well-being scores. These individuals have missing job performance ratings, so it is no longer possible to use all three variables to compute the log-likelihood. The missing data log-likelihood accommodates this situation by ignoring the parameters that correspond to the missing job performance ratings. For example, consider the individual with IQ and well-being scores of 94 and 3, respectively. Eliminating the job performance parameters from the mean vector and the covariance matrix leaves the following subset of parameter estimates.

$$\hat{\boldsymbol{\mu}}_i = \begin{bmatrix} \hat{\mu}_{IQ} \\ \hat{\mu}_{WB} \end{bmatrix} = \begin{bmatrix} 100.00 \\ 10.27 \end{bmatrix}$$

$$\hat{\boldsymbol{\Sigma}}_i = \begin{bmatrix} \hat{\sigma}^2_{IQ} & \hat{\sigma}_{IQ,WB} \\ \hat{\sigma}_{WB,IQ} & \hat{\sigma}^2_{WB} \end{bmatrix} = \begin{bmatrix} 189.60 & 12.21 \\ 12.21 & 11.04 \end{bmatrix}$$

The log-likelihood computations use only these parameter values, as follows:

$$\log L_i = -\frac{k_i}{2}\log(2\pi) - \frac{1}{2}\log\begin{vmatrix} \hat{\sigma}^2_{IQ} & \hat{\sigma}_{IQ,WB} \\ \hat{\sigma}_{WB,IQ} & \hat{\sigma}^2_{WB} \end{vmatrix}$$

$$-\frac{1}{2}\left(\begin{bmatrix} IQ_i \\ WB_i \end{bmatrix} - \begin{bmatrix} \hat{\mu}_{IQ} \\ \hat{\mu}_{WB} \end{bmatrix}\right)^T \begin{bmatrix} \hat{\sigma}^2_{IQ} & \hat{\sigma}_{IQ,WB} \\ \hat{\sigma}_{WB,IQ} & \hat{\sigma}^2_{WB} \end{bmatrix}^{-1} \left(\begin{bmatrix} IQ_i \\ WB_i \end{bmatrix} - \begin{bmatrix} \hat{\mu}_{IQ} \\ \hat{\mu}_{WB} \end{bmatrix}\right)$$

$$= -\frac{2}{2}\log(2\pi) - \frac{1}{2}\log\begin{vmatrix} 189.60 & 12.21 \\ 12.21 & 11.04 \end{vmatrix}$$

$$-\frac{1}{2}\left(\begin{bmatrix} 94 \\ 3 \end{bmatrix} - \begin{bmatrix} 100.00 \\ 10.27 \end{bmatrix}\right)^T \begin{bmatrix} 189.60 & 12.21 \\ 12.21 & 11.04 \end{bmatrix}^{-1} \left(\begin{bmatrix} 94 \\ 3 \end{bmatrix} - \begin{bmatrix} 100.00 \\ 10.27 \end{bmatrix}\right) = -8.03$$

Notice that the log-likelihood equation no longer contains any reference to the job performance variable. Thus, the resulting log-likelihood value is the relative probability of drawing the two scores from a bivariate normal distribution with a mean vector and covariance matrix equal to $\hat{\boldsymbol{\mu}}_i$ and $\hat{\boldsymbol{\Sigma}}_i$, respectively. Again, the log-likelihood computations for the remaining cases that share this missing data pattern follow the same approach.

As a final example, consider the subsample of cases that have data on the IQ variable only. Consistent with the previous example, the log-likelihood computations ignore the parameters that correspond to the missing variables, leaving only the IQ parameters.

$$\hat{\boldsymbol{\mu}}_i = [\hat{\mu}_{IQ}] = [100.00]$$

$$\hat{\boldsymbol{\Sigma}}_i = [\hat{\sigma}^2_{IQ}] = [189.60]$$

To illustrate, the log-likelihood for the employee with an IQ score of 87 is as follows:

$$\log L_i = -\frac{k_i}{2}\log(2\pi) - \frac{1}{2}\log|\hat{\sigma}^2_{IQ}| - \frac{1}{2}(IQ_i - \hat{\mu}_{IQ})^T(\hat{\sigma}^2_{IQ})^{-1}(IQ_i - \hat{\mu}_{IQ})$$

$$= -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log|189.60| - \frac{1}{2}(87 - 100)^T(189.60)^{-1}(87 - 100)$$

$$= -3.99$$

The log-likelihood value is now the relative probability of drawing an IQ score of 87 from a univariate normal distribution with a mean of 100 and a variance of 189.60.

Table 4.2 shows the log-likelihood values for all 20 employees. Consistent with complete-data estimation, the sample log-likelihood is the sum of the individual log-likelihood values. For example, summing the log-likelihood values in Table 4.2 gives $\log L = -146.443$. Despite the missing values, the sample log-likelihood is still a summary measure that quantifies the joint probability of drawing the observed data from a normally distributed population with a particular mean vector and covariance matrix (e.g., the previous estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$). Furthermore, the estimation process follows the same logic as Chapter 3. Conceptually, an iterative optimization algorithm repeats the log-likelihood computations many times, each time with different estimates of the population parameters. Each unique combination of parameter estimates yields a different log-likelihood value. The goal of estimation is to identify

TABLE 4.2. Individual Log-Likelihood Values

IQ	Psychological well-being	Job performance	$\log L_i$
78	13	—	–7.73904
84	9	—	–6.30206
84	10	—	–6.32745
85	10	—	–6.24113
87	—	—	–3.98707
91	3	—	–8.02047
92	12	—	–6.03874
94	3	—	–8.02968
94	13	—	–6.19267
96	—	—	–3.58359
99	6	7	–8.37010
105	12	10	–7.66375
105	14	11	–8.07781
106	10	15	–9.54606
108	—	10	–5.64284
112	10	10	–7.88229
113	14	12	–8.23350
115	14	14	–8.33434
118	12	16	–9.49084
134	11	12	–10.73921

the particular constellation of estimates that produce the highest log-likelihood and thus the best fit to the data. Importantly, the estimation algorithm does not need to impute or replace the missing values. Rather, it uses all of the available data to estimate the parameters and the standard errors.

As an aside, maximum likelihood missing data handling is 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. 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). I illustrate some of these more advanced applications of maximum likelihood estimation later in the chapter.

4.3 HOW DO THE INCOMPLETE DATA RECORDS IMPROVE ESTIMATION?

Using all of the available data to estimate the parameters is an intuitively appealing approach, but it is not necessarily obvious why including the incomplete data records improves the accuracy of the resulting parameter estimates. A bivariate analysis in which one of the variables has missing data may provide deeper insight into the estimation process. Returning to the data in Table 4.1, suppose that the company wants to estimate the IQ and job performance means. Table 4.3 shows the maximum likelihood estimates along with those of listwise deletion. By virtue of the selection process, listwise deletion discards the entire lower half of the IQ distribution (the company only hires applicants with high IQ scores, so low-scoring applicants do not contribute to the analysis). Because IQ scores and job performance ratings are positively correlated, listwise deletion also excludes cases from the lower tail of the job performance distribution. Not surprisingly, the remaining cases are unrepresentative of the hypothetically complete data set because they have systematically higher scores on both variables. Consequently, the listwise deletion mean estimates are too high. In contrast, the maximum likelihood estimates are relatively similar to those of the complete data. An analysis based on a sample size of 20 does not provide compelling evidence in favor of maximum

TABLE 4.3. IQ and Job Performance Means from the Employee Selection Data

Estimator	$\hat{\mu}_{IP}$	$\hat{\mu}_{JP}$
Complete data	100.00	10.35
Maximum likelihood	100.00	10.28
Listwise deletion	111.50	11.70

Note. The complete data estimates are from the data in Table 3.1.

likelihood estimation, but the estimates in Table 4.3 are consistent with Rubin's (1976) theoretical predictions for an MAR mechanism.

The log-likelihood equation can provide some insight into the differences between the maximum likelihood and listwise deletion parameter estimates. With missing data, the individual log-likelihood computations depend only on the variables and the parameter estimates for which a case has data. Because the bivariate analysis has just two missing data patterns (i.e., cases with complete data and cases with data on IQ only), there are two log-likelihood formulas. The individual log-likelihood equation for the subsample of employees with complete data is

$$\begin{aligned}\log L_i = &-\frac{k_i}{2}\log(2\pi) - \frac{1}{2}\log\begin{vmatrix}\sigma^2_{IQ} & \sigma_{JP,IQ}\\ \sigma_{JP,IQ} & \sigma^2_{JP}\end{vmatrix}\\ &-\frac{1}{2}\left(\begin{bmatrix}IQ_i\\ JP_i\end{bmatrix}-\begin{bmatrix}\mu_{IQ}\\ \mu_{JP}\end{bmatrix}\right)^T\begin{bmatrix}\sigma^2_{IQ} & \sigma_{JP,IQ}\\ \sigma_{JP,IQ} & \sigma^2_{JP}\end{bmatrix}^{-1}\left(\begin{bmatrix}IQ_i\\ WB_i\end{bmatrix}-\begin{bmatrix}\mu_{IQ}\\ \mu_{JP}\end{bmatrix}\right)\end{aligned} \tag{4.3}$$

and eliminating the job performance parameters gives the individual log-likelihood equation for the applicants with incomplete data, as follows:

$$\log L_i = -\frac{k_i}{2}\log(2\pi) - \frac{1}{2}\log|\sigma^2_{IQ}| - \frac{(IQ_i-\mu_{IQ})^2}{2\sigma^2_{IQ}} \tag{4.4}$$

Finally, summing the previous equations across the entire sample gives the sample log-likelihood

$$\begin{aligned}\log L = &\left\{-n_C\left(\frac{k_i}{2}\log[2\pi] - \frac{1}{2}\log|\boldsymbol{\Sigma}|\right) - \frac{1}{2}\sum_{i=1}^{n_C}(\mathbf{Y}_i-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i-\boldsymbol{\mu})\right\}\\ &-n_M\left(\frac{k_i}{2}\log[2\pi] - \frac{1}{2}\log|\sigma^2_{IQ}|\right) - \frac{1}{2\sigma^2_{IQ}}\sum_{i=1}^{n_M}(IQ_i-\mu_{IQ})^2 \qquad (4.5)\\ &= \{\log L_{\text{Complete}}\} + \log L_{\text{Incomplete}}\end{aligned}$$

where n_C is the number of complete cases, and n_M is the number of incomplete cases. To make the equation more compact, I do not display the individual matrix elements from Equation 4.3 (e.g., $\boldsymbol{\mu}$ replaces the vector in Equation 4.3 that contains μ_{IQ} and μ_{JP}).

Equation 4.5 is useful because it partitions the sample log-likelihood into two components. The bracketed terms reflect the contribution of the complete cases to the sample log-likelihood, and remaining terms contain the additional information from the incomplete data records. A maximum likelihood analysis based on the 10 complete cases (i.e., an analysis that uses only the bracketed terms) would produce the listwise estimates in Table 4.3. This implies that the incomplete data records are solely responsible for differences between the listwise deletion and maximum likelihood parameter estimates. In some sense, the portion of the log-likelihood equation for the incomplete cases serves as a correction factor that steers the estimator to a more accurate set of parameter estimates.

The fact that maximum likelihood better estimates the IQ mean should come as no surprise because the variable is complete. The accuracy of the job performance mean is less intuitive when you consider that the incomplete cases have no job performance ratings. To illustrate how the incomplete data records affect estimation, Table 4.4 shows the sample log-likelihood for different combinations of the IQ and job performance means. For simplicity, I limited the IQ estimates to values of 100.00 and 111.50 (these are the maximum likelihood and listwise deletion estimates, respectively). The column labeled $\log L_{\text{Complete}}$ contains the sample log-likelihood values from a listwise deletion analysis (i.e., maximum likelihood estimation based only on the bracketed terms in Equation 4.5); the column labeled $\log L_{\text{Incomplete}}$ shows the log-likelihood contribution for the incomplete data records; and the $\log L$ column gives the sample log-likelihood values for maximum likelihood missing data handling (i.e., the sum of $\log L_{\text{Complete}}$ and $\log L_{\text{Incomplete}}$).

Recall that the goal of estimation is to identify the constellation of parameter values that produces the highest log-likelihood and thus the best fit to the data. As seen in the $\log L_{\text{Complete}}$ column, a listwise deletion analysis would produce estimates of $\hat{\mu}_{IQ} = 111.50$ and $\hat{\mu}_{JP} = 11.75$ because this combination of parameter values has the highest (i.e., least negative) log-likelihood value. (Had I used smaller increments for the job performance mean, these estimates would exactly match the listwise estimates in Table 4.3.) Next, the $\log L_{\text{Incomplete}}$ column gives the contribution of the 10 incomplete cases to the sample log-likelihood. Because these applicants do not have job performance ratings, the log-likelihood values are constant across different estimates of the job performance mean (i.e., Equation 4.4 depends only on the IQ parameters). However, the incomplete data records do carry information about the IQ mean, and the log-likelihood values suggest that $\mu_{IQ} = 100.00$ is more plausible than $\mu_{IQ} = 111.50$ (i.e., the log-likelihood for $\mu_{IQ} = 100.00$ is higher than that of $\mu_{IQ} = 111.50$). Finally, the $\log L$ column gives the sample log-likelihood values for maximum likelihood missing data handling. As you can see, $\mu_{IQ} = 100.00$ and $\mu_{JP} = 10.25$ provide the best fit to the data because this combination of parameter values has the highest log-likelihood.

Mathematically, the goal of maximum likelihood estimation is to identify the parameter values that minimize the standardized distances between the data points and the center of a multivariate normal distribution. Whenever the estimation process involves a set of model parameters, fine-tuning one estimate can lead to changes in the other estimates. This is precisely what happened in the bivariate analysis example. Specifically, the log-likelihood values in the $\log L_{\text{Incomplete}}$ column of Table 4.4 strongly favor a lower value for the IQ mean. Including these incomplete data records in the analysis therefore pulls the IQ mean down to a value that is identical to that of the complete data. Higher values for the job performance mean (e.g., $\mu_{JP} = 11.75$) are an unlikely match for an IQ mean of 100, so the downward adjustment to the IQ average effectively steers the estimator toward a job performance mean that more closely matches that of the complete data. In effect, maximum likelihood estimation improves the accuracy of the parameter estimates by "borrowing" information from the observed data (e.g., the IQ scores), some of which is contained in the incomplete data records. Although it is difficult to illustrate with equations, the same process applies to complex multivariate analyses with general missing data patterns.

TABLE 4.4. Sample Log-Likelihood Values for Different Combinations of the IQ and Job Performance Means

		Log-likelihood		
μ_{IQ}	μ_{JP}	$\log L_{\text{Complete}}$	$\log L_{\text{Incomplete}}$	$\log L$
100.00	10.00	–63.754	–39.694	–103.449
	10.25	–63.681	–39.694	–103.376
	10.50	–63.726	–39.694	–103.420
	10.75	–63.888	–39.694	–103.582
	11.00	–64.167	–39.694	–103.861
	11.25	–64.564	–39.694	–104.258
	11.50	–65.079	–39.694	–104.773
	11.75	–65.711	–39.694	–105.405
	12.00	–66.460	–39.694	–106.154
111.50	10.00	–62.909	–50.157	–113.066
	10.25	–62.169	–50.157	–112.326
	10.50	–61.547	–50.157	–111.703
	10.75	–61.041	–50.157	–111.198
	11.00	–60.654	–50.157	–110.810
	11.25	–60.383	–50.157	–110.540
	11.50	–60.231	–50.157	–110.387
	11.75	–60.195	–50.157	–110.352
	12.00	–60.278	–50.157	–110.434

4.4 AN ILLUSTRATIVE COMPUTER SIMULATION STUDY

The preceding bivariate analysis is useful for illustration purposes, but it does not offer compelling evidence about the performance of maximum likelihood missing data handling. To better illustrate the properties of maximum likelihood estimation, I conducted a series of Monte Carlo computer simulations. The simulation programs generated 1,000 samples of $N = 250$ from a population model that mimicked the IQ and job performance data in Table 4.1. The first simulation created MCAR data by randomly deleting 50% of the job performance ratings. The second simulation modeled MAR data and eliminated job performance scores for the cases in the lower half of the IQ distribution. The final simulation generated MNAR data by deleting the job performance scores for the cases in the lower half of the job performance distribution. After generating each data set, the simulation programs used maximum likelihood missing data handling to estimate the mean vector and the covariance matrix.

Table 4.5 shows the average parameter estimates from the simulations and uses bold typeface to highlight severely biased estimates. For comparison purposes, the table also shows the corresponding estimates from listwise deletion. As seen in the table, maximum likelihood and listwise deletion produced unbiased estimates in the MCAR simulation, and both sets of estimates were virtually identical. Although not shown in the table, the listwise deletion standard errors were generally 7 to 40% larger than those of maximum likelihood estimation. Not surprisingly, this translates into a substantial power advantage for maximum likelihood. The MAR simulation produced dramatic differences between the two missing data techniques,

TABLE 4.5. Average Parameter Estimates from the Illustrative Computer Simulation

Parameter	Population value	Maximum likelihood	Listwise deletion
	MCAR simulation		
μ_{IQ}	100.00	100.02	100.00
μ_{JP}	12.00	11.99	11.99
σ^2_{IQ}	169.00	168.25	166.94
σ^2_{JP}	9.00	8.96	8.94
$\sigma_{IQ,JP}$	19.50	19.48	19.31
	MAR simulation		
μ_{IQ}	100.00	100.01	110.35
μ_{JP}	12.00	12.01	13.18
σ^2_{IQ}	169.00	168.50	61.37
σ^2_{JP}	9.00	8.96	7.49
$\sigma_{IQ,JP}$	19.50	19.15	6.99
	MNAR simulation		
μ_{IQ}	100.00	100.00	105.19
μ_{JP}	12.00	14.12	14.38
σ^2_{IQ}	169.00	169.11	141.41
σ^2_{JP}	9.00	3.33	3.25
$\sigma_{IQ,JP}$	19.50	8.55	7.14

such that listwise deletion produced substantial bias, and the maximum likelihood estimates were quite accurate. Finally, both maximum likelihood and listwise deletion produced biased estimates in the MNAR simulation, although the bias in the maximum likelihood estimates was restricted to a subset of the parameter estimates. These simulation studies are limited in scope, but the results in Table 4.5 are predictable based on Rubin's (1976) missing data theory and are consistent with a number of published computer simulation studies (e.g., Arbuckle, 1996; Enders, 2001; Enders & Bandalos, 2001; Gold & Bentler, 2000; Muthén et al., 1987; Olinsky, Chen, & Harlow, 2003; Wothke, 2000).

You might recall from Chapter 2 that stochastic regression imputation is the only traditional missing data handling technique that also produces unbiased parameter estimates under an MCAR or MAR mechanism (see Table 2.5). The downside of stochastic regression is that it underestimates standard errors, potentially by a substantial amount. If its assumptions (multivariate normality and an MAR mechanism) are met, maximum likelihood estimation does not suffer from this same problem. To illustrate, I computed the confidence interval coverage rates from the MAR simulation. Confidence interval coverage quantifies the percentage of samples where the 95% confidence interval contains the true population parameter. If standard errors are accurate, confidence interval coverage should equal 95%. In contrast, if the standard errors are too low, confidence intervals will not capture the population parameter as frequently as they should, and coverage rates will drop below 95%. Confidence interval coverage rates are a useful indicator of standard error bias because they directly relate to type I error rates (e.g., a confidence interval coverage value of 90% suggests a twofold increase in type I errors). The confidence interval coverage values from the MAR simulation were quite close to the optimal 95% rate, which implies that the standard errors were relatively free of bias. In contrast, using stochastic regression imputation to analyze the same simulation data produced coverage rates between 60 and 70% (i.e., on average, standard errors were far too small). All things considered, the simulation results clearly favor maximum likelihood estimation, despite the fact that stochastic regression imputation requires identical assumptions.

4.5 ESTIMATING STANDARD ERRORS WITH MISSING DATA

Chapter 3 described the important role that second derivatives play in the computation of standard errors. Recall that the standard error computations begin with the matrix of second derivatives, the so-called Hessian matrix. Multiplying the Hessian by negative 1 yields the information matrix, and computing the inverse of the information gives the parameter covariance matrix. The diagonal elements of the parameter covariance matrix contain the sampling variances of the parameter estimates, and taking the square root of these elements gives the standard errors. The computational steps are identical with missing data, except that it is necessary to distinguish between standard errors based on the observed information matrix versus those based on the expected information matrix.

Recall that the information matrix contains values that quantify the curvature of the log-likelihood function. The magnitude of the curvature directly influences standard errors, such that peaked functions produce large information values and small standard errors, whereas flat functions produce small information values and large standard errors. In a missing data analysis, two approaches can be used to convert second derivatives into information values, and thus two approaches have developed for computing standard errors (with complete data, the observed and the expected information matrices tend to yield the same standard errors). The distinction between the two computational approaches is important because the expected information matrix yields standard errors that require the MCAR assumption, whereas the observed information matrix gives standard errors that are appropriate with MAR data (Kenward & Molenberghs, 1998; Little & Rubin, 2002; Molenberghs & Kenward, 2007). The next few sections describe the differences between these two procedures in more detail.

As an aside, some of the subsequent information is relatively technical in nature. For readers who are not interested in the mathematical details behind the two computational approaches, there is a simple take-home message: whenever possible, use the observed information matrix to compute standard errors. Many (but not all) software packages implement this method, although it may not be the default analysis option. Later in the chapter I present some simulation results that strongly favor standard errors based on the observed information matrix. It is therefore a good idea to consider this computational option when choosing a software package.

4.6 OBSERVED VERSUS EXPECTED INFORMATION

The **expected information** matrix replaces certain terms in the second derivative formulas with their expected values (i.e., long-run averages), whereas the **observed information** uses the realized data values to compute these terms. Before describing how this process applies to missing data, it is useful to demonstrate the computational approaches in the context of a complete-data scenario. Efron and Hinkley (1978) use an intuitive example that involves the weighted mean to illustrate the distinction between the observed and the expected information. In their example, one of two different measurement instruments generates a score for each case, and a coin toss determines which device generates each score. Because a coin toss dictates the use of each measurement instrument, the two instruments should generate the same number of scores over the long run, even though the observed frequency is likely to deviate from a 50/50 split in any given sample.

The standard error of the weighted mean relies on the score variance from each measurement instrument (i.e., σ_1^2 and σ_2^2) as well as on the number of observations that each device generates (i.e., n_1 and n_2). There are two options for computing the standard error of the weighted mean. Because the two instruments should generate the same number of observations over the long run, one approach is to weight the variances equally in the standard error computations. Weighting the variances by the realized values of n_1 and n_2 is also appropriate because the observed frequencies are unlikely to be exactly equal in any given sample. These two strategies are consistent with the notion of expected and observed information, respectively.

Computing the information (and thus the standard error) requires the second derivative of the log-likelihood function. The second derivative formula for the weighted mean is $-n_1/\sigma_1^2 - n_2/\sigma_2^2$. Because a random process with a probability of .50 dictates the values of n_1 and n_2, the expectation (i.e., long-run average) of these two values is $(n_1 + n_2)/2 = N/2$. Substituting this expectation into the second derivative formula in place of n_1 and n_2 and multiplying the derivative by negative 1 yields the following equation for the expected information

$$I_E = -E\left\{\frac{\partial^2 \log L}{\partial^2 \mu}\right\} = -E\left\{-\frac{n_1}{\sigma_1^2} - \frac{n_2}{\sigma_2^2}\right\} = -\left(-\frac{N/2}{\sigma_1^2} - \frac{N/2}{\sigma_2^2}\right) = \frac{N/2}{\sigma_1^2} + \frac{N/2}{\sigma_2^2} \tag{4.6}$$

where ∂^2 denotes the second derivative, and E is the expectation symbol. In contrast, the observed information relies on the realized values of n_1 and n_2, as follows:

$$I_O = -\left\{\frac{\partial^2 \log L}{\partial^2 \mu}\right\} = -\left\{-\frac{n_1}{\sigma_1^2} - \frac{n_2}{\sigma_2^2}\right\} = \frac{n_1}{\sigma_1^2} + \frac{n_2}{\sigma_2^2} \tag{4.7}$$

Following the procedures from Chapter 3, computing the inverse (i.e., reciprocal) of the information gives the sampling variance of the mean, and taking the square root of the sampling variance returns the standard error. As you can see, the two information equations will yield the same standard error only if the observed data (i.e., the values of n_1 and n_2) match the long-run expectation (i.e., $N/2$).

How Does the Observed and Expected Information Apply to Missing Data?

The previous example is useful for understanding the conceptual difference between the observed and the expected information, but it does not illustrate how these concepts apply to missing data analyses. Applying the previous ideas to missing data, we find that the realized values of n_1 are n_2 are roughly analogous to the observed missing data pattern. In the weighted mean example, the expected information yields standard errors that do not depend on the values of n_1 and n_2, whereas the observed information uses the values of n_1 and n_2 to compute standard errors. In the context of a missing data analysis, the expected information produces standard errors that effectively ignore the pattern of missing values, whereas standard errors based on the observed information depend on the missing data pattern. This distinction has important practical implications because the two computational approaches make different assumptions about the missing data mechanism.

The missing data literature refers to the MAR mechanism as ignorable missingness because the distribution of missing data carries no information about the analysis model parameters. Interestingly, the realized missing data pattern does contain information that influences the information matrix, and thus the standard errors (Kenward & Molenberghs, 1998; Little, 1976). Specifically, the expected information matrix yields standard errors that require the MCAR assumption, whereas the observed information matrix produces standard errors that are appropriate with MCAR and MAR data (Kenward & Molenberghs, 1998; Little & Rubin, 2002; Molenberghs & Kenward, 2007). Kenward and Molenberghs (1998) provide a detailed discussion of this issue, and I summarize their main points in the next section.

4.7 A BIVARIATE ANALYSIS EXAMPLE

To illustrate the difference between the observed and expected information, suppose that it is of interest to use the IQ scores and job performance ratings from Table 4.1 to estimate the mean vector and the covariance matrix. The matrix of second derivatives (i.e., the Hessian) for this analysis is a 5 by 5 symmetric matrix in which each row and column corresponds to one of the estimated parameters (there are two means and three unique covariance matrix elements). Furthermore, the diagonal elements of the Hessian matrix contain the second derivatives for each parameter, and the off-diagonal elements quantify the extent to which the log-likelihood functions for two parameters share similar curvature. Collectively, the elements in the Hessian matrix are the building blocks of maximum likelihood standard errors.

The observed and the expected information matrices differ in how they treat the deviation scores (i.e., $y_i - \mu$) that appear in certain second derivative formulas. In particular, the two computational approaches produce different values for the off-diagonal elements of the Hessian that involve a mean parameter and a covariance matrix parameter. To illustrate, consider the second derivative formula for the off-diagonal element that involves the mean and the variance of the IQ scores (i.e., μ_{IQ} and σ^2_{IQ}, respectively). The second derivative formula is

$$\frac{\partial^2 \log L}{\partial \mu_{IQ} \partial \sigma^2_{IQ}} = \left\{ -[1 \quad 0] \Sigma^{-1} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \Sigma^{-1} \sum_{i=1}^{n_C} \left(\begin{bmatrix} IQ_i \\ JP_i \end{bmatrix} - \begin{bmatrix} \mu_{IQ} \\ \mu_{JP} \end{bmatrix} \right) \right\} - \frac{1}{(\sigma^2_{IQ})^2} \sum_{i=1}^{n_M} (IQ_i - \mu_{IQ}) \quad (4.8)$$

where ∂^2 denotes a second derivative, n_C is the number of complete cases, and n_M is the number of incomplete cases. Note that the bracketed terms reflect the contribution of the complete cases to the second derivative value, and the remaining terms contain the additional information from the incomplete cases. Although the derivative formula is relatively complex, the deviation scores and their sums are the key to understanding the distinction between the observed and the expected information.

Consider what happens to Equation 4.8 when the data are complete. In this situation, the bracketed terms alone generate the second derivative and the remaining terms vanish. The observed information uses the realized data values (i.e., IQ_i and JP_i) to compute the second derivative. By definition, the sum of the deviation scores equals zero, so the entire second derivative equation returns a value of zero. In contrast, the expected information replaces the observed scores with their expected values (i.e., long-run averages). The expected value of a random variable is the mean, and so the data values in Equation 4.8 get replaced by their respective averages. In this situation, the sum of the deviation scores also equals zero, as does the value of the second derivative. With complete data, all of the second derivative equations that involve a mean parameter and a covariance matrix parameter work in the same fashion and return a value of zero (i.e., the mean parameters are independent of the covariance matrix parameters).

Thus far, using the observed data or the expected values to compute the second derivative formulas leads to the same answer. However, the two computational approaches diverge with missing data, and the second derivative values depend on the missing data mechanism. Consider what happens to Equation 4.8 when the job performance ratings are MCAR. If the values are missing in a purely random fashion, the observed job performance scores should be equally dispersed above and below the mean. Using the realized data values to compute the sums should therefore still produce a value of zero, on average. Consistent with the complete-data scenario, the expected information replaces the observed data values with their respective averages; thus, the deviation terms vanish and the entire equation returns a value of zero. Consequently, the observed and the expected information should produce the same second derivative value (and thus the same standard error), on average. Again, this result holds for any off-diagonal element of the Hessian that involves a mean parameter and a covariance matrix parameter.

The situation changes with MAR data. By virtue of the employee selection process, the job performance ratings in Table 4.1 are primarily missing from the lower tail of the score distribution. This implies that the observed data points are not equally dispersed above and below the mean. For example, a quick inspection of the data in Table 4.1 shows that the majority of the observed job performance ratings are above the maximum likelihood estimate of the mean, which is $\hat{\mu}_{JP} = 10.28$. Consequently, the sum of the deviation scores (and thus the value of the second derivative) no longer equals zero. In contrast, because the expected information replaces the observed data values with their respective averages, the second derivative formula will always return a value of zero, regardless of the missing data mechanism.

To numerically illustrate the differences between the observed and the expected information, Table 4.6 shows the information matrices and the parameter covariance matrices from the bivariate analysis. First, notice that the expected information matrix contains values of zero for the off-diagonal elements that involve a mean parameter and a covariance matrix

TABLE 4.6. Information and Parameter Covariance Matrices from the Bivariate Analysis Example

Parameter	1	2	3	4	5
			Information matrix (observed)		
1: μ_{IQ}	0.134132				
2: μ_{JP}	–0.232050	1.879713			
3: σ^2_{IQ}	0.001738	–0.014075	0.000492		
4: $\sigma_{IQ,JP}$	–0.014075	0.114012	–0.002065	0.022111	
5: σ^2_{JP}	0.000000	0.000001	0.002692	–0.043618	0.176666
			Parameter covariance matrix (observed)		
1: μ_{IQ}	**9.479986**				
2: μ_{JP}	1.170302	**1.507618**			
3: σ^2_{IQ}	–0.000020	0.000059	**3594.794000**		
4: $\sigma_{IQ,JP}$	–0.000044	–13.703090	443.774440	**280.705520**	
5: σ^2_{JP}	–0.000011	–3.383282	54.783599	62.542877	**20.267296**
			Information matrix (expected)		
1: μ_{IQ}	0.134132				
2: μ_{JP}	–0.232050	1.879713			
3: σ^2_{IQ}	0	0	0.000470		
4: $\sigma_{IQ,JP}$	0	0	–0.001889	0.020684	
5: σ^2_{JP}	0	0	0.002692	–0.043619	0.176666
			Parameter covariance matrix (expected)		
1: μ_{IQ}	**9.479986**				
2: μ_{JP}	1.170299	**0.676469**			
3: σ^2_{IQ}	0	0	**3594.805000**		
4: $\sigma_{IQ,JP}$	0	0	443.776810	**155.650330**	
5: σ^2_{JP}	0	0	54.784017	31.666846	**12.644013**

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

parameter. Again, this is a consequence of replacing the observed data values with their expectations. In contrast, the observed information matrix has nonzero off-diagonal elements, which suggests that the mean parameters and the covariance matrix parameters are no longer independent. Computing the inverse of the information matrix gives the parameter covariance matrix, the diagonal elements of which contain sampling variances (i.e., squared standard errors). Notice that the expected information matrix produces smaller sampling variances for the parameters affected by missing data. Considering the job performance mean, the observed information matrix gives a sampling variance of 1.508, whereas the expected information matrix produces an estimate of .676. Not surprisingly, the disparity between these two values translates into a marked difference in the standard errors. For example, the observed information matrix yields a standard error of 1.228, whereas the expected information matrix yields a standard error of 0.822.

The sampling variances in Table 4.6 illustrate that the two computational approaches can produce very different standard errors, particularly with MAR data. At an intuitive level, using the observed information is desirable because the standard errors take into account the realized missing data pattern. The methodological literature clearly favors this approach because the resulting standard errors are accurate with MAR data. Referring to the observed information matrix, Kenward and Molenberghs (1998, p. 238) stated that "its use in missing data problems should be the rule rather than the exception." Other authors have echoed this sentiment (Laird, 1988; Little & Rubin, 2002; Molenberghs & Kenward, 2007). Fortunately, many software packages can compute standard errors from the observed information matrix, although this may not be the default analysis option.

4.8 AN ILLUSTRATIVE COMPUTER SIMULATION STUDY

The results in Table 4.6 are useful for illustration purposes, but they do not provide strong evidence about the differences that can result from using the observed versus the expected information to compute standard errors. To better illustrate the performance of these computational approaches, I conducted a series of Monte Carlo computer simulations. The simulation programs generated 1,000 samples of $N = 250$ from a population model that mimicked the IQ and job performance data in Table 4.1. The first simulation created MCAR data by randomly deleting 50% of the job performance ratings, and the second simulation mimicked an MAR mechanism by eliminating the job performance scores for the cases in the lower half of the IQ distribution. After generating each data set, the simulation programs used maximum likelihood missing data handling to estimate the mean vector and the covariance matrix. They subsequently computed standard errors using both the observed and the expected information matrix.

Table 4.7 shows the average standard error for each parameter estimate. To gauge the accuracy of the standard errors, the table also gives the standard deviation of the parameter estimates across the 1,000 samples, along with the confidence interval coverage values. The standard deviations quantify the actual sampling fluctuation of the estimates and provide a benchmark for assessing the average standard errors. Confidence interval coverage quantifies the percentage of samples where the 95% confidence interval contains the true population parameter. If standard errors are accurate, confidence interval coverage should equal 95%. In contrast, if the standard errors are too low, confidence intervals will not capture the population parameter as frequently as they should, and coverage rates will drop below 95%. Confidence interval coverage rates are a useful indicator of standard error bias because they directly relate to type I error rates (e.g., a confidence interval coverage value of 90% suggests a twofold increase in type I errors).

As seen in the table, the two computational approaches produced nearly identical results in the MCAR simulation, and the standard errors from both methods were quite accurate (i.e., the average standard errors were quite close to the standard deviations, and the coverage values were roughly 95%). In the MAR simulation, the observed information matrix produced standard errors that closely resembled the standard deviation values (i.e., the true standard errors), and the corresponding confidence interval coverage values were quite close

TABLE 4.7. Simulation Results Comparing Observed and Expected Standard Errors

		Observed information		Expected information	
Parameter	SD	Average *SE*	Coverage	Average *SE*	Coverage
		MCAR simulation			
μ_{IQ}	0.791	0.820	0.963	0.820	0.963
μ_{JP}	0.247	0.250	0.951	0.250	0.951
σ^2_{IQ}	14.777	15.049	0.948	15.049	0.948
σ^2_{JP}	1.105	1.117	0.939	1.114	0.937
$\sigma_{IQ,JP}$	3.434	3.484	0.949	3.465	0.946
		MAR simulation			
μ_{IQ}	0.806	0.820	0.947	0.820	0.947
μ_{JP}	0.394	0.395	0.953	0.249	0.804
σ^2_{IQ}	15.074	15.071	0.949	15.071	0.949
σ^2_{JP}	1.490	1.439	0.920	1.112	0.851
$\sigma_{IQ,JP}$	5.275	5.283	0.959	3.463	0.795

to the optimal 95% rate. In contrast, the expected information matrix produced inaccurate standard errors for the parameters affected by missing data. For example, the standard error of the job performance mean was too small, on average, and had a coverage value of approximately 80%. From a practical standpoint, a confidence interval coverage value of 80% represents a type I error rate of approximately 20%, which is a fourfold increase over the nominal 5% type I error rate.

It is difficult to say whether the simulation results in Table 4.7 are representative of real-world analysis examples, but they clearly suggest that standard errors based on the expected information matrix are prone to severe bias and are only valid with MCAR data. Many (but not all) software programs can compute standard errors from the observed information matrix, so you should consider this option when choosing a software package. If you do not have access to software that computes the observed information matrix, you can always use the likelihood ratio statistic to perform significance tests (e.g., by fitting two models, one of which constrains the parameter of interest to zero during estimation) because the likelihood ratio is unaffected by the choice of information matrix.

4.9 AN OVERVIEW OF THE EM ALGORITHM

Certain complete-data applications of maximum likelihood estimation (e.g., the estimation of means, variances, covariances, and regression coefficients) are straightforward because familiar equations define the maximum likelihood parameter estimates. With few exceptions, missing data analyses require iterative optimization algorithms, even for very simple estimation problems. The **EM algorithm** is one such procedure that is particularly important for missing data analyses. The origins of EM date back to the 1970s (Beale & Little, 1975; Dempster et al., 1977; Orchard & Woodbury, 1972), with Dempster et al. (1977) playing a key role in developing the algorithm. The early applications of EM primarily focused on estimating a mean vector and a covariance matrix with missing data, but methodologists have since extended the algorithm to address a variety of difficult complete-data estimation problems, including multilevel models, finite mixtures, and structural equation models, to name a few (Jamshidian & Bentler, 1999; Liang & Bentler, 2004; McLachlan & Krishnan, 1997; Muthén & Shedden, 1999; Raudenbush & Bryk, 2002). To keep things simple, I describe the estimation process for a mean vector and a covariance matrix, but the EM algorithm is readily suited for more complex missing data problems (e.g., structural equation models with missing data; Jamshidian & Bentler, 1999).

The EM algorithm is a two-step iterative procedure that consists of an E-step and an M-step (E and M stand for expectation and maximization, respectively). The iterative process starts with an initial estimate of the mean vector and the covariance matrix (e.g., a listwise deletion estimate of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$). The **E-step** uses the elements in the mean vector and the covariance matrix to build a set of regression equations that predict the incomplete variables from the observed variables. The purpose of the E-step is to fill in the missing values in a manner that resembles stochastic regression imputation (I use the words "fill in" loosely here, because the algorithm does not actually impute the missing values). The **M-step** subsequently applies standard complete-data formulas to the filled-in data to generate updated estimates of the mean vector and the covariance matrix. The algorithm carries the updated parameter estimates forward to the next E-step, where it builds a new set of regression equations to predict the missing values. The subsequent M-step then re-estimates the mean vector and the covariance matrix. EM repeats these two steps until the elements in $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ no longer change between consecutive M-steps, at which point the algorithm has converged on the maximum likelihood estimates. These estimates might be of substantive interest in and of themselves, or they can serve as input data for other multivariate statistical procedures (Enders, 2003; Enders & Peugh, 2004; Yuan & Bentler, 2000). It is important to reiterate that the algorithm does not impute or replace the missing values. Rather, it uses all of the available data to estimate the mean vector and the covariance matrix.

In Chapter 3, I used a hill-climbing analogy to introduce iterative optimization algorithms. In this analogy, the goal of estimation is to locate the peak of the log-likelihood function (i.e., climb to the top of a hill) where the maximum likelihood estimates are located. In an EM analysis, the initial estimates of the mean vector and the covariance matrix effectively serve as the starting coordinates for the climb, and a single iteration (i.e., one E- and one M-step) represents a step toward the top of the hill. Numerically, the goal of each iteration is to adjust the parameter values in a direction that increases the log-likelihood value (i.e., the algorithm should climb in a vertical direction). The regression-based procedure at each E-step does just that, and the updated parameter estimates at each M-step will produce a higher log-likelihood value than the estimates from the preceding M-step. As the climb nears the plateau, the adjustments to the parameter estimates are very small and the log-likelihood effectively remains the same across successive M-steps. When the difference between successive estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ falls below some very small threshold (software programs often refer to this threshold as the **convergence criterion**), the iterative process stops. At this point, the algorithm has located the peak of the log-likelihood function, and the values of the mean vector and the covariance matrix from the final M-step serve as the maximum likelihood estimates.

4.10 A DETAILED DESCRIPTION OF THE EM ALGORITHM

The previous description of EM is conceptual in nature and omits many of the mathematical details of the procedure. This section expands the previous ideas and gives a more precise explanation of the E-step and the M-step. To illustrate the mechanics of EM, I use a bivariate analysis example where one of the variables is incomplete. Throughout this section, I use *X* to denote the complete variable (e.g., IQ scores) and *Y* to represent the incomplete variable (e.g., job performance ratings). This is a relatively simple estimation problem, but the basic ideas readily extend to multivariate analyses with general patterns of missing data.

With complete data, the following formulas generate the maximum likelihood estimates of the mean, the variance, and the covariance.

$$\hat{\mu}_Y = \frac{1}{N}\sum Y \tag{4.9}$$

$$\hat{\sigma}_Y^2 = \frac{1}{N}\left(\sum Y^2 - \frac{(\sum Y)^2}{N}\right) \tag{4.10}$$

$$\hat{\sigma}_{X,Y} = \frac{1}{N}\left(\sum XY - \frac{\sum X \sum Y}{N}\right) \tag{4.11}$$

Notice that the sum of the scores (i.e., $\sum X$ and $\sum Y$), the sum of the squared scores (i.e., $\sum X^2$ and $\sum Y^2$), and the sum of the cross product terms (i.e., $\sum XY$) are the basic building blocks of the previous equations. Collectively, these quantities are known as **sufficient statistics** because they contain all of the necessary information to estimate the mean vector and the covariance matrix. As you will see, these sufficient statistics play an important role in the E-step.

The purpose of the E-step is to "fill in" the missing values so that the M-step can use Equations 4.9 through 4.11 to generate parameter estimates. More accurately, the E-step fills in each case's contribution to the sufficient statistics (Dempster et al., 1977). The E-step uses the elements in the mean vector and the covariance matrix to build a set of regression equations that predict the incomplete variables from the observed variables. In a bivariate data set with missing value on *Y*, the necessary equations are

$$\hat{\beta}_1 = \frac{\hat{\sigma}_{X,Y}}{\hat{\sigma}_X^2} \tag{4.12}$$

$$\hat{\beta}_0 = \hat{\mu}_Y - \hat{\beta}_1\hat{\mu}_X \tag{4.13}$$

$$\hat{\sigma}^2_{Y|X} = \hat{\sigma}^2_Y - \hat{\beta}^2_1\hat{\sigma}^2_X \tag{4.14}$$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \tag{4.15}$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the intercept and slope coefficients, respectively, $\hat{\sigma}^2_{Y|X}$ is the residual variance from the regression of Y on X, and $\hat{Y}_i$ is the predicted Y score for a given value of X. The means, variances, and covariances that appear on the right side of the equations are elements from the mean vector and the covariance matrix.

The missing data complicate an otherwise straightforward analysis because the incomplete cases have nothing to contribute to $\sum Y$, $\sum Y^2$, and $\sum XY$. The E-step replaces the missing components of these sufficient statistics with their expected values (i.e., long-run averages). EM borrows information from other variables, so the algorithm actually uses so-called **conditional expectations** to replace the missing components of the formulas. To illustrate, consider the sum of the scores and the sum of the cross product terms (i.e., $\sum Y$ and $\sum XY$, respectively). The expected value of Y is the predicted score from Equation 4.15, so the E-step replaces the missing components of $\sum Y$ and $\sum XY$ with $\hat{Y}_i$. Next, consider the sum of the squared scores, $\sum Y^2$. The expected value of a squared variable is $\hat{Y}_i^2 + \hat{\sigma}^2_{Y|X}$, where $\hat{Y}_i^2$ is the squared predicted score, and $\hat{\sigma}^2_{Y|X}$ is the residual variance from the regression of Y on X. The E-step replaces the missing components of $\sum Y^2$ with this expectation.

Notice that the E-step does not actually impute the raw data. Rather, it fills in the computational building blocks for the mean, the variance, and the covariance (i.e., the sufficient statistics). Once this process is complete, the M-step becomes a straightforward estimation problem that uses the filled-in sufficient statistics to compute Equations 4.9 through 4.11. The resulting parameter estimates carry forward to the next E-step, where the process begins anew.

4.11 A BIVARIATE ANALYSIS EXAMPLE

Having outlined the necessary mathematical details, I use the IQ and job performance scores in Table 4.1 to illustrate a worked analysis example. Software programs that implement the EM algorithm fully automate the estimation procedure, so there is no need to perform the computational steps manually. Nevertheless, examining what happens at each step of the process is instructive and gives some insight into the inner workings of the algorithm.

EM requires an initial estimate of the mean vector and the covariance matrix. A number of traditional missing data techniques can generate these starting values, including deletion methods and single imputation (Little & Rubin, 2002, p. 225). To be consistent with statistical software packages (e.g., the SAS MI procedure), I use pairwise deletion estimates of the means and the variances and set the covariance to zero, as follows:

$$\hat{\boldsymbol{\mu}}_0 = \begin{bmatrix}\hat{\mu}_{IQ}\\ \hat{\mu}_{JP}\end{bmatrix} = \begin{bmatrix}\hat{\mu}_X\\ \hat{\mu}_Y\end{bmatrix} = \begin{bmatrix}100.000\\ 11.700\end{bmatrix}$$

$$\hat{\boldsymbol{\Sigma}}_0 = \begin{bmatrix}\hat{\sigma}^2_{IQ} & \hat{\sigma}_{IQ,JP}\\ \hat{\sigma}_{JP,IQ} & \hat{\sigma}^2_{JP}\end{bmatrix} = \begin{bmatrix}\hat{\sigma}^2_X & \hat{\sigma}_{X,Y}\\ \hat{\sigma}_{Y,X} & \hat{\sigma}^2_Y\end{bmatrix} = \begin{bmatrix}199.579 & 0\\ 0 & 7.344\end{bmatrix}$$

Throughout the example, I use a numeric subscript to index each EM cycle, and a value of zero denotes the fact that these parameter values precede the first E-step. Finally, to maintain consistency with the previous notation, I use X and Y to denote the IQ and job performance scores, respectively.

The first E-step uses the elements in the mean vector and the covariance matrix to build a regression equation that predicts the incomplete variable (e.g., job performance) from the complete variable (e.g., IQ). Substituting the appropriate elements from $\hat{\boldsymbol{\mu}}_0$ and $\hat{\boldsymbol{\Sigma}}_0$ into Equations 4.12 through 4.14 yields the following estimates: $\hat{\beta}_0 = 11.700$, $\hat{\beta}_1 = 0$, and $\hat{\sigma}^2_{Y|X} = 7.344$. Because the regression slope is zero, all of the predicted values happen to be the same, $\hat{Y}_1 = 11.700$. The ultimate goal of the E-step is to fill in the missing components of $\sum Y$, $\sum Y^2$, and $\sum XY$. Specifically, the predicted values fill in the missing components of $\sum Y$ and $\sum XY$, and $\hat{Y}_i^2 + \hat{\sigma}^2_{Y|X} = 11.700^2 + 7.344 = 144.234$ replaces the missing parts of $\sum Y^2$. Table 4.8 shows the computations for the first E-step, and the resulting sufficient statistics appear in the bottom row of the table.

Having dealt with the missing values in the E-step, the M-step uses standard complete-data formulas to update the mean vector and the covariance matrix. Substituting the sufficient statistics from Table 4.8 into Equations 4.9 through 4.11 updates the parameter estimates, as follows.

$$\hat{\boldsymbol{\mu}}_1 = \begin{bmatrix} \hat{\mu}_{IQ} \\ \hat{\mu}_{JP} \end{bmatrix} = \begin{bmatrix} \hat{\mu}_X \\ \hat{\mu}_Y \end{bmatrix} = \begin{bmatrix} 100.000 \\ 11.700 \end{bmatrix}$$

$$\hat{\boldsymbol{\Sigma}}_1 = \begin{bmatrix} \hat{\sigma}^2_{IQ} & \hat{\sigma}_{IQ,JP} \\ \hat{\sigma}_{JP,IQ} & \hat{\sigma}^2_{JP} \end{bmatrix} = \begin{bmatrix} \hat{\sigma}^2_X & \hat{\sigma}_{X,Y} \\ \hat{\sigma}_{Y,X} & \hat{\sigma}^2_Y \end{bmatrix} = \begin{bmatrix} 189.600 & 5.200 \\ 5.200 & 6.977 \end{bmatrix}$$

Notice that the job performance mean did not change, even though this variable has missing values. Because the initial regression slope is zero, the intercept (i.e., the mean job performance rating) replaces the missing Y values. Consequently, the mean does not change in the first step, although it will in subsequent steps. In addition, notice that the IQ variance changed, even though this variable is complete. This change occurred because the maximum likelihood estimate uses N rather than $N - 1$ in the denominator (the usual formula for the sample variance generated the initial estimate).

With the first cycle completed, the updated parameter estimates carry forward to the next E-step, where EM builds a new regression equation. Substituting the appropriate elements from $\hat{\boldsymbol{\mu}}_1$ and $\hat{\boldsymbol{\Sigma}}_1$ into Equations 4.12 through 4.14 yields the following estimates: $\hat{\beta}_0 = 8.957$, $\hat{\beta}_1 = 0.027$, and $\hat{\sigma}^2_{Y|X} = 6.834$. Consistent with the previous E-step, expected values replace the missing components of the sufficient statistics. For example, the individual with an IQ score of 78 contributes a predicted job performance rating of $\hat{Y}_i = 8.975 + 0.027(78) = 11.063$ to the computation of $\sum Y$ and $\sum XY$. Similarly, this case's contribution to $\sum Y^2$ is $11.063^2 + 6.834 = 129.224$. Table 4.9 shows the computations for the second E-step, with the sufficient statistics in the bottom row of the table.

As before, the M-step uses the sufficient statistics from the preceding E-step to update the mean vector and the covariance matrix. The sufficient statistics in Table 4.9 produce the following estimates.

TABLE 4.8. Computation of the Sufficient Statistics for the First E-Step

X	X^2	Y	Y^2	XY
78	6084	11.700	$11.700^2 + 7.344$	912.600
84	7056	11.700	$11.700^2 + 7.344$	982.800
84	7056	11.700	$11.700^2 + 7.344$	982.800
85	7225	11.700	$11.700^2 + 7.344$	994.500
87	7569	11.700	$11.700^2 + 7.344$	1017.900
91	8281	11.700	$11.700^2 + 7.344$	1064.700
92	8464	11.700	$11.700^2 + 7.344$	1076.400
94	8836	11.700	$11.700^2 + 7.344$	1099.800
94	8836	11.700	$11.700^2 + 7.344$	1099.800
96	9216	11.700	$11.700^2 + 7.344$	1123.200
99	9801	7	49	693
105	11025	10	100	1050
105	11025	11	121	1155
106	11236	15	225	1590
108	11664	10	100	1080
112	12544	10	100	1120
113	12769	12	144	1356
115	13225	14	196	1610
118	13924	16	256	1888
134	17956	12	144	1608
$\sum X =$ 2000.000	$\sum X^2 =$ 203792.000	$\sum Y =$ 234.000	$\sum Y^2 =$ 2877.340	$\sum XY =$ 23504.500

Note. X = IQ and Y = job performance. Bold typeface denotes imputed values.

$$\hat{\boldsymbol{\mu}}_2 = \begin{bmatrix} \hat{\mu}_{IQ} \\ \hat{\mu}_{JP} \end{bmatrix} = \begin{bmatrix} \hat{\mu}_{X} \\ \hat{\mu}_{Y} \end{bmatrix} = \begin{bmatrix} 100.000 \\ 11.523 \end{bmatrix}$$

$$\hat{\boldsymbol{\Sigma}}_2 = \begin{bmatrix} \hat{\sigma}^2_{IQ} & \hat{\sigma}_{IQ,JP} \\ \hat{\sigma}_{JP,IQ} & \hat{\sigma}^2_{JP} \end{bmatrix} = \begin{bmatrix} \hat{\sigma}^2_{X} & \hat{\sigma}_{X,Y} \\ \hat{\sigma}_{Y,X} & \hat{\sigma}^2_{Y} \end{bmatrix} = \begin{bmatrix} 189.600 & 7.663 \\ 7.663 & 6.764 \end{bmatrix}$$

Notice that the IQ mean and variance do not change because these parameters immediately converge to the maximum likelihood estimates in the first EM cycle. However, the parameters affected by missing data do change, and they continue to do so from one M-step to the next.

As you might have guessed, $\hat{\boldsymbol{\mu}}_2$ and $\hat{\boldsymbol{\Sigma}}_2$ carry forward to the next E-step, where the algorithm generates a new set of regression estimates that fill in the missing components of the sufficient statistics. The following M-step then uses the sufficient statistics to update the parameter values. EM repeats these two steps until the elements in the mean vector and the covariance matrix no longer change (or change by a trivially small amount) between consecutive M-steps, at which point the algorithm has converged on the maximum likelihood estimates. This example requires 59 cycles to converge and yields the following parameter estimates.

$$\hat{\boldsymbol{\mu}} = \begin{bmatrix} \hat{\mu}_{IQ} \\ \hat{\mu}_{JP} \end{bmatrix} = \begin{bmatrix} 100.000 \\ 10.281 \end{bmatrix}$$

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} \hat{\sigma}^2_{IQ} & \hat{\sigma}_{IQ,JP} \\ \hat{\sigma}_{JP,IQ} & \hat{\sigma}^2_{JP} \end{bmatrix} = \begin{bmatrix} 189.600 & 23.393 \\ 23.393 & 8.206 \end{bmatrix}$$

TABLE 4.9. Computation of the Sufficient Statistics for the Second E-Step

X	X^2	Y	Y^2	XY
78	6084	11.063	$11.063^2 + 6.834$	862.914
84	7056	11.225	$11.225^2 + 6.834$	942.900
84	7056	11.225	$11.225^2 + 6.834$	942.900
85	7225	11.252	$11.252^2 + 6.834$	956.420
87	7569	11.306	$11.306^2 + 6.834$	983.622
91	8281	11.414	$11.414^2 + 6.834$	1038.674
92	8464	11.441	$11.441^2 + 6.834$	1052.572
94	8836	11.495	$11.495^2 + 6.834$	1080.530
94	8836	11.495	$11.495^2 + 6.834$	1080.530
96	9216	11.549	$11.549^2 + 6.834$	1108.704
99	9801	7	49	693
105	11025	10	100	1050
105	11025	11	121	1155
106	11236	15	225	1590
108	11664	10	100	1080
112	12544	10	100	1120
113	12769	12	144	1356
115	13225	14	196	1610
118	13924	16	256	1888
134	17956	12	144	1608
$\sum X =$ 2000.000	$\sum X^2 =$ 203792.000	$\sum Y =$ 230.465	$\sum Y^2 =$ 2790.990	$\sum XY =$ 23199.766

Note. X = IQ; Y = job performance. **Bold** typeface denotes imputed values.

The regression-based procedure that EM uses to update the parameters largely obscures the fact that the estimates are incrementally improving from one step to the next. To illustrate how EM "climbs" to the top of the log-likelihood function, I used the parameter estimates from each iteration to compute the sample log-likelihood values. (EM does not actually manipulate the log-likelihood equation, so the log-likelihood values are not an automatic by-product of the analysis.) For example, substituting the starting values (i.e., $\hat{\mu}_0$ and $\hat{\Sigma}_0$) and the observed data into Equation 4.2 yields an initial log-likelihood value of $\log L = -76.9318195$. Similarly, substituting $\hat{\mu}_1$ and $\hat{\Sigma}_1$ into Equation 4.2 gives the log-likelihood for the first EM cycle, and so on. Table 4.10 shows the log-likelihood values and the job performance parameters from selected cycles of the bivariate EM analysis. As you can see, the first few EM cycles produce the largest changes in the log-likelihood, whereas the latter steps yield much smaller changes. The same is also true for the parameter estimates. 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 log-likelihood function, each additional cycle produces a very small improvement in the log-likelihood value, and the adjustments to the parameters are so small that the estimates effectively remain the same between successive M-steps. For example, in the final three EM cycles, the changes to the job performance mean occur in the fourth decimal, and the changes to the sample log-likelihood occur past the seventh decimal. At this point, the hill climb is effectively over, and the algorithm has converged on the maximum likelihood estimates.

TABLE 4.10. Sample Log-Likelihood Values across EM Cycles

EM cycle	Log-likelihood	$\hat{\mu}_{JP}$	$\hat{\sigma}^2_{JP}$	$\hat{\sigma}_{IQ,JP}$
0	–76.9318195	11.7000000	7.3440000	0.0000000
1	–76.5939005	11.7000000	6.9772220	5.2002527
2	–76.4254785	11.5225410	6.7641355	7.6631331
3	–76.2929150	11.3944910	6.6538592	9.3296347
4	–76.1883350	11.2643060	6.6285983	10.9748205
5	–76.1059020	11.1493190	6.6552569	12.4275358
6	–76.0410225	11.0477700	6.7152964	13.7104777
7	–75.9900400	10.9580870	6.7959299	14.8434952
8	–75.9500360	10.8788850	6.8882473	15.8441088
9	–75.9186850	10.8089380	6.9860245	16.7277910
10	–75.8941405	10.7471650	7.0849410	17.5082066
...	...	...	...	...
50	–75.8064920	10.2835690	8.1993288	23.3651099
51	–75.8064915	10.2831910	8.2005058	23.3698912
52	–75.8064905	10.2828570	8.2015456	23.3741137
53	–75.8064900	10.2825610	8.2024642	23.3778428
54	–75.8064895	10.2823010	8.2032757	23.3811362
55	–75.8064890	10.2820710	8.2039925	23.3840446
56	–75.8064890	10.2818670	8.2046257	23.3866132
57	–75.8064885	10.2816880	8.2051849	23.3888817
58	–75.8064885	10.2815290	8.2056789	23.3908850
59	–75.8064885	10.2813890	8.2061153	23.3926542

As an aside, the EM differs from other optimization algorithms (e.g., the scoring algorithm; Hartley & Hocking, 1971; Trawinski & Bargmann, 1964) because it does not require the computation of first or second derivatives. Consequently, the EM algorithm does not automatically produce the basic building blocks of maximum likelihood standard errors. Methodologists have outlined approaches for generating standard errors in an EM analysis (Little & Rubin, 2002; Meng & Rubin, 1991), but these methods require additional computational steps that are not implemented in all software packages. Bootstrap resampling is a simulation-based approach that is particularly useful for estimating standard errors with nonnormal data, but it is also applicable to an EM analysis. I give a detailed description of bootstrap in Chapter 5.

4.12 EXTENDING EM TO MULTIVARIATE DATA

The preceding bivariate analysis is relatively straightforward because the missing values are isolated to a single variable. Applying EM to multivariate data is typically more complex because the E-step requires a unique regression equation (or set of equations) for each missing data pattern. Despite this complication, the basic logic of EM remains the same and requires just a few additional details. To illustrate the changes to the E-step, I use the full data set in Table 4.1. EM with three variables is still relatively straightforward, but the logic of this example generalizes to data sets with any number of variables. Finally, note that the procedural details of the M-step do not change because this step always uses the standard complete-data formulas in Equations 4.9 through 4.11 to update the parameter estimates. Consequently, the following discussion focuses solely on the E-step. To maintain consistent notation, X denotes the IQ scores, Y represents the job performance ratings, and Z corresponds to the well-being scores.

Applying the E-step to the data in Table 4.1 requires the following set of sufficient statistics: $\sum X$, $\sum X^2$, $\sum Y$, $\sum Y^2$, $\sum Z$, $\sum Z^2$, $\sum XY$, $\sum XZ$, and $\sum YZ$. Notice that these quantities are the same as those from the previous bivariate example (i.e., the sum of the scores, the sum of the squared scores, and the sum of the cross product terms). As before, the purpose of the E-step is to replace the missing components of the sufficient statistics with expectations, but this now requires a unique set of regression equations for missing data pattern. Returning to the data in Table 4.1, note that there are four missing data patterns: (1) cases with only IQ scores, (2) cases with IQ and well-being scores, (3) cases with IQ and job performance scores, and (4) cases with complete data on all three variables. The complete cases are not a concern, so the E-step only deals with the three patterns that have missing data. Table 4.11 shows the missing sufficient statistics and the relevant expectation terms for each missing data pattern.

Consider the subsample of cases with missing job performance ratings (i.e., missing Y values). These individuals have complete data on the IQ and psychological well-being variables (i.e., X and Z, respectively), so the problematic sufficient statistics are $\sum Y$, $\sum XY$, $\sum YZ$, and $\sum Y^2$. Following the logic from the bivariate example, predicted scores replace the missing components of the variable sums and sums of products. This missing data pattern has two complete variables, so a multiple regression equation generates the predicted scores, as follows:

$$\hat{Y}_{i|X,Z} = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 Z_i \qquad (4.16)$$

where $\hat{Y}_{i|X,Z}$ is the predicted Y value for case i (the vertical bar denotes the fact that the predicted score is conditional on both X and Z). Consistent with the bivariate analysis, the expectation for $\sum Y^2$ involves a squared predicted score and a residual variance estimate.

TABLE 4.11. Expectations for a Multivariate Application of the EM Algorithm

Missing variable	Missing sufficient statistics	Imputed expectations
Y (job performance)	$\sum Y, \sum XY, \sum YZ$ $\sum Y^2$	$\hat{Y}_{i\|X,Z}$ $\hat{Y}^2_{i\|X,Z} + \hat{\sigma}^2_{Y\|X,Z}$
Z (well-being)	$\sum Z, \sum XZ, \sum YZ$ $\sum Z^2$	$\hat{Z}_{i\|X,Y}$ $\hat{Z}^2_{i\|X,Y} + \hat{\sigma}^2_{Z\|X,Y}$
Y and Z (job performance and well-being)	$\sum Y, \sum XY, \sum YZ$ $\sum Y^2$ $\sum Z, \sum XY, \sum YZ$ $\sum Z^2$ $\sum YZ$	$\hat{Y}_{i\|X}$ $\hat{Y}^2_{i\|X} + \hat{\sigma}^2_{Y\|X}$ $\hat{Z}_{i\|X}$ $\hat{Z}^2_{i\|X} + \hat{\sigma}^2_{Z\|X}$ $(\hat{Y}_{i\|X})(\hat{Z}_{i\|X}) + \hat{\sigma}_{Y,Z\|X}$

Consequently, $\hat{Y}^2_{i|X,Z} + \hat{\sigma}^2_{Y|X,Z}$ replaces the missing components of $\sum Y^2$, where $\hat{\sigma}^2_{Y|X,Z}$ is the residual variance from the regression of Y on X and Z.

Next, consider the individual with a missing well-being score (i.e., missing Z value). Again, a multiple regression equation generates the predicted score

$$\hat{Z}_{i|X,Y} = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 Y_i, \tag{4.17}$$

and this predicted value replaces the missing components of $\sum Z$, $\sum XZ$, and $\sum YZ$. As seen in Table 4.11, the expectation for the missing Z^2 value is similar to the previous missing data pattern and equals the squared predicted score plus the residual variance from the regression of Z on X and Y.

Thus far, the E-step has not changed very much. Each missing data pattern requires a unique set of regression equations and expectations, but the underlying logic is the same as it was in the bivariate example. The only additional nuance occurs with patterns that have two or more missing variables. For example, consider the subsample of cases with missing job performance ratings and well-being scores (i.e., Y and Z, respectively). As before, regression equations generate predicted scores for each missing variable, as follows:

$$\hat{Y}_{i|X} = \hat{\beta}_0 + \hat{\beta}_1 X_i \tag{4.18}$$

$$\hat{Z}_{i|X} = \hat{\beta}_2 + \hat{\beta}_3 X_i \tag{4.19}$$

As seen in Table 4.11, the predicted scores and corresponding residual variances fill in all but one of the sufficient statistics. The cross product term for the two missing variables (i.e., $\sum YZ$) involves a new expectation, $(\hat{Y}_{i|X})(\hat{Z}_{i|X}) + \hat{\sigma}_{Y,Z|X}$, where the terms in parentheses are the predicted values from previous regression equations, and $\sigma_{Y,Z|X}$ is the residual covariance between job performance and well-being, which is $\hat{\sigma}_{Y,Z|X} = \hat{\sigma}_{Y,Z} - \hat{\beta}_1\hat{\beta}_3\hat{\sigma}^2_X$.

Extending the E-step computations to complex multivariate analyses with general missing data patterns is straightforward because the relevant expectations are identical to those in Table 4.11. The main procedural difficulty is the computation of regression equations for each missing data pattern. Not surprisingly, the number of missing data patterns (and thus the number of regression equations) can get quite large as the number of variables increases. Although it sounds tedious to construct a set of regressions for each missing data pattern, a computational algorithm called the sweep operator can automate this process. The sweep operator combines the mean vector and the covariance matrix into a single augmented matrix and applies a series of transformations that produce the desired regression coefficients and residual variances. A number of sources give detailed descriptions of the sweep operator (Dempster, 1969; Goodnight, 1979; Little & Rubin, 2002).

4.13 MAXIMUM LIKELIHOOD ESTIMATION SOFTWARE OPTIONS

Although the mathematical foundations of maximum likelihood missing data handling have been in the literature for many years, estimation routines have only recently become widely available in statistical software packages. In the late 1980s, methodologists outlined techniques that effectively tricked complete-data software packages into implementing maximum likelihood missing data handling by treating each missing data pattern as a subpopulation in a multiple group structural equation model (Allison, 1987; Muthén, Kaplan, and Hollis, 1987). However, these approaches did not enjoy widespread use because they were complicated to program and were unwieldy to implement with more than a small handful of missing data patterns. Fortunately, this approach is no longer necessary.

Many of the recent software innovations have occurred within the latent variable modeling framework, and virtually every structural equation modeling software package now implements maximum likelihood missing data handling. (This approach is often referred to as full information maximum likelihood estimation, or simply FIML.) The latent variable modeling framework encompasses a vast number of analytic methods that researchers use on a routine basis (e.g., correlation, regression, factor analysis, path analysis, structural equation models, mixture models, multilevel models). Structural equation modeling software is therefore an ideal tool for many missing data problems. Structural equation modeling programs have undergone dramatic improvements in the number of and type of missing data analyses that they are capable of performing, and these packages continue to evolve at a rapid pace. Because of their flexibility and breadth, I rely heavily on structural equation programs to generate the analysis examples throughout the book. I discuss the capabilities of specific packages in more detail in Chapter 11.

As an aside, a word of caution is warranted concerning software programs that implement the EM algorithm. Some popular packages (e.g., LISREL and SPSS) offer the option of imputing the raw data after the final EM cycle. This is somewhat unfortunate because it gives the impression that a maximum likelihood approach has properly handled the missing values. In reality, this imputation scheme is nothing more than regression imputation. The only difference between EM imputation and regression imputation is that the EM approach uses a maximum likelihood estimate of the mean vector and the covariance matrix to generate the regression equations, whereas standard regression imputation schemes tend to use listwise deletion estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ to build the regressions. Although it may sound appealing to base the imputation process on maximum likelihood estimates, doing so leads to the same negative outcomes described in Chapter 2, namely, biased parameter estimates and attenuated standard errors (von Hippel, 2004). Consequently, it is a good idea to avoid EM-based single imputation routines. In situations that necessitate a filled-in data set, multiple imputation is a much better option.

4.14 DATA ANALYSIS EXAMPLE 1

This section describes a data analysis that uses the EM algorithm to generate maximum likelihood estimates of a mean vector, covariance matrix, and correlation matrix.* The data for this analysis consist of scores from 480 employees on eight work-related variables: gender, age, job tenure, IQ, psychological well-being, job satisfaction, job performance, and turnover

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

TABLE 4.12. Mean, Covariance, and Correlation Estimates from Data Analysis Example 1

Variable	1	2	3	4	5	6	7	8
			Missing data maximum likelihood					
1: Age	28.908	0.504	–0.010	0.182	0.136	–0.049	–0.150	0.015
2: Tenure	8.459	9.735	–0.034	0.155	0.154	0.016	0.011	0.001
3: Female	–0.028	–0.052	0.248	0.115	0.047	–0.015	0.005	0.068
4: Well-being	1.148	0.569	0.067	1.382	0.322	0.456	–0.257	0.291
5: Satisfaction	0.861	0.565	0.028	0.446	1.386	0.184	–0.234	0.411
6: Performance	–0.330	0.061	–0.009	0.671	0.271	1.570	–0.346	0.426
7: Turnover	–0.377	0.016	0.001	–0.141	–0.129	–0.203	0.218	–0.180
8: IQ	0.674	0.026	0.284	2.876	4.074	4.496	–0.706	70.892
Means	37.948	10.054	0.542	6.288	5.950	6.021	0.321	100.102
			Complete data 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

Note. Correlations are shown in the upper diagonal in bold typeface. Elements affected by missing data are enclosed in the shaded box.

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). The data have three missing data patterns, each of which contains one-third of the sample. The first pattern consists of cases with complete data, and the remaining two patterns have missing data on either well-being or job satisfaction. These patterns mimic a situation in which the data are missing by design (e.g., to reduce the cost of data collection).

Table 4.12 shows the maximum likelihood estimates, along with the corresponding estimates from the complete data. To facilitate comparison, a shaded box encloses the parameter estimates affected by the missing data. As seen in the table, the missing data estimates are quite similar to those of the complete data. For example, the two sets of correlation values typically differ by approximately .02, and the largest difference is .04 (the correlation between well-being and turnover intentions). The similarity of the two sets of estimates might seem somewhat remarkable given that 33% of the satisfaction and well-being scores are missing.

4.15 DATA ANALYSIS EXAMPLE 2

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

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

The top panel of Figure 4.1 shows the path diagram of the regression model. I used a structural equation modeling program to estimate the regression model because these packages offer a convenient platform for implementing maximum likelihood estimation with missing data.* Finally, note that I requested standard errors based on the observed information matrix.

Researchers typically begin a regression analysis by examining the omnibus F test. The likelihood ratio statistic and the multivariate Wald test are analogous procedures in a maximum likelihood analysis. The procedural details of both tests are identical with or without missing data. Recall from Chapter 3 that the likelihood ratio test involves a pair of nested models. The full model corresponds to the multiple regression in the top panel of Figure 4.1, 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 $\log L_{\text{Full}} = -1753.093$ and $\log L_{\text{Restricted}} = -1793.181$, 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. Using the log-likelihood values to compute the likelihood ratio test (see Equation 3.16) yields $LR = 80.18$. The models differ by two parameters (i.e., the restricted model constrains two coefficients two zero), so 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 an F statistic, this suggests that at least one of the regression coefficients is significantly different from zero.

Researchers typically follow up a significant omnibus test by examining partial regression coefficients. Table 4.13 gives the regression model estimates along with those of the corresponding complete-data analysis from Chapter 3. As seen in the table, psychological well-being was a significant predictor of job performance, $\hat{\beta}_1 = 0.476$, $z = 8.66$, $p < .001$, but job satisfaction was not, $\hat{\beta}_2 = 0.027$, $z = 0.45$, $p = 0.66$. Notice that the missing data estimates are quite similar to those of the complete data, despite the fact that each predictor variable has a missing data rate of 33%. The missing data analysis produced somewhat larger standard errors, but this is to be expected. Finally, note that the interpretation of the regression coefficients is the same as it is in a complete-data regression analysis. For example, holding job satisfaction constant, a one-point increase in psychological well-being yields a .476 increase in job satisfaction, on average.

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

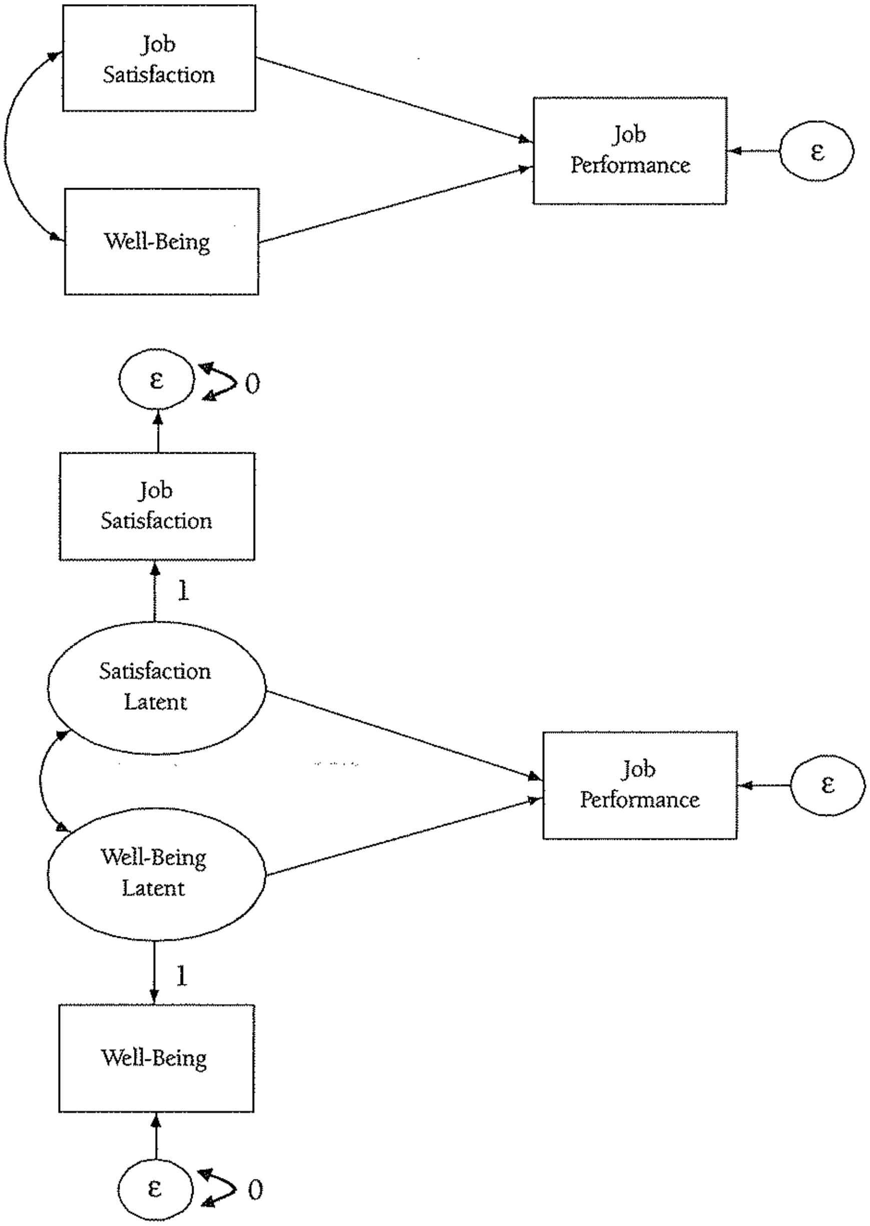

FIGURE 4.1. A path diagram for the multiple regression model. The single-headed straight lines represent regression coefficients, the double-headed curved arrow is a correlation, the rectangles are manifest variables, and the ellipse is a latent variable. The top panel of the figure shows the manifest variable regression model. The bottom panel of the figure shows the regression model recast as a latent variable model, where the two latent variables have a single manifest indicator. The factor loadings are fixed at values of one, and the residual variances (the doubled-headed curved arrows the manifest variable residual terms to themselves) are constrained to zero.

A Note on Missing Explanatory Variables

Before proceeding to the next analysis example, it is important to note that software packages are not uniform in their treatment of missing explanatory variables. Specifically, some software programs exclude cases that have incomplete data on explanatory variables, while others do not. To understand why this is the case, reconsider the log-likelihood formula in Equation 4.2.

TABLE 4.13. Regression Model Estimates from Data Analysis Example 2

Parameter	Estimate	*SE*	*z*
	Missing data maximum likelihood		
β_0 (intercept)	6.021	0.053	113.123
β_1 (well-being)	0.476	0.055	8.664
β_2 (satisfaction)	0.027	0.060	0.445
$\hat{\sigma}^2_e$ (residual)	1.243	0.087	14.356
R^2	0.208		
	Complete data 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
$\hat{\sigma}^2_e$ (residual)	1.256	0.081	15.492
R^2	0.200		

Note. Predictors were centered at the maximum likelihood estimates of the mean.

The log-likelihood quantifies the standardized distance between the outcome variables (i.e., the Y vector) and the population mean. Depending on the software package and the underlying statistical model, the explanatory variables may not be included in the score vector. For example, in a regression analysis, some software platforms specify the explanatory variables as part of the population mean vector, such that $\beta_0 + \beta_1(X_1) + \beta_2(X_2)$ replaces the $\boldsymbol{\mu}$ term in Equation 4.2. In these situations, the software program is likely to exclude the cases with the missing explanatory variables. To further complicate matters, a given software program may not be consistent in its treatment of missing explanatory variables across different analyses. For example, a package might include the incomplete cases in a regression model but exclude those data records in more complex models.

Structural equation modeling programs incorporate some flexibility for dealing with incomplete explanatory variables. Specifically, recasting an incomplete predictor variable as the sole manifest indicator of a latent variable effectively tricks the software program into treating the explanatory variable as an outcome, while still maintaining the variable's exogenous status in the model. For example, the bottom panel of Figure 4.1 shows the previous regression analysis as a latent variable model. In the latent variable specification, the factor loadings are constrained to one (this equates the latent variable's metric to the manifest variable's metric) and the residual variances are constrained to zero (this equates the latent variable's variance to the manifest variable's variance). Because the latent variables predict the explanatory variables, the incomplete predictors become part of the Y vector in Equation 4.2. Importantly, this programming trick does not change the interpretation of the model parameters (e.g., the arrow that connects the latent job satisfaction variable to the job performance variable is still a partial regression coefficient). Readers interested in more details on single-indicator latent variables can consult any number of structural equation modeling textbooks (e.g., see Kline, 2005, pp. 229–231).

As an important aside, the single-indicator latent variable approach can have a bearing on the likelihood ratio test. As an illustration, reconsider the likelihood ratio test from the multiple regression analysis. I specified the restricted model by constraining both regression slopes to zero during estimation. Had the data been complete, I could have specified an equivalent restricted model by simply excluding the explanatory variables from the analysis. However, this approach would not produce a nested model if the manifest explanatory variables are both indicators of a latent variable, because the two models will have different sets of variables that contribute to the Y vector in the log-likelihood equation. Consequently, the only correct way to specify a nested model is to constrain parameters from the full model to zero. Returning to the latent variable model in the bottom panel of Figure 4.1, note that constraining the arrows that connect the latent variables to the job performance variable to zero during estimation produces an appropriate nested model, whereas excluding job satisfaction and well-being from the model does not.

4.16 DATA ANALYSIS EXAMPLE 3

The third analysis example uses maximum likelihood to estimate a multiple regression model with an interaction term. The analysis uses the employee data set from the previous examples and involves the regression of job performance on well-being, gender, and the interaction between well-being and gender. The goal of the analysis is to determine whether gender moderates the association between psychological well-being and job performance. The multiple regression equation is as follows:

$$JP_i = \beta_0 + \beta_1(WB_i) + \beta_2(FEMALE_i) + \beta_3(WB_i)(FEMALE_i) + \varepsilon$$

and Figure 4.2 shows the corresponding path diagram of the model. Notice that the interaction term (i.e., the product of gender and well-being) simply serves as an additional explanatory variable in the model. Using maximum likelihood to estimate a model with an interaction term is straightforward and follows the same procedure as any multiple regression analysis. I include this example as a point of contrast with multiple imputation. As you will see in Chapter 9, multiple imputation requires special procedures to deal with interactive effects such as this. Consistent with the previous analysis example, I used structural equation software to estimate the regression model and requested standard errors based on the observed information matrix.*

Prior to conducting the analysis, I centered the psychological well-being scores at the maximum likelihood estimate of the grand mean from Table 4.12. Next, I created a product term by multiplying gender (0 = male, 1 = female) and the centered well-being scores. The resulting product term is missing for any case with a missing well-being score. Because males have a gender code of zero, their product terms should always equal zero, regardless of whether the well-being variable is complete. Consequently, I recoded the missing product terms to have a value of zero within the male subsample.

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

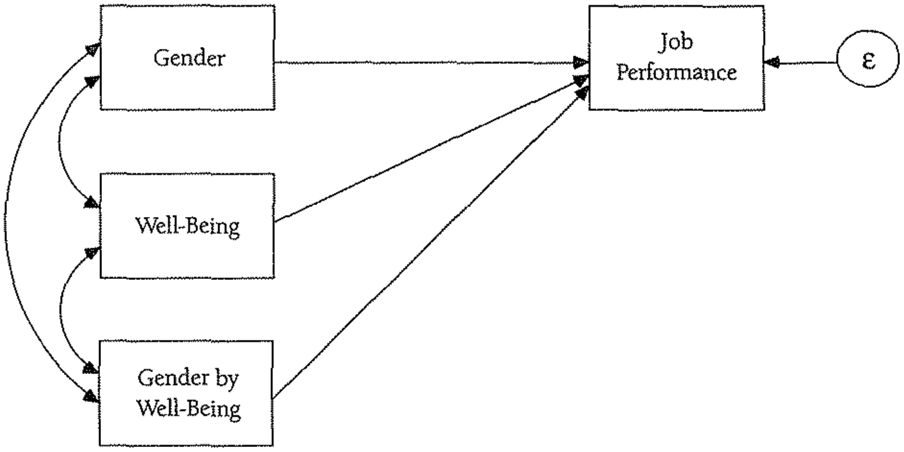

FIGURE 4.2. A path diagram for the multiple regression model. The single-headed straight lines represent regression coefficients, the double-headed curved arrow is a correlation, the rectangles are manifest variables, and the ellipse is a latent variable.

Because the previous analysis illustrates the use of the likelihood ratio test, there is no need to demonstrate the procedure further. Table 4.14 gives the regression model estimates along with those of the corresponding complete-data analysis. The analysis results suggest that males and females do not differ with respect to their mean job performance ratings, $\hat{\beta}_2 = -0.167$, $z = -1.59$, $p = .11$, but the significant interaction term indicates that the association between well-being and performance is different for males and females, $\hat{\beta}_3 = 0.362$, $z = 3.43$, $p < .001$. Because the gender variable is coded such that female = 1 and male = 0, the sign of the interaction coefficient indicates that the relationship is stronger for females. Notice that the interpretation of the regression coefficients is identical to what it would have been had the data been complete. In addition, the computation of simple slopes is identical to that of a complete-data analysis. For example, the regression equation for the subsample of males (the group coded 0) is $\hat{Y}_M = \hat{\beta}_0 + \hat{\beta}_1(WB)$, and the corresponding equation for females (the group coded 1) is $\hat{Y}_F = (\hat{\beta}_0 + \hat{\beta}_2) + (\hat{\beta}_1 + \hat{\beta}_3)(WB)$. Finally, notice that the missing data estimates are quite similar to those of the complete data, but they have larger standard errors. The increase in the standard errors is not surprising given that the well-being variable and the interaction term have a substantial proportion of missing values.

4.17 DATA ANALYSIS EXAMPLE 4

This section presents a data analysis example that illustrates how to use an EM covariance matrix to conduct an exploratory factor analysis and an internal consistency reliability analysis.* The analyses use artificial data from a questionnaire on eating disorder risk. Briefly, the data contain the responses from 400 college-aged women on 10 questions from the Eating Attitudes Test (EAT; Garner, Olmsted, Bohr, & Garfinkel, 1982), a widely used measure of eating disorder risk. The 10 questions measure two constructs: Drive for Thinness (e.g., "I avoid

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

TABLE 4.14. Regression Model Estimates from Data Analysis Example 3

Parameter	Estimate	*SE*	*z*
	Missing data maximum likelihood		
β_0 (intercept)	6.091	0.076	79.755
β_1 (well-being)	0.337	0.071	4.723
β_2 (gender)	–0.167	0.105	–1.587
β_3 (interaction)	0.362	0.106	3.426
$\hat{\sigma}^2_\varepsilon$ (residual)	1.234	0.084	14.650
R^2	.214		
	Complete data maximum likelihood		
β_0 (intercept)	6.080	0.075	81.536
β_1 (well-being)	0.304	0.057	5.339
β_2 (gender)	–0.146	0.101	–1.438
β_3 (interaction)	0.326	0.082	3.975
$\hat{\sigma}^2_\varepsilon$ (residual)	1.211	0.078	15.492
R^2	.229		

Note. Predictors were centered at the maximum likelihood estimates of the mean.

eating when I'm hungry") and Food Preoccupation (e.g., "I find myself preoccupied with food"), and they mimic the two-factor structure proposed by Doninger, Enders, and Burnett (2005). Figure 4.3 shows a graphic of the EAT factor structure and abbreviated descriptions of the item stems. The data set also contains an anxiety scale score, a variable that measures beliefs about Western standards of beauty (e.g., high scores indicate that respondents internalize a thin ideal of beauty), and body mass index (BMI) values.

Variables in the EAT data set are missing for a variety of reasons. I simulated MCAR data by randomly deleting scores from the anxiety variable, the Western standards of beauty scale, and two of the EAT questions (EAT_2 and EAT_{21}). It seems reasonable to expect a relationship between body weight and missingness, so I created MAR data on five variables (EAT_1, EAT_{10}, EAT_{12}, EAT_{18}, and EAT_{24}) by deleting the EAT scores for a subset of cases in both tails of the BMI distribution. These same EAT questions were also missing for individuals with elevated anxiety scores. Finally, I introduced a small amount of MNAR data by deleting a number of the high body mass index scores (e.g., to mimic a situation in which females with high BMI values refuse to be weighed). The deletion process typically produced a missing data rate of 5 to 10% on each variable.

Most software packages use deletion methods to handle missing data in factor analyses and reliability analyses. The same software programs can usually accommodate a covariance matrix as input data, so you can effectively implement maximum likelihood by estimating the mean vector and the covariance matrix (e.g., using the EM algorithm) and using the resulting estimates as input data for the analysis. The problem with using an EM covariance matrix as input data is that no single value of *N* is applicable to the entire matrix (Enders & Peugh, 2004). This poses a problem for standard error computations and requires corrective proce-

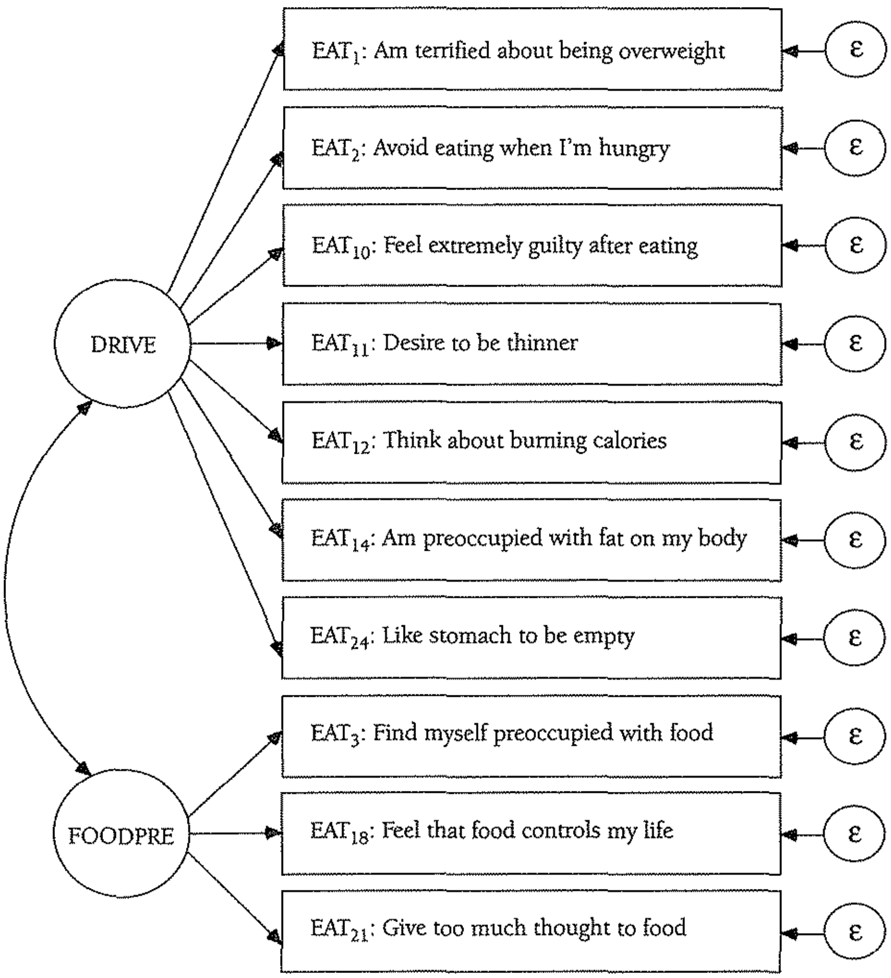
EAT$_1$: Am terrified about being overweight
EAT$_2$: Avoid eating when I'm hungry
EAT$_{10}$: Feel extremely guilty after eating
EAT$_{11}$: Desire to be thinner
EAT$_{12}$: Think about burning calories
EAT$_{14}$: Am preoccupied with fat on my body
EAT$_{24}$: Like stomach to be empty
EAT$_3$: Find myself preoccupied with food
EAT$_{18}$: Feel that food controls my life
EAT$_{21}$: Give too much thought to food
DRIVE
FOODPRE
ε

FIGURE 4.3. A path diagram for the two-factor confirmatory factor analysis model. The single-headed straight lines represent regression coefficients, the double-headed curved arrow is a correlation, the rectangles are manifest variables, and the ellipses are latent variables.

dures such as bootstrap resampling. However, software programs typically do not report standard errors for exploratory factor analyses and reliability analyses. Therefore, specifying a sample size is not a concern for the analyses in this section.

Table 4.15 shows the maximum likelihood estimates of the variable means, covariances, and correlations for the EAT questionnaire items. Although the factor analysis and the reliability analysis rely only on the 10 questionnaire items, I included all 13 variables in the initial EM analysis. Chapter 1 introduced the idea of an inclusive analysis strategy that utilizes auxiliary variables that are correlates of missingness or correlates of the analysis variables (Collins, Schafer, & Kam, 2001). The three additional variables effectively served as auxiliary variables in the initial EM analysis. Excluding these variables from the EM analysis would have been detrimental to the accuracy of the parameter estimates because body mass index and anxiety scores determine missingness. Adopting an inclusive analysis strategy is nearly always beneficial because it can improve the chances of satisfying the MAR assumption and can fine-tune the resulting parameter estimates by decreasing bias or increasing power.

I used the correlations in Table 4.15 as input data for an exploratory factor analysis. The principal axis factor analysis produced two factors with eigenvalues greater than one, which

TABLE 4.15. Mean, Covariance, and Correlation Estimates from Data Analysis Example 4

Variable	1	2	3	4	5	6	7	8	9	10
1: EAT_1	1.158	0.508	0.548	0.553	0.512	0.593	0.435	0.362	0.268	0.365
2: EAT_2	0.536	0.960	0.521	0.554	0.479	0.576	0.387	0.288	0.260	0.374
3: EAT_{10}	0.585	0.506	0.983	0.648	0.539	0.727	0.506	0.430	0.449	0.502
4: EAT_{11}	0.560	0.511	0.604	0.886	0.569	0.720	0.529	0.352	0.334	0.404
5: EAT_{12}	0.545	0.465	0.529	0.531	0.981	0.562	0.435	0.255	0.264	0.345
6: EAT_{14}	0.654	0.578	0.738	0.694	0.570	1.049	0.563	0.439	0.412	0.495
7: EAT_{24}	0.467	0.378	0.500	0.496	0.430	0.575	0.994	0.190	0.241	0.264
8: EAT_3	0.392	0.285	0.429	0.334	0.254	0.452	0.191	1.014	0.583	0.656
9: EAT_{18}	0.291	0.257	0.449	0.317	0.264	0.426	0.242	0.593	1.020	0.637
10: EAT_{21}	0.395	0.368	0.500	0.382	0.344	0.510	0.265	0.664	0.647	1.011
Means	4.010	3.940	3.950	3.940	3.930	3.960	3.990	3.970	3.980	3.950

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

suggests the presence of two underlying dimensions. I subsequently used direct oblimin rotation to examine the relationships between the factors and the questionnaire items; Table 4.16 shows the resulting pattern weights and structure coefficients. The structure coefficients are correlations between the questionnaire items and the factors, whereas the pattern weights are partial regression coefficients that quantify the influence of a factor on an item after partialling out the influence of the other factor. Both the pattern weights and structure coefficients suggest a two-factor solution, although the structure coefficients are less clear owing to the strong correlation between the factors ($r = .55$). The first factor consists of seven questions that measure a construct that the eating disorder literature refers to as Drive for Thinness, and the remaining three items form a Food Preoccupation factor. Finally, I used the EM correlations as input data for an internal consistency reliability analysis and computed the coefficient alpha for the two EAT subscales (Enders, 2003, 2004). The coefficient alpha reliability estimates for the Drive for Thinness and Food Preoccupation subscale scores are .893 and .834, respectively.

4.18 DATA ANALYSIS EXAMPLE 5

The final data analysis example illustrates a confirmatory factor analysis. I used structural equation modeling software to fit the two-factor model in Figure 4.3 to the EAT questionnaire data set.* Estimating a confirmatory factor analysis model with missing data is largely the same as it is with complete data, and software packages typically invoke maximum likelihood missing data handling with a single additional keyword or line of code. Consistent with the previous analyses, I requested standard errors based on the observed information matrix.

Researchers have traditionally used a covariance matrix as input data for structural equation modeling analyses. A complete data set simplifies the estimation process because the

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

TABLE 4.16. Factor Analysis Estimates from Data Analysis Example 4

	Pattern weights		Structure coefficients	
Variable	DT	FP	DT	FP
EAT_1	0.684	0.030	0.701	0.409
EAT_2	0.664	0.014	0.671	0.381
EAT_{10}	0.691	0.190	0.796	0.572
EAT_{11}	0.825	–0.008	0.820	0.448
EAT_{12}	0.714	–0.042	0.690	0.353
EAT_{14}	0.803	0.114	0.866	0.558
EAT_{24}	0.686	–0.093	0.634	0.286
EAT_3	–0.008	0.785	0.426	0.780
EAT_{18}	–0.010	0.758	0.410	0.753
EAT_{21}	0.061	0.807	0.508	0.841

Note. DT = drive for thinness; FP = food preoccupation.

sample log-likelihood is less complex and does not require raw data (Kaplan, 2000, pp. 25–27). The missing data log-likelihood in Equation 4.2 necessitates the use of raw data, adding a mean structure that is not usually present in standard structural equation models. The key part of the missing data log-likelihood is the collection of terms that form Mahalanobis distance, $(Y_i - \boldsymbol{\mu}_i)^T \Sigma_i^{-1}(Y_i - \boldsymbol{\mu}_i)$. A confirmatory factor analysis expresses $\boldsymbol{\mu}_i$ as a model-implied mean vector that depends on the measurement intercepts, factor loadings, and latent variable means (i.e., $\boldsymbol{\mu} = \boldsymbol{\upsilon} + \boldsymbol{\Lambda\kappa}$, where $\boldsymbol{\upsilon}$ is the vector of measurement intercepts, $\boldsymbol{\Lambda}$ is the factor loading matrix, and $\boldsymbol{\kappa}$ is the vector of latent variable means). The measurement intercepts and the latent variable means are parameter estimates that you may not be accustomed to seeing on a confirmatory factor analysis printout. These additional parameters are a technical nuance associated with the missing data handling procedure; they may or may not be of substantive interest. However, the mean structure does require its own identification constraint, and constraining the latent variable means to zero during estimation is a straightforward way to achieve model identification. Consistent with a complete-data analysis, fixing the latent factor variance to unity or setting one of the factor loadings to one identifies the covariance structure portion of the model (Bollen, 1989; Kline, 2005).

Table 4.17 shows the confirmatory factor analysis parameter estimates along with those from a corresponding complete-data analysis. The factor loadings quantify the expected change in the questionnaire items for a one-standard-deviation increase in the latent construct, and the measurement intercepts are the expected scores for a case that has a value of zero on the latent factor (i.e., is at the mean of the latent variable). Because the factor means equal zero, the measurement intercepts estimate the item means. A complete-data confirmatory factor analysis model would not ordinarily include the measurement intercepts, but I estimated these parameters for comparability.

The two-factor model fits the data well according to conventional standards (Hu & Bentler, 1998, 1999), $\chi^2(34) = 47.10$, $p = .07$, CFI = 0.993, RMSEA = 0.031, SRMR = 0.029, and all of the factor loadings are statistically significant at $p < .001$. The missing data estimates are quite similar to those of the complete data (the EAT_{18} loading is a notable exception)

TABLE 4.17. Confirmatory Factor Analysis Estimates from Data Analysis Example 5

	Loadings		Intercepts		Residuals	
Variable	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*
			Missing data maximum likelihood			
EAT$_1$	0.741	0.050	4.002	0.055	0.602	0.048
EAT$_2$	0.650	0.045	3.934	0.050	0.534	0.042
EAT$_{10}$	0.807	0.043	3.955	0.050	0.329	0.030
EAT$_{11}$	0.764	0.040	3.937	0.047	0.300	0.026
EAT$_{12}$	0.662	0.047	3.926	0.051	0.538	0.043
EAT$_{14}$	0.901	0.041	3.962	0.051	0.235	0.025
EAT$_{24}$	0.623	0.048	3.980	0.051	0.597	0.047
EAT$_3$	0.772	0.046	3.967	0.050	0.416	0.041
EAT$_{18}$	0.749	0.048	3.974	0.052	0.453	0.044
EAT$_{21}$	0.862	0.045	3.950	0.051	0.262	0.039
			Complete data maximum likelihood			
EAT$_1$	0.731	0.048	3.995	0.053	0.600	0.046
EAT$_2$	0.638	0.045	3.940	0.049	0.534	0.041
EAT$_{10}$	0.797	0.042	3.943	0.049	0.344	0.029
EAT$_{11}$	0.763	0.040	3.938	0.047	0.302	0.026
EAT$_{12}$	0.692	0.047	3.965	0.051	0.570	0.044
EAT$_{14}$	0.901	0.041	3.963	0.051	0.235	0.025
EAT$_{24}$	0.630	0.046	3.995	0.050	0.603	0.045
EAT$_3$	0.780	0.046	3.967	0.050	0.404	0.041
EAT$_{18}$	0.700	0.047	3.970	0.050	0.494	0.043
EAT$_{21}$	0.855	0.045	3.953	0.050	0.275	0.039

but have larger standard errors. It is important to point out that this analysis does not satisfy the MAR assumption because the "causes" of missing data (i.e., body mass index and anxiety) do not appear in the model. Collins et al. (2001) show that omitting a cause of missingness tends to be problematic if the correlation between the omitted variable and the analysis variables is relatively strong (e.g., $r > .40$) or if the missing data rate is greater than 25%. The body mass index and anxiety variables are not that highly correlated with the EAT questionnaire items, which probably explains why the missing data estimates are similar to those of the complete data. Chapter 5 illustrates how to incorporate correlates of missingness into a maximum likelihood analysis, and doing so would satisfy the MAR assumption for this analysis.

As a final note, the model fit statistics and the standard errors from this analysis are not entirely trustworthy because the data do not satisfy the multivariate normality assumption. (The EAT questionnaire items use a discrete Likert-type scale and are a somewhat positively skewed and kurtotic.) Methodological studies have repeatedly shown that normality violations can distort model fit statistics and standard errors, with or without missing data (Enders, 2001; Finney & DiStefano, 2006; West, Finch, & Curran, 1995). The next chapter describes corrective techniques that remedy these problems.

4.19 SUMMARY

This chapter describes how maximum likelihood estimation applies to missing data problems. The methodological literature regards maximum likelihood estimation as a state-of-the-art missing data technique because it yields unbiased parameter estimates with MAR data. From a practical standpoint, this means that maximum likelihood will produce accurate parameter estimates when traditional approaches fail. Even if the data are MCAR, maximum likelihood is still superior to traditional techniques because it maximizes statistical power by borrowing information from the observed data. Despite these desirable properties, maximum likelihood estimation is not a perfect solution and will yield biased parameter estimates when the data are MNAR. However, this bias tends to be isolated to a subset of the analysis model parameters, whereas traditional techniques are more apt to propagate bias throughout the entire model.

Maximum likelihood estimation repeatedly auditions different combinations of population parameter values until it identifies the particular constellation of values that produce the highest log-likelihood value (i.e., the best fit to the data). Conceptually, the estimation process is the same with or without missing data. However, the incomplete data records require a slight alteration to the individual log-likelihood equation. The missing data log-likelihood does not require each case to have the same number of observed data points, and the computation of the individual log-likelihood uses only the variables and parameters for which a case has complete data. Although the log-likelihood formula looks slightly different for each missing data pattern, the individual log-likelihood still quantifies the relative probability that an individual's scores originate from a multivariate normal distribution with a particular mean vector and covariance matrix. Consistent with a complete-data analysis, the sample log-likelihood is the sum of the individual log-likelihoods, and the goal of estimation is to identify the parameter estimates that maximize the sample log-likelihood.

The process of computing maximum likelihood standard errors does not change much with missing data, except that it is necessary to distinguish between standard errors that are based on the observed information matrix and the expected information matrix. The expected information matrix replaces certain terms in the second derivative formulas with their expected values (i.e., long-run averages), whereas the observed information uses the realized data values to compute these terms. This is an important distinction because the expected information matrix yields standard errors that require the MCAR assumption, whereas the observed information matrix gives standard errors that are appropriate with MAR data. Because they make less stringent assumptions, the missing data literature clearly favors standard errors based on the observed information matrix.

With few exceptions, missing data analyses require iterative optimization algorithms, even for very simple estimation problems. This chapter described one such algorithm that is particularly important for missing data analyses, the EM algorithm. The EM algorithm is a two-step iterative procedure that consists of an E-step and an M-step. The E-step uses the elements from the mean vector and the covariance matrix to derive regression equations that predict the incomplete variables from the complete variables, and the M-step subsequently uses standard complete-data formulas to generate updated estimates of the mean vector and the covariance matrix. The algorithm carries these updated parameter values forward to the next E-step, where the process begins anew. EM repeats these two steps until the elements in the mean vector and the covariance matrix no longer change between consecutive M-steps, at which point the algorithm has converged on the maximum likelihood estimates.

With the basic principles of maximum likelihood estimation established in this chapter, the next chapter describes procedures useful for fine-tuning a maximum likelihood analysis. Specifically, the chapter outlines auxiliary variable models that incorporate correlates of missingness into a maximum likelihood analysis. Adopting this so-called inclusive analysis strategy can decrease bias, increase power, and improve the chances of satisfying the MAR assumption. The chapter also outlines corrective procedures that remedy the negative effects of nonnormal data.

4.20 RECOMMENDED READINGS

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling* (pp. 243–277). Mahwah, NJ: Erlbaum.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39*, 1–38.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling: An Interdisciplinary Journal, 8*, 430–457.

Kenward, M. G., & Molenberghs, G. (1998). Likelihood based frequentist inference when data are missing at random. Statistical Science, 13, 236–247.