8

The Analysis and Pooling Phases of Multiple Imputation

8.1 CHAPTER OVERVIEW

A multiple imputation analysis consists of three distinct steps: the imputation phase, the analysis phase, and the pooling phase. Chapter 7 described the mechanics of the imputation phase, and the purpose of this chapter is to outline the analysis and pooling phases. The purpose of the analysis phase is to analyze the filled-in data sets from the preceding imputation phase. This step consists of *m* statistical analyses, one for each imputed data set. The analysis phase yields several sets of parameter estimates and standard errors, so the goal of the pooling phase is to combine everything into a single set of results. Rubin (1987) outlined relatively straightforward formulas for pooling parameter estimates and standard errors. For example, the pooled parameter estimate is simply the arithmetic average of the estimates from the analysis phase. Combining the standard errors is slightly more complex but follows the same logic. The analysis phase is probably the easiest aspect of multiple imputation and requires very little explanation. Consequently, the majority of this chapter is devoted to the pooling phase, including the various significance testing procedures that are available at this step.

As an advance warning, this chapter is relatively dense with equations, largely due to the complexity of the multiple imputation significance tests. Not all of these formulas are equally important. For example, understanding Rubin's (1987) equations for combining parameter estimates and standard errors is probably far more important than trying to digest the different test statistics and their degrees of freedom. Software packages implement the majority of the significance testing procedures that I outline in this chapter, so there is usually no need to compute the formulas by hand. Nevertheless, I felt that it was important for this chapter to serve as a comprehensive reference, so I included more equations than usual. The abundance of equations should not hinder readers who are interested primarily in applying multiple imputation to their own research because the majority of the text does not require an in-depth understanding of the formulas.

TABLE 8.1. 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

I use the small data set in Table 8.1 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 that 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 missing at random (MAR) because they are systematically missing as a function of IQ (i.e., individuals in the lower half of the IQ distribution were never hired, and thus have no performance rating). In addition, I randomly deleted three of the well-being scores in order to mimic a situation where the applicant's well-being questionnaire is inadvertently lost.

8.2 THE ANALYSIS PHASE

The analysis phase is probably the easiest aspect of a multiple imputation analysis. The imputation phase generates *m* imputed data sets, each of which contains different estimates of the missing values. The purpose of the analysis phase, as noted earlier, is to analyze the filled-in data sets. This step consists of *m* statistical analyses, one for each imputed data set. For example, suppose that a researcher had previously generated 20 imputations and is now interested in estimating a multiple regression equation. In the analysis phase, she would simply repeat the regression analysis 20 times, once for each data set. The researcher can employ the

same analysis procedures and the same software package that she would have used had the data been complete. Of course, repeating the analysis 20 times sounds incredibly tedious, but an increasing number of software packages have built-in routines that automate this process.

As an important aside, auxiliary variables play no role in the analysis phase. Multiple imputation can readily accommodate auxiliary variables, but this is handled in the imputation phase. The imputation process infuses the imputed values with the information from the auxiliary variables, so there is no need to include the additional variables in the subsequent analysis step. This is in contrast to maximum likelihood estimation, which uses the somewhat awkward saturated correlates approach to incorporate auxiliary variables. Although multiple imputation is arguably more difficult to implement, it holds a clear advantage over maximum likelihood when it comes to dealing with auxiliary variables.

8.3 COMBINING PARAMETER ESTIMATES IN THE POOLING PHASE

The analysis phase yields *m* different estimates of each parameter, any one of which is unbiased if the data are MAR. Rather than rely on the results from any single data set, a multiple imputation analysis pools the *m* parameter values into a single point estimate. Rubin (1987) defined the **multiple imputation point estimate** as the arithmetic average of the *m* estimates

$$\bar{\theta} = \frac{1}{m}\sum_{t=1}^{m}\hat{\theta}_t \tag{8.1}$$

where $\hat{\theta}_t$ is the parameter estimate from data set *t* and $\bar{\theta}$ is the pooled estimate. Notice that Equation 8.1 is the usual formula for the sample mean, where the parameter estimates serve as data points. Although Rubin (1987) developed multiple imputation in the Bayesian framework, the pooled point estimate is meaningful from either a Bayesian or a frequentist perspective. From the frequentist standpoint, $\bar{\theta}$ is a point estimate of the fixed population parameter, whereas the Bayesian paradigm views $\bar{\theta}$ as the mean of the observed-data posterior distribution (Little & Rubin, 2002, pp. 210–211; Rubin, 1987).

A Bivariate Analysis Example

To illustrate the pooling process, suppose that it is of interest to use the data in Table 8.1 to estimate the regression of job performance on IQ. After generating 20 imputed data sets, I fit an ordinary least squares regression model to each data set and saved the estimates and the standard errors to a file for further analysis. Table 8.2 shows the regression slopes from the analysis phase. As seen in the table, the regression coefficients ranged between –0.025 and 0.239. Substituting the 20 estimates into Equation 8.1 yields a pooled point estimate of $\bar{\theta}$ = 0.105. The fact that the pooled estimate is an average of 20 different values has no bearing on its interpretation. Consistent with a complete-data regression analysis, 0.105 is the expected change in job performance for a one-point increase in IQ.

TABLE 8.2. Regression Coefficients and Sampling Variances from the Bivariate Analysis Example

Imputation	$\hat{\theta}_t$	SE_t	SE_t^2
1	0.12630	0.03639	0.00132
2	0.09499	0.04978	0.00248
3	0.05515	0.08348	0.00697
4	0.06942	0.03509	0.00123
5	0.16699	0.03901	0.00152
6	0.02960	0.06283	0.00395
7	0.20581	0.04523	0.00205
8	0.02627	0.03739	0.00140
9	0.05293	0.03456	0.00119
10	0.15939	0.05294	0.00280
11	0.18642	0.03604	0.00130
12	0.14726	0.03933	0.00155
13	0.23944	0.03601	0.00130
14	0.04638	0.04718	0.00223
15	0.10295	0.05341	0.00285
16	0.07162	0.04275	0.00183
17	0.20742	0.03783	0.00143
18	–0.02501	0.04752	0.00226
19	0.09447	0.03839	0.00147
20	0.04705	0.04372	0.00191

8.4 TRANSFORMING PARAMETER ESTIMATES PRIOR TO COMBINING

The pooling formula in Equation 8.1 assumes that the parameter estimates are asymptotically (i.e., in very large samples) normally distributed. However, some parameters meet this requirement better than others do, particularly in small and moderate samples. For example, the sampling distribution of Pearson's correlation is normal when the population correlation equals zero but becomes increasingly skewed as ρ approaches plus or minus one. Many common variance estimates (e.g., R^2 statistics, standard deviations, estimates of variances, and covariances) also have skewed sampling distributions (or from the Bayesian framework, skewed posterior distributions). These distributions eventually normalize as the sample size gets very large, but they can be markedly non-normal in small and moderate samples. Averaging m parameter values into a single estimate is asymptotically valid for any parameter, but applying normalizing transformations prior to the pooling phase may improve the accuracy of certain estimates (Schafer, 1997).

To illustrate the use of normalizing transformations, consider Pearson's correlation coefficient. Fisher's (1915) z transformation is a natural choice for pooling correlations because it places the estimates on a metric that more closely approximates a normal distribution. The transformation is

$$z_t = \frac{1}{2}\log\left(\frac{r_t + 1}{r_t - 1}\right) \tag{8.2}$$

where r_t is the correlation coefficient from data set t and z_t is the corresponding transformed coefficient. Substituting the transformed correlations into Equation 8.1 expresses the average correlation on the z score metric, and the equation below transforms the pooled estimate back to the correlation metric.

$$\bar{r} = \left(\frac{e^{2\bar{\theta}} - 1}{e^{2\bar{\theta}} + 1} \right) \tag{8.3}$$

Applying normalizing transformations to variances and covariances is more complex because the appropriate transformation may not be immediately obvious. For example, a logarithmic transformation may work best for a distribution with substantial positive skewness, whereas a square root transformation may be more appropriate for a moderately skewed distribution. When transforming raw data, methodologists often recommend experimenting with different transformations to identify the one that best normalizes the data, but this exploratory approach is unlikely to work well in the pooling phase. Given the potential difficulties associated with specifying an appropriate transformation, it is reasonable to ask whether the use of transformations makes any practical difference. Because parameter distributions tend to normalize as N increases, it is also important to determine whether there is a sample size at which transformations are no longer necessary. I am unaware of any studies that have systematically evaluated the use of transformations at the pooling phase, so I performed some computer simulations to examine this issue.

Briefly, the computer simulations generated 1,000 samples of bivariate normal data from a population with a correlation of $\rho = .50$. I subsequently imposed missing completely at random (MCAR) data by randomly deleting 25% of the values from one of the variables. Because the sample size plays an important role, I examined six different sample size conditions ($N = 50$, 100, 200, 300, 500, and 1,000). Finally, I created $m = 10$ imputations for each sample and applied logarithmic and square root transformations prior to pooling variances, covariances, and R^2 statistics. Although my simulations were very limited in scope, they do suggest that normalizing transformations tend to make very little difference, particularly when the sample size exceeds $N = 200$. Averaging the transformed estimates did reduce bias, but the mean squared errors of the transformed estimates were virtually identical to those of the raw estimates (the mean squared error is an overall measure of accuracy that combines bias and sampling error). The mean squared error results are interesting because they suggest that normalizing transformations increase sampling error to a degree that effectively negates the reduction in bias. Consequently, there may be little or no practical advantage to transforming estimates prior to combining them. (Fisher's transformation is a notable exception because it provides a convenient mechanism for significance testing.) As a caveat, my simulations were very limited in scope, so it is a good idea to view the results with some caution. Further methodological research should attempt to clarify this issue.

8.5 POOLING STANDARD ERRORS

The analysis phase also yields m estimates of each standard error. Pooling standard errors is not as simple as computing an arithmetic average, but Rubin's (1987) combining rules are

still relatively straightforward. Multiple imputation standard errors combine two sources of sampling fluctuation: the sampling error that would have resulted had the data been complete, and the additional sampling error that results from missing data. As an aside, Rubin's pooling formulas operate on the sampling variance metric rather than on the standard error metric. However, the sampling variance is simply the squared standard error, so switching to the standard error metric is an easy conversion.

Within-Imputation Variance

A multiple imputation standard error consists of two sources of sampling fluctuation: within-imputation variance and between-imputation variance. The **within-imputation variance** is the arithmetic average of the *m* sampling variances

$$V_W = \frac{1}{m}\sum_{t=1}^{m} SE_t^2 \tag{8.4}$$

where V_W denotes the within-imputation variance, and SE_t^2 is the squared standard error (i.e., sampling variance) from data set *t*. Notice that Equation 8.4 is the usual formula for the sample mean, where the sampling variances serve as data points. Equation 8.4 averages complete-data sampling variances, so the within-imputation variance effectively estimates the sampling variability that would have resulted had there been no missing data.

Between-Imputation Variance

At an intuitive level, missing values should increase standard errors because they add an additional layer of noise to the parameter estimates. Single imputation techniques fail to address this issue because they treat the filled-in values as real data. Consequently, even the best single imputation technique (e.g., stochastic regression imputation) will underestimate standard errors. Analyzing multiply imputed data sets solves this problem because it provides a mechanism for estimating the additional source of sampling error. As an illustration, reconsider the regression coefficients in Table 8.2. The variation in the regression coefficients from one data set to the next (the estimates range between –0.025 and 0.239) is solely due to the use of different imputed values. Consequently, the variability of the parameter values across the *m* data sets estimates the additional sampling fluctuation that results from the missing data.

More formally, the **between-imputation variance** quantifies the variability of a parameter estimate across the *m* data sets, as follows:

$$V_B = \frac{1}{m-1}\sum_{t=1}^{m} (\hat{\theta}_t - \bar{\theta})^2 \tag{8.5}$$

where V_B denotes the between-imputation variance, $\hat{\theta}_t$ is the parameter estimate from data set *t*, and $\bar{\theta}$ is the average point estimate from Equation 8.1. Notice that Equation 8.5 is the

usual formula for the sample variance, where the parameter estimates serve as data points. Again, the between-imputation variance represents the additional sampling error that results from the missing data because the fluctuation of the $\hat{\theta}_t$ values from one data set to the next is solely due to the use of different imputed values.

Total Sampling Variance

Equations 8.4 and 8.5 decompose sampling error into two components: the sampling fluctuation that would have resulted had the data been complete (i.e., the within-imputation variance) and the additional sampling error that results from the missing data (i.e., the between-imputation variance). The **total sampling variance** combines these two components into a single quantity, as follows:

$$V_T = V_W + V_B + \frac{V_B}{m} \tag{8.6}$$

You might have anticipated that the total sampling variance is just the sum of the within- and between-imputation components, but the equation has an additional term, V_B / m. The between-imputation variance in Equation 8.5 requires the average parameter estimate (i.e., $\bar{\theta}$), and this mean is also subject to sampling error. The right-most term in Equation 8.6 quantifies the sampling variance (i.e., squared standard error) of the mean and essentially serves as a correction factor for using a finite number of imputations. (As m goes to infinity, this term vanishes and the total variance becomes the sum of V_W and V_B.)

Researchers are generally accustomed to reporting their results on the standard error metric rather than on the variance metric. Therefore, taking the square root of the total variance gives the multiple imputation standard error, as follows:

$$SE = \sqrt{V_T} \tag{8.7}$$

Throughout this section, I have been referring to various quantities as sampling variances, which implies repeated sampling (i.e., a frequentist interpretation). However, the total variance is meaningful from either a Bayesian or a frequentist perspective. From a frequentist perspective, the total variance estimates the variability of a parameter estimate across repeated samples. In contrast, the Bayesian paradigm views V_T as the variance of the observed-data posterior distribution. The difference in terminology is not just semantics and represents an important philosophical difference between the two paradigms (see Chapter 6). Because the standard error is a familiar concept, I use this term throughout the remainder of the book (much of the multiple imputation literature follows the same convention).

An ANOVA Analogy

Partitioning a parameter's sampling variance into within- and between-imputation components is very similar to what happens in an analysis of variance (ANOVA). ANOVA partitions

score variation into two orthogonal sources: explained variability that is attributable to an explanatory variable (i.e., between-group variability) and residual variation that remains after accounting for the explanatory variable (i.e., within-group variability). The pooling phase partitions variance in a manner that closely resembles an ANOVA analysis, but it does so using the variation in a *parameter* distribution rather than a score distribution.

To align the previous concepts with an ANOVA analysis, you can think of missingness as an explanatory variable and the total sampling variance as the variability in the outcome variable. In this analogy, the between-imputation variance quantifies the portion of the parameter's variance that is due to the explanatory variable (i.e., the missing data) and is akin to the between-group mean square from an ANOVA analysis. The within-imputation variance is the residual variation that remains after subtracting out the explanatory variable's influence (i.e., the sampling variation that would result had there been no missing data) and is analogous to the mean square error in an ANOVA. Thinking about V_W and V_B in ANOVA terms puts Rubin's (1987) combining rules in a familiar context, but it also leads to an intuitive interpretation of some important quantities that I define later in the chapter.

A Bivariate Analysis Example

To illustrate the process of combining standard errors, reconsider the regression of job performance on IQ. Table 8.2 also shows the standard errors and the sampling variances from the 20 regression analyses. Averaging the squared standard errors in the right-most column of the table yields a within-imputation variance of $V_W = 0.00215$. Again, this is an estimate of the sampling variability that would have resulted had the data been complete. Next, using Equation 8.5 to compute the variance of the regression coefficients across the 20 imputations gives a between-imputation variance of $V_B = 0.00515$. As I explained previously, the between-imputation variance represents the additional uncertainty that results from the missing data. Finally, substituting V_W and V_B into Equation 8.6 yields the total variance, $V_T = 0.00756$, and taking the square root of this value gives the multiple imputation standard error, $SE = 0.087$. Notice that the pooled standard error is considerably larger than most of the individual standard errors in Table 8.2. (The average complete-data standard error is 0.045.) This makes intuitive sense because multiple imputation explicitly incorporates the additional sampling error that accrues from the missing data.

8.6 THE FRACTION OF MISSING INFORMATION AND THE RELATIVE INCREASE IN VARIANCE

The within-imputation variance, between-imputation variance, and the total variance define two useful diagnostic measures, the fraction of missing information and the relative increase in variance due to nonresponse. These measures are important because they (1) quantify the influence of missing data on the standard errors, (2) dictate the convergence speed of the data augmentation algorithm, and (3) help define the significance tests outlined later in the chapter.

The Fraction of Missing Information

I briefly introduced the fraction of missing information in Chapter 7, where I described it as a diagnostic measure that adjusts the missing data rate by the correlations among the variables. More specifically, the **fraction of missing information** quantifies the missing data's influence on the sampling variance of a parameter estimate. An intuitive expression for the fraction of missing information is as follows.

$$\text{FMI} = \frac{V_B + V_B/m}{V_T} \tag{8.8}$$

Equation 8.8 assumes that the number of imputations is very large; thus, an alternate expression that adjusts for a finite number of imputations is

$$\text{FMI}_1 = \frac{V_B + V_B/m + 2/(\nu + 3)}{V_T} \tag{8.9}$$

where ν is a degrees of freedom value that is defined later in Equation 8.12. The value of ν increases to infinity as m goes to infinity, so the additional terms in the numerator essentially vanish with a very large number of imputations. The result is the more straightforward expression in Equation 8.8.

Focusing on Equation 8.8, the fraction of missing information has an intuitive interpretation. The denominator is the total sampling variance (i.e., squared standard error), and the numerator quantifies the additional sampling variation that accrues from the missing data. Consequently, the fraction of missing information is the proportion of the total sampling variance that is due to the missing data. If you think of between- and within-imputation variance as being similar to the between- and within-group variation from ANOVA, then the fraction of missing information is analogous to an R^2 statistic. In the context of multiple imputation, the pooling phase partitions the variation in a parameter distribution rather than a score distribution, but the R^2 analogy is useful for understanding the equation.

With regard to the previous regression example, substituting $V_B = 0.00515$ and $V_T = 0.00755$ into Equation 8.8 yields FMI = 0.715. This value indicates that 71.5% of the regression coefficient's sampling variance is attributable to the missing data. Using the more complex expression in Equation 8.9 gives an estimate that is more appropriate for a finite number of imputations, but the interpretation remains the same (i.e., $\text{FMI}_1 = 0.729$, so approximately 73% of the sampling variance is due to the missing data). I previously described missing information as a summary measure that combines the missing data rate and the correlations among the variables. The missing information is typically lower than the missing data rate, particularly when the variables in the imputation model are predictive of the missing values. In this situation, the correlations among the variables mitigate the information loss, such that the increase in sampling error is not completely commensurate with the overall reduction in the sample size. The regression analysis produced a fraction of missing information that exceeds the missing data rate, but this is likely an artifact of the sample size and the number of imputations (accurate FMI estimates require far more than 20 data sets).

The fraction of missing information is also a useful diagnostic tool because it influences the convergence of the data augmentation algorithm. (Parameters with high rates of missing information tend to converge slowly.) Consequently, paying especially close attention to parameters with large fractions of missing information is a good strategy when examining the graphical diagnostics from the imputation phase. Because some multiple imputation software packages report the fraction of missing information as a by-product of the imputation phase, usually these estimates are readily available. As an aside, methodologists have noted that the fraction of missing information tends to be noisy and somewhat untrustworthy (Harel, 2007; Schafer, 1997), particularly with fewer than 100 imputations (Harel, 2007). However, estimating the fraction of missing information is usually not the primary analytic goal, so approximate estimates are often acceptable.

Relative Increase in Variance

Like the fraction of missing information, the relative increase in variance quantifies the missing data's influence on the sampling variance of a parameter estimate, but it does so in a slightly different fashion. A standard formulation of the **relative increase in variance** is

$$\text{RIV} = \frac{V_B + V_B/m}{V_W} = \frac{\text{FMI}}{1 - \text{FMI}} \tag{8.10}$$

To understand the relative increase in variance, consider the meaning of its component parts. The denominator of Equation 8.10 estimates the sampling variance that would have resulted had there been no missing data, and the numerator of the equation quantifies the additional sampling variation that accrues from the missing data. Consequently, the relative increase in variance gives proportional increase in the sampling variance that is due to the missing data. For example, if the missing data have no influence on the sampling error of a particular parameter, the between-imputation variance is zero, as is the relative increase in variance. In contrast, if the between-imputation variance is equal to the within-imputation variance, then the relative increase in variance equals one. Returning to the previous regression analysis, note that the between- and the within-imputation variance estimates are $V_B = 0.00515$ and $V_W = 0.00215$, respectively. Substituting these values into Equation 8.10 yields RIV = 2.51. This means that the sampling fluctuation due to the missing data is two and a half times larger than the sampling variance of a complete-data analysis.

Like the fraction of missing information, the relative increase in variance dictates the convergence speed of the data augmentation algorithm. Equation 8.10 shows that FMI and RIV are one-to-one transformations, so it makes little difference which measure you choose to examine. Several of the significance tests outlined in subsequent sections rely on the relative increase in variance (or equivalently, the fractional missing information), so these concepts will resurface throughout the remainder of the chapter.

8.7 WHEN IS MULTIPLE IMPUTATION COMPARABLE TO MAXIMUM LIKELIHOOD?

Having gained an understanding of all three phases in a multiple imputation analysis, it is useful to consider the comparability of maximum likelihood and multiple imputation. Maximum likelihood and multiple imputation are equivalent techniques in the sense that they both assume multivariate normality and MAR data. Despite making the same assumptions, the two approaches may or may not yield similar parameter estimates and standard errors. Assuming that the sample size and the number of imputations are both large enough to eliminate idiosyncratic performance differences, the set of input variables and the relative complexity of the imputation model and the analysis model largely determine whether the two procedures produce similar results (Collins, Schafer, & Kam, 2001; Schafer, 2003).

When comparing multiple imputation and maximum likelihood, the first thing to consider is whether the imputation phase uses the same set of variables as the analysis phase. To illustrate, consider an analysis model that involves three variables, *X*, *M*, and *Y*. A researcher could use maximum likelihood to directly estimate the analysis model, or she could impute the data and analyze the *m* complete data sets. If the imputation phase includes additional variables that are not part of maximum likelihood analysis (e.g., a set of auxiliary variables), then the two procedures can yield different estimates, standard errors, or both. If the imputation phase includes only *X*, *M*, and *Y*, then multiple imputation and maximum likelihood are seemingly on an equal footing because they make the same assumptions and use the same set of input variables. However, the comparability of the two procedures still depends on the relative complexity of the imputation and the analysis models.

Recall from Chapter 7 that the imputation phase of a multiple imputation analysis uses a multiple regression model to fill in the missing values. A multiple regression model is known as a **saturated model** because the number of parameters in the model exactly equals the number of elements in the mean vector and the covariance matrix (i.e., there is a one-to-one transformation that links the regression model parameters to the elements in $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$). In practical terms, this means that the imputation phase uses the most complex model possible to impute the missing values (i.e., estimating the regression model expends all of the information present in the mean vector and the covariance matrix). The subsequent analysis model may or may not be as complex as the imputation regression model, and the relative parsimony of these two models has a bearing on the comparability of multiple imputation and maximum likelihood.

To illustrate the parsimony issue, consider a mediation analysis in which *X* predicts *M*, *M* predicts *Y*, and *X* also has a direct influence on *Y*. The top panel of Figure 8.1 shows a path diagram of this model. To begin, notice that the mediation model is saturated because it, too, estimates every possible association among the variables. Assuming that the imputation phase includes only three variables, then a maximum likelihood analysis of the mediation model estimates the same number of parameters as the imputation phase (i.e., the number of parameters in the mediation model equals the number of elements in $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$). When the imputation and analysis models use the same set of input variables and estimate the same number of parameters, the two models are said to be **congenial** (Meng, 1994). In this situation, multiple imputation and maximum likelihood should produce very similar estimates

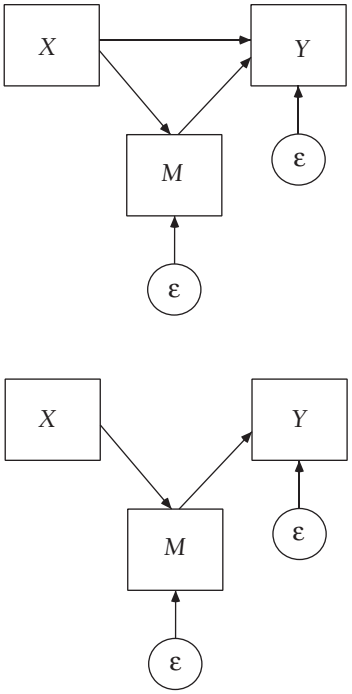

FIGURE 8.1. Path diagram of a mediation analysis model. The top panel shows a model where *X* has a direct relationship with *Y* and is also related to *Y* via a mediating variable, *M*. The bottom panel shows a model where *X* and *Y* are only related via their mutual association with *M*.

and standard errors (Collins et al., 2001; Schafer, 2003). All things being equal, Bayesian estimation is asymptotically (i.e., in large samples) equivalent to maximum likelihood (Gelman, Carlin, Stern, & Rubin, 1995), so there is no theoretical reason for the procedures to produce different results.

Next, consider an analysis model that restricts the association between *X* and *Y* to zero during estimation, such that the relationship between *X* and *Y* is completely mediated by *M*. The bottom panel of Figure 8.1 shows a path diagram of this model. Unlike the previous example, the imputation and analysis models are now **uncongenial** because they differ in complexity. That is, the analysis model restricts the association between *X* and *Y*, whereas the imputation regression model does not. When the imputation and analysis models are uncongenial but use the same set of input variables, multiple imputation and maximum likelihood should produce very similar parameter estimates, but multiple imputation standard errors may be slightly larger (Collins et al., 2001; Schafer, 2003). In effect, the imputation phase uses an unnecessarily complex model to deal with the missing data, and this additional complexity can add a small amount of noise to the resulting estimates. However, the difference between the two sets of standard errors is usually trivial, so uncongeniality is not necessarily a reason to favor maximum likelihood estimation.

The previous example might suggest that uncongeniality is detrimental to a multiple imputation analysis. However, uncongeniality can be beneficial when it results from an inclu-

sive analysis strategy that incorporates auxiliary variables that are correlates of missingness or correlates of the incomplete analysis model variables. Because a single set of imputations can serve as input data for a variety of different analyses, it is natural for the imputation phase to include a much larger set of variables than would appear in any single analysis model. Returning to the analysis models in Figure 8.1, note that an ideal imputation model would include the mediation model variables, variables from other analyses, and a set of auxiliary variables. When the imputation phase includes additional variables that are not part of the analysis model, multiple imputation and maximum likelihood can yield different parameter estimates, standard errors, or both. Idiosyncratic features of the data influence these discrepancies, so it is difficult to make predictions about the pattern and the magnitude of the differences (e.g., some estimates may be similar, others may be different; one procedure may produce smaller standard errors for some parameters but not others).

A final situation in which multiple imputation and maximum likelihood can differ occurs when the imputation model is more restrictive than the analysis model. Returning to the mediation example, suppose that it is of interest to determine whether the regression coefficient between *X* and *M* is different for males and females (e.g., using a multiple group path analysis model or a regression model with interaction terms). Furthermore, suppose that the imputation phase includes *X*, *M*, *Y*, and a gender dummy code. In this situation, including the dummy code in the imputation phase accounts for mean differences between males and females, but omitting the gender by *X* product term effectively assumes that the gender groups have the same covariance between *X* and *M*. This is a potentially harmful form of uncongeniality because the subsequent analyses can attenuate the interaction effect. Maximum likelihood estimation would not suffer from this problem, so it is possible for the two approaches to produce very different estimates and standard errors. This example underscores the well-established but important point that omitting analysis variables from the imputation phase can produce biased parameter estimates, regardless of the missing data mechanism (Meng, 1994; Rubin, 1996).

8.8 AN ILLUSTRATIVE COMPUTER SIMULATION STUDY

In Chapter 4, I illustrated the accuracy of maximum likelihood analyses using computer simulations. Having outlined the analysis and pooling phases, I repeated these simulations, this time using multiple imputation to deal with missing data. The simulation programs generated 1,000 samples of $N = 250$ from a population model that mimicked the IQ and job performance data in Table 8.1. 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 missing not at random (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, I used the data augmentation algorithm from Chapter 7 to create $m = 10$ imputed data sets for each sample. Next, I estimated the mean vector and the covariance matrix from each imputed data set and used Equation 8.1 to pool the resulting estimates. Table 8.3 shows the average multiple-imputation estimates from the simulations and uses **bold** typeface to

TABLE 8.3. Average Parameter Estimates from the Illustrative Computer Simulation

Parameter	Population value	Multiple imputation	Maximum likelihood
		MCAR simulation	
μ_{IQ}	100.00	99.98	100.02
μ_{JP}	12.00	11.99	11.99
σ^2_{IQ}	169.00	169.34	168.25
σ^2_{JP}	9.00	9.08	8.96
$\sigma_{IQ,JP}$	19.50	19.51	19.48
		MAR simulation	
μ_{IQ}	100.00	100.00	100.01
μ_{JP}	12.00	12.00	12.01
σ^2_{IQ}	169.00	168.46	168.50
σ^2_{JP}	9.00	9.23	8.96
$\sigma_{IQ,JP}$	19.50	19.43	19.15
		MNAR simulation	
μ_{IQ}	100.00	100.02	100.00
μ_{JP}	12.00	**14.13**	**14.12**
σ^2_{IQ}	169.00	170.37	169.11
σ^2_{JP}	9.00	**3.42**	**3.33**
$\sigma_{IQ,JP}$	19.50	**8.51**	**8.55**

highlight severely biased estimates. For comparison purposes, the table also shows the corresponding maximum likelihood estimates.

As seen in the table, the multiple imputation and maximum likelihood parameter estimates are virtually indistinguishable in all three simulations, which is not surprising given that the imputation and analysis models are congenial (i.e., they include the same variables and estimate the same number of parameters). Consistent with maximum likelihood estimation, multiple imputation produced unbiased estimates in the MCAR and MAR simulations but gave biased estimates in the MNAR simulation. However, it is important to point out that the MNAR bias was confined to the parameters that were affected by missing data. Although these simulations were limited in scope, the results are consistent with missing data theory (Rubin, 1976; Schafer, 1997) and with previous simulation studies (e.g., Allison, 2000; Collins et al., 2001; Graham & Schafer, 1999; Newman, 2003).

8.9 SIGNIFICANCE TESTING USING THE *t* STATISTIC

The next few sections outline a number of multiple imputation significance tests. Again, the subsequent sections are relatively dense with equations, but not all of these formulas are

equally important (e.g., the degrees of freedom equations are complex and not very intuitive). The abundance of equations should not hinder readers who are primarily interested in applying multiple imputation to their own research because the majority of the text does not require an in-depth understanding of the formulas.

In the context of a maximum likelihood analysis, the Wald *z* test provides a mechanism for assessing whether a parameter estimate is statistically different from some hypothesized value. Multiple imputation analyses use an analogous *t* statistic. Like the Wald test, the numerator of the *t* statistic compares the point estimate to some hypothesized value, and the denominator contains the standard error, as follows:

$$t = \frac{\bar{\theta} - \theta_0}{\sqrt{V_T}} \tag{8.11}$$

where $\bar{\theta}$ is the pooled point estimate, and θ_0 is the hypothesized parameter value. Researchers typically test whether a parameter is significantly different from zero, in which case the *t* statistic reduces to the ratio of the point estimate to its standard error.

Many complete-data statistical procedures employ a *t* statistic similar to that in Equation 8.11, but the multiple imputation test statistic uses a complex expression for the degrees of freedom (Rubin, 1987; Rubin & Schenker, 1986).

$$\nu = (m - 1)\left(1 + \frac{V_W}{V_B + V_B/m}\right)^2 = (m - 1)\left(\frac{1}{FMI^2}\right) \tag{8.12}$$

With complete data, the *t* sampling distribution converges to a normal curve as the sample size becomes very large (i.e., the degrees of freedom approach infinity). Interestingly, the sample size does not directly influence the value of ν. Instead, the degrees of freedom increase as the number of imputations increase or as the fraction of missing information decreases. For example, substituting $m = 20$ and FMI = .25 (e.g., a 25% missing data rate) into Equation 8.12 yields $\nu = 304$, whereas $m = 20$ and FMI = .05 gives a degrees of freedom value of $\nu = 7600$.

In small to moderate samples, ν can substantially exceed the degrees of freedom that would have resulted had the data been complete. Returning to the previous regression example, observe that the complete-data regression of job performance on IQ would have $N - k - 1 = 18$ degrees of freedom, where *k* is the number of predictor variables. In contrast, Equation 8.12 yields a value of $\nu = 37.148$. To correct this problem, Barnard and Rubin (1999) proposed the following adjusted degrees of freedom value

$$\nu_1 = \left(\frac{1}{\nu} + \frac{1}{\tilde{\nu}}\right)^{-1} \tag{8.13}$$

where

$$\tilde{\nu} = (1 - FMI)\left(\frac{df_{com} + 1}{df_{com} + 3}\right)df_{com} \tag{8.14}$$

and df_{com} is the degrees of freedom that would have resulted had the data been complete. Unlike ν, the adjusted degrees of freedom value increases as the sample size increases and never exceeds the complete-data degrees of freedom. For example, the adjusted degrees of freedom for the previous regression example is $\nu_1 = 4.124$ as opposed to $\nu = 37.148$. Barnard and Rubin's (1999) computer simulations suggest that ν_1 improves the accuracy of confidence intervals in small samples, so you should use the adjusted degrees of freedom whenever possible.

Confidence Intervals

Establishing a confidence interval around a multiple imputation point estimate requires the appropriate critical values from a *t* distribution with ν_1 degrees of freedom. To get the upper and lower confidence interval limits, you multiply the standard error by the appropriate critical value and add the resulting product to the pooled point estimate, as follows:

$$\bar{\theta} + (t_{\nu_{1,1-\alpha/2}})\left(\sqrt{V_T}\right) \tag{8.15}$$

where $t_{\nu_{1,1-\alpha/2}}$ is the *t* critical value that separates the desired proportion of the distribution. For example, the 95% confidence interval requires the *t* critical value that separates the upper and the lower 2.5% of a *t* sampling distribution with ν_1 degrees of freedom. Multiple imputation software programs generally report confidence intervals, but you can obtain the *t* critical values from other software programs (e.g., Excel), if need be.

A Bivariate Analysis Example

Returning to the previous bivariate analysis, note that the regression of job performance on IQ produced a slope estimate of $\bar{\theta} = 0.105$ and a standard error of $SE = 0.087$. The test statistic for the regression coefficient is $t = 1.207$, and referencing the statistic to a *t* distribution with $\nu_1 = 4.124$ degrees of freedom returns a probability value of $p = .29$. With an alpha level of 0.05, the two-tailed critical value for a *t* distribution with 4.124 degrees of freedom is 2.776, therefore, substituting the appropriate values into Equation 8.15 gives upper and lower confidence limits of 0.347 and –0.137, respectively. Aside from using a fractional degrees of freedom value, the significance testing procedure is virtually identical to that of a complete-data analysis.

Revisiting the Number of Imputations

Recall from Chapter 7 that the number of imputations has an impact on the power of multiple imputation significance tests, such that power improves as *m* increases. The equations in this section illustrate that increasing the number of imputations can improve power in two ways. First, reconsider the expression for the total sampling variance (i.e., squared standard error) in Equation 8.6. The formula includes a correction factor (i.e., V_B / m) that quantifies the sampling error of the pooled point estimate. Increasing the number of imputations decreases the value of the correction factor and thus decreases the standard error. Increasing

the number of imputations also improves power in a more subtle fashion. Equations 8.12 and 8.13 show that the degrees of freedom value increases as the number of imputations increases. As the degrees of freedom increase, the *t* critical value decreases, making it easier to reject the null hypothesis. Consequently, all things being equal, analyses with a large number of imputations will produce more powerful significance tests than analyses with a small number of imputations. Computer simulation studies suggest that $m = 20$ is a good rule of thumb for many situations (Graham, Olchowski, & Gilreath, 2007), but increasing the number of imputations beyond this point is certainly a good idea, if processing time permits. Using a large number of imputations will also improve the performance of the multiple-parameter significance tests that are described next, although much less is known about the impact of *m* on these tests.

8.10 AN OVERVIEW OF MULTIPARAMETER SIGNIFICANCE TESTS

In many situations it is of interest to determine whether a set of parameters is significantly different from zero. For example, in a multiple regression analysis, researchers are often interested in testing whether two or more regression slopes are different from zero. In an ordinary least squares analysis with complete data, it is standard practice to use an omnibus *F* test for this purpose. In the context of maximum likelihood estimation, the multivariate Wald test and the likelihood ratio statistic are analogous procedures. Multiple imputation also offers different mechanisms for testing a set of parameter estimates (the literature sometimes refers to these procedures as **multiparameter inference** or **multivariate inference**), although relatively little is known about the performance of these tests.

The subsequent sections describe three different multiparameter significance tests. Following Schafer (1997), I refer to these tests as D_1, D_2, and D_3. The D_1 statistic uses the pooled parameter estimates and the pooled sampling variances to construct a test that closely resembles the multivariate Wald statistic from Chapter 3. In contrast, D_2 and D_3 pool significance tests from the analysis phase; the D_2 statistic pools Wald tests, and the D_3 statistic pools likelihood ratio tests. Although these procedures accomplish the same task, they are not equally trustworthy, nor are they equally easy to implement. The D_1 and D_3 statistics are asymptotically equivalent, but D_1 is easier to implement because it is readily available in multiple imputation software programs. (At the time of this writing, relatively few programs compute D_3.) Computing D_2 is straightforward, but it appears to be the least trustworthy of the three test statistics.

8.11 TESTING MULTIPLE PARAMETERS USING THE D_1 STATISTIC

The D_1 statistic uses the pooled parameter estimates and the pooled sampling variances to construct a test that closely resembles the multivariate Wald statistic. Recall from Chapter 3 that the Wald test is

$$\omega = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \text{var}(\hat{\boldsymbol{\theta}})^{-1}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \tag{8.16}$$

where $\hat{\boldsymbol{\theta}}$ is a vector of parameter estimates, $\boldsymbol{\theta}_0$ is a vector of hypothesized values (typically zeros), and var($\hat{\boldsymbol{\theta}}$) contains the appropriate elements from the parameter covariance matrix. In order to construct an analogous test for a multiple imputation analysis, it is first necessary to extend Rubin's (1987) pooling equations to multiple parameters and parameter covariance matrices.

Pooling Multiple Parameter Estimates

Because Rubin's (1987) procedure for combining parameter estimates is unaffected by the shift to multiple parameters, the multiple imputation point estimate is still the arithmetic average of the *m* sets of estimates (see Equation 8.1). Constructing a test that resembles the Wald statistic requires matrix computations, so a column vector $\boldsymbol{\theta}_t$ contains the set of estimates from data set *t*, and the vector $\bar{\boldsymbol{\theta}}$ holds the pooled point estimates.

Pooling Parameter Covariance Matrices

The Wald test in Equation 8.16 uses elements from the parameter covariance matrix to standardize the deviations between the parameter estimates and the hypothesized values. The D_1 statistic uses the same procedure, so it is necessary to extend Rubin's (1987) variance partitioning formulas to multiple parameters. The basic logic of the pooling process remains the same, but covariance matrices quantify the within- and between-imputation variability.

With a single parameter, the within-imputation variance is the arithmetic average of the *m* sampling variances. In the multivariate context, the **within-imputation covariance matrix** is the average of the *m* parameter covariance matrices, as follows:

$$\mathbf{V}_{\mathrm{W}} = \frac{1}{m}\sum_{t=1}^{m} \mathrm{var}(\hat{\boldsymbol{\theta}}_t) \tag{8.17}$$

where $\mathbf{V}_{\mathrm{W}}$ is the average within-imputation covariance matrix, and var($\hat{\boldsymbol{\theta}}_t$) is the parameter covariance matrix from data set *t*. Consistent with the single parameter case, $\mathbf{V}_{\mathrm{W}}$ estimates the parameter covariance matrix that would have resulted had the data been complete.

Filling in the data with different sets of imputed values causes the parameter estimates to vary across the *m* analyses, and this between-imputation variability is an important component of the total sampling error. The **between-imputation covariance matrix** quantifies this variation, as follows:

$$\mathbf{V}_{\mathrm{B}} = \frac{1}{m-1}\sum_{t=1}^{m} (\hat{\boldsymbol{\theta}}_t - \bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_t - \bar{\boldsymbol{\theta}})^T \tag{8.18}$$

where $\mathbf{V}_{\mathrm{B}}$ is the between-imputation covariance matrix, $\hat{\boldsymbol{\theta}}_t$ contains the parameter estimates from data set *t*, and $\bar{\boldsymbol{\theta}}$ is the vector of pooled point estimates (i.e., the arithmetic average of the $\hat{\boldsymbol{\theta}}_t$ vectors). The diagonal elements of $\mathbf{V}_{\mathrm{B}}$ contain the between-imputation variance estimates for individual parameters, and the off-diagonal elements quantify the extent to which

the between-imputation fluctuation in one parameter is related to the between-imputation fluctuation in another parameter. Considered as a whole, the between-imputation covariance matrix represents the additional sampling fluctuation that results from the missing data.

Finally, the **total parameter covariance matrix** combines the within- and between-imputation covariance matrices, as follow:

$$\mathbf{V}_T = \mathbf{V}_W + \mathbf{V}_B + \frac{1}{m}\mathbf{V}_B \tag{8.19}$$

The matrix $\mathbf{V}_T$ reflects the total sampling fluctuation in a set of parameter estimates. Like the parameter covariance matrix from a maximum likelihood analysis, the diagonal elements of $\mathbf{V}_T$ contain sampling variances, and the off-diagonals contain covariances between pairs of estimates.

An Alternate Estimate of the Total Covariance Matrix

The between-imputation covariance matrix in Equation 8.18 is prone to a great deal of sampling error when the number of imputations is small, and this results in a poor estimate of the total parameter covariance matrix. Consequently, using the total covariance matrix in Equation 8.19 to construct a Wald-like test statistic can produce inaccurate inferences. Li, Raghunathan, and Rubin (1991) proposed a solution to this problem that requires an alternate expression for the total covariance matrix.

$$\tilde{\mathbf{V}}_T = (1 + \text{ARIV})\mathbf{V}_W \tag{8.20}$$

Earlier in the chapter, I introduced the relative increase in variance due to nonresponse. The ARIV term in the equation above estimates the **average relative increase in variance** across the k parameter estimates in $\bar{\boldsymbol{\theta}}$ and is defined by

$$\text{ARIV} = \frac{(1 + m^{-1})\text{tr}(\mathbf{V}_B\mathbf{V}_W^{-1})}{k} \tag{8.21}$$

where tr denotes the trace operator (i.e., the sum of the diagonal elements).

To better understand $\tilde{\mathbf{V}}_T$, reconsider the total sampling variance for a single parameter. Applying some algebra to Equation 8.6 gives

$$V_T = (1 + \text{RIV})V_W \tag{8.22}$$

where RIV is the relative increase in variance from Equation 8.10. Defining the total variance in this way makes it clear that $\tilde{\mathbf{V}}_T$ is a matrix analog of V_T, where the average relative increase in variance replaces RIV. Because ARIV condenses the information in the between-imputation covariance matrix into a single numeric value (i.e., ARIV), $\tilde{\mathbf{V}}_T$ can provide a more stable estimate of the total parameter covariance matrix.

The D_1 Statistic

Li, Raghunathan, et al. (1991) proposed the following test statistic:

$$D_1 = \frac{1}{k}(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T(\tilde{\mathbf{V}}_T)^{-1}(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \tag{8.23}$$

where k is the number of parameters in $\bar{\boldsymbol{\theta}}$. Although D_1 closely resembles the Wald test in Equation 8.16, its sampling distribution is far more complex. Li, Raghunathan, et al. suggest using an F distribution with k numerator degrees of freedom and ν_2 denominator degrees of freedom to obtain a probability value, where

$$\nu_2 = 4 + (km - k - 4)\left[1 + \left(1 - \frac{2}{km - k}\right)\frac{1}{\text{ARIV}}\right]^2 \tag{8.24}$$

In a situation where $km - k$ is less than or equal to 4, they recommend an alternate expression for ν_2, as follows.

$$\nu_2 = \frac{(km - k)\left(1 + \frac{1}{k}\right)\left(1 + \frac{1}{\text{ARIV}}\right)^2}{2} \tag{8.25}$$

The D_1 statistic uses a total parameter covariance matrix based on the average relative increase in variance. This formulation of the test statistics assumes that the relative increase in variance (or equivalently, the fraction of missing information) is the same for all parameters (i.e., ARIV is representative of each parameter's RIV value). This assumption is unlikely to hold in practice because it essentially requires that the analysis variables have the same missing data rates and the same correlations. Li, Raghunathan, et al. (1991) used Monte Carlo simulations to study the performance of the D_1 statistic under a variety of different conditions. Their simulation results suggest that D_1 has type I error rates close to the nominal 0.05 level, but it lacks power when the number of parameters is large or the number of imputations is small. For example, they show that an analysis that uses $m = 4$ imputations has approximately 10% less power than a hypothetical analysis based on an infinite number of imputations. The authors only report power levels for $m = 4$ imputations, but it is reasonable to expect power to improve as the number of imputations increases. As a final note, the derivation of D_1 assumes a very large sample size, but no research to date has investigated its performance in small to moderate samples. Consequently, it is difficult to assess the trustworthiness of the D_1 statistic in realistic research scenarios.

An Analysis Example

To illustrate the D_1 statistic, suppose that it is of interest to use the data in Table 8.1 to estimate the regression of job performance on IQ and psychological well-being. After generating

20 imputations, I fit an ordinary least squares regression model to each data set and saved the estimates and parameter covariance matrices to a file for further analysis. In a multiple regression analysis, researchers typically use an omnibus *F* test to determine whether two or more coefficients are significantly different from zero, and the D_1 statistic can serve a similar role in a multiple imputation analysis. Table 8.4 shows the parameter estimates and the parameter covariance matrices from the 20 analyses. Note that I excluded the regression intercept and its covariance matrix elements from the table because the intercept is not part of the usual omnibus test. Consequently, the diagonal elements of each parameter covariance matrix contain the sampling variances (i.e., squared standard errors), and the off-diagonal is the covariance between the two regression slopes.

To begin, averaging the 20 sets of regression coefficients gives the following vector of point estimates.

$$\bar{\boldsymbol{\theta}} = \begin{bmatrix} \bar{\beta}_{IQ} \\ \bar{\beta}_{WB} \end{bmatrix} = \begin{bmatrix} .083 \\ .365 \end{bmatrix}$$

The interpretation of the regression coefficients is identical to that of a complete-data analysis. For example, holding psychological well-being constant, a one-point increase in IQ is associated with a 0.083 increase in job performance ratings, on average.

Next, I computed the pooled parameter covariance matrix. Averaging the covariance matrices in Table 8.4 yields the pooled within-imputation covariance matrix.

$$\mathbf{V}_W = \begin{bmatrix} .00159 & -.00176 \\ -.00176 & .02708 \end{bmatrix}$$

Again, $\mathbf{V}_W$ estimates the parameter covariance matrix that would have resulted had there been no missing data. Next, I used the *m* sets of regression coefficients and the corresponding pooled values to compute the between-imputation covariance matrix that quantifies the additional sampling fluctuation that accrues from the missing data.

$$\mathbf{V}_B = \begin{bmatrix} .00689 & -.01723 \\ -.01723 & .11446 \end{bmatrix}$$

Computing the total covariance matrix requires the average relative increase in variance. Substituting the previous estimates of $\mathbf{V}_W$ and $\mathbf{V}_B$ into Equation 8.21 gives ARIV = 4.042. This value suggests that the sampling variance due to the missing data is, on average, four times larger than the sampling variance that would have resulted had the data been complete. Next, substituting ARIV and $\mathbf{V}_W$ into Equation 8.20 gives the total parameter covariance matrix as follows:

$$\tilde{\mathbf{V}}_T = \begin{bmatrix} .00802 & -.00889 \\ -.00889 & .13654 \end{bmatrix}$$

TABLE 8.4. Coefficients and Parameter Covariance Matrices from the Multiple Regression Example

Imputation	Estimate		Covariance matrix		Imputation	Estimate		Covariance matrix	
1	$\hat{\beta}_{IQ}$	0.12714	0.00153	–0.00183	11	$\hat{\beta}_{IQ}$	0.16918	0.00137	–0.00164
	$\hat{\beta}_{WB}$	–0.01162	–0.00183	0.02521		$\hat{\beta}_{WB}$	0.19610	–0.00164	0.01861
2	$\hat{\beta}_{IQ}$	0.06811	0.00105	–0.00067	12	$\hat{\beta}_{IQ}$	0.13184	0.00137	–0.00119
	$\hat{\beta}_{WB}$	0.66786	–0.00067	0.01670		$\hat{\beta}_{WB}$	0.32367	–0.00119	0.02498
3	$\hat{\beta}_{IQ}$	–0.02464	0.00383	–0.00448	13	$\hat{\beta}_{IQ}$	0.23750	0.00152	–0.00200
	$\hat{\beta}_{WB}$	1.06036	–0.00448	0.05952		$\hat{\beta}_{WB}$	0.02645	–0.00200	0.02741
4	$\hat{\beta}_{IQ}$	0.06189	0.00126	–0.00099	14	$\hat{\beta}_{IQ}$	0.02960	0.00226	–0.00250
	$\hat{\beta}_{WB}$	0.16877	–0.00099	0.02218		$\hat{\beta}_{WB}$	0.29091	–0.00250	0.04341
5	$\hat{\beta}_{IQ}$	0.16090	0.00172	–0.00206	15	$\hat{\beta}_{IQ}$	0.04459	0.00121	–0.00149
	$\hat{\beta}_{WB}$	0.09261	–0.00206	0.03131		$\hat{\beta}_{WB}$	0.76175	–0.00149	0.01943
6	$\hat{\beta}_{IQ}$	–0.00272	0.00183	–0.00118	16	$\hat{\beta}_{IQ}$	0.04217	0.00161	–0.00210
	$\hat{\beta}_{WB}$	0.82762	–0.00118	0.03015		$\hat{\beta}_{WB}$	0.39766	–0.00210	0.02830
7	$\hat{\beta}_{IQ}$	0.22419	0.00220	–0.00296	17	$\hat{\beta}_{IQ}$	0.18720	0.00112	–0.00097
	$\hat{\beta}_{WB}$	–0.24891	–0.00296	0.04012		$\hat{\beta}_{WB}$	0.34958	–0.00097	0.01684
8	$\hat{\beta}_{IQ}$	0.03216	0.00159	–0.00199	18	$\hat{\beta}_{IQ}$	–0.07194	0.00163	–0.00234
	$\hat{\beta}_{WB}$	–0.08738	–0.00199	0.02956		$\hat{\beta}_{WB}$	0.56429	–0.00234	0.02812
9	$\hat{\beta}_{IQ}$	0.03843	0.00112	–0.00114	19	$\hat{\beta}_{IQ}$	0.07507	0.00109	–0.00086
	$\hat{\beta}_{WB}$	0.26239	–0.00114	0.02068		$\hat{\beta}_{WB}$	0.37283	–0.00086	0.01653
10	$\hat{\beta}_{IQ}$	0.09957	0.00139	–0.00197	20	$\hat{\beta}_{IQ}$	0.02076	0.00113	–0.00089
	$\hat{\beta}_{WB}$	0.77943	–0.00197	0.02568		$\hat{\beta}_{WB}$	0.50011	–0.00089	0.01690

Finally, substituting the parameter estimates and the total covariance matrix into Equation 8.23 yields $D_1 = 1.245$, as follows:

$$D_1 = \frac{1}{2}\left(\begin{bmatrix}.083\\.365\end{bmatrix} - \begin{bmatrix}0\\0\end{bmatrix}\right)^T \begin{bmatrix}.00802 & -.00889\\-.00889 & .13654\end{bmatrix}^{-1}\left(\begin{bmatrix}.083\\.365\end{bmatrix} - \begin{bmatrix}0\\0\end{bmatrix}\right) = 1.45$$

Referencing D_1 against an F distribution with $k = 2$ and $\nu_2 = 55.806$ degrees of freedom returns a probability value of $p = .30$. D_1 is analogous to an omnibus F statistic, so the lack of significance suggests that the pair of regression coefficients is not statistically different from zero (i.e., considered as a set, the explanatory variables do not predict job performance). Fortunately, the D_1 statistic is available in a number of software programs, so performing the tedious matrix computations is rarely necessary.

8.12 TESTING MULTIPLE PARAMETERS BY COMBINING WALD TESTS

A second approach for conducting multiparameter significance tests is to pool significance tests from the analysis phase. Li, Meng, Raghunathan, and Rubin (1991) outlined a procedure for pooling Wald tests, which I henceforth refer to as the D_2 statistic. To begin, D_2 requires the arithmetic average of the m Wald tests, as follows:

$$\bar{\omega} = \frac{1}{m}\sum_{t=1}^{m}\omega_t \tag{8.26}$$

where ω_t is the Wald statistic from data set t and $\bar{\omega}$ is the mean test statistic. Similar to the D_1 statistic, D_2 also requires an estimate of the average relative increase in variance. Li, Meng, et al. (1991) provide an expression that relies only on the m Wald statistics

$$\text{ARIV}_1 = (1 + m^{-1})\left[\frac{1}{m-1}\sum_{t=1}^{m}\left(\sqrt{\omega_t} - \overline{\sqrt{\omega}}\right)^2\right] \tag{8.27}$$

where $\sqrt{\omega_t}$ is the square root of the Wald statistic from data set t, and $\overline{\sqrt{\omega}}$ is the average of the $\sqrt{\omega_t}$ values. (Collectively, the terms in brackets quantify the variance of the square root of the Wald statistics.) Although it does not resemble its previous counterpart, ARIV_1 has the same interpretation as ARIV. Finally, the D_2 statistic is as follows:

$$D_2 = \frac{\bar{\omega}k^{-1} - (m+1)(m-1)^{-1}\text{ARIV}_1}{1 + \text{ARIV}_1} \tag{8.28}$$

To generate a probability value, Li, Meng, et al. recommend an F reference distribution with k numerator degrees of freedom and ν_3 denominator degrees of freedom, where

$$\nu_3 = k^{-3/m}(m-1)\left(1 + \frac{1}{\text{ARIV}_1}\right)^2 \tag{8.29}$$

The interpretation of D_2 is similar to that of a complete-data Wald statistic. That is, a statistically significant test statistic indicates that the parameter estimates differ from their hypothesized values.

Li, Meng, et al. (1991) used Monte Carlo simulations to study the performance of D_2 statistic under a variety of conditions. Their results suggest that type I error rates can either be too high or too low, depending on the fraction of missing information (e.g., when the fraction of missing information was less than 20%, type I errors dropped below the nominal 0.05 level). Their simulations also indicate that D_2 has lower power than D_1. Considered as a whole, these simulation results suggest that D_2 does not yield accurate inferences, and the authors recommend using the procedure "primarily as a screening test statistic" (p. 83). You should use the D_1 statistic whenever possible, but a custom program for computing D_2 is available on the companion website, if necessary.

8.13 TESTING MULTIPLE PARAMETERS BY COMBINING LIKELIHOOD RATIO STATISTICS

A final option for conducting multiparameter significance tests is to combine likelihood ratio test statistics from the analysis phase. Meng and Rubin (1992) outline such a procedure, and I subsequently refer to their test statistic as D_3. As a brief reminder, recall that the likelihood ratio statistic uses the log-likelihood value to compare the relative fit of two nested models, as follows:

$$\text{LR} = -2(\log L_{\text{Restricted}} - \log L_{\text{Full}}) \tag{8.30}$$

where $\log L_{\text{Full}}$ and $\log L_{\text{Restricted}}$ are the log-likelihood values from the full and the restricted models, respectively. The restricted model may include a subset of the parameters from the full model (e.g., a regression model where the slopes are constrained to zero during estimation), or it can differ from the full model by a set of complex parameter constraints (e.g., a confirmatory factor analysis model is a restricted model that expresses the population covariance matrix as a function of the factor model parameters).

To begin, the D_3 requires the average likelihood ratio test from the analysis phase, as follows:

$$\overline{\text{LR}} = \frac{1}{m}\sum_{t=1}^{m} \text{LR}_t \tag{8.31}$$

where $\overline{\text{LR}}$ is the arithmetic average of the m likelihood ratio statistics, and LR_t is the likelihood ratio test from data set t. The computations also require the pooled parameter estimates from both models. I denote these parameter vectors as $\bar{\boldsymbol{\theta}}_{\text{F}}$ and $\bar{\boldsymbol{\theta}}_{\text{R}}$ for the full and the restricted models, respectively.

After pooling the test statistics and the parameter estimates, the next step is to re-estimate the full and the restricted models, this time constraining the model parameters to their pooled values (i.e., estimate the full model m times, each time fixing the model parameters to the

values in $\bar{\boldsymbol{\theta}}_F$). Estimating the models with parameter constraints yields a second set of m likelihood ratio tests that compare the relative fit of the constrained models. The purpose of this step is to obtain the arithmetic average of these likelihood ratio tests (e.g., by substituting the LR values into Equation 8.31). I denote this average as $\overline{\text{LR}}_{\text{Constrained}}$ in order to differentiate it from $\overline{\text{LR}}$.

Finally, the D_3 test statistic is as follows:

$$D_3 = \frac{\overline{\text{LR}}_{\text{Constrained}}}{k(1 + \text{ARIV}_2)} \tag{8.32}$$

where ARIV_2 is yet another estimate of the average relative increase in variance.

$$\text{ARIV}_2 = \frac{m + 1}{k(m - 1)}(\overline{\text{LR}} - \overline{\text{LR}}_{\text{Constrained}}) \tag{8.33}$$

To obtain a probability value for D_3, Meng and Rubin recommend an F reference distribution with k numerator degrees of freedom and ν_4 denominator degrees of freedom. In this context, k is the number of parameter constraints (i.e., the degrees of freedom for the complete-data likelihood ratio test), and ν_4 is

$$\nu_4 = 4 + (km - k - 4)\left[1 + \left(1 - \frac{2}{km - k}\right)\frac{1}{\text{ARIV}_2}\right]^2 \tag{8.34}$$

In the situation where km - k is less than or equal to four, Meng and Rubin recommend an alternate expression for ν_4, as follows.

$$\nu_4 = \frac{(km - k)\ (1\ + k^{-1})\left(1 + \dfrac{1}{\text{ARIV}_3}\right)^2}{2} \tag{8.35}$$

Meng and Rubin show that D_3 is asymptotically equivalent to D_1, so the two tests should yield similar conclusions in large samples. However, because virtually no research studies have compared the two test statistics, it is difficult to assess their relative performance in realistic research scenarios. All things being equal, D_1 is more convenient because it is readily available in a number of popular software programs. However, D_3 is potentially useful in structural equation modeling analyses because it provides a mechanism for assessing model fit (e.g., by pooling the chi-square tests of model fit). In the structural equation modeling context, the full model is a saturated model (e.g., a model that estimates the sample covariance matrix), and the restricted model is the hypothesized model (e.g., a confirmatory factor analysis model that expresses the population covariance matrix as a function of the factor model parameters). The so-called chi-square test of model fit is a likelihood ratio test that compares the relative fit of these two models. Methodologists have yet to develop procedures for pooling structural equation modeling fit indices (e.g., the CFI, RMSEA), so the D_3 statistic

is currently the only formal option for assessing fit. I illustrate the use of D_3 for this purpose in one of the subsequent data analysis examples.

8.14 DATA ANALYSIS EXAMPLE 1

In the remainder of the chapter, I use three data analysis examples to illustrate various aspects of a multiple imputation analysis. Chapter 7 did not include data analysis examples, so the subsequent examples illustrate all three phases of a multiple imputation analysis. To facilitate comparisons between maximum likelihood estimation and multiple imputation, the analysis examples are identical to those from Chapter 4.

The first analysis example illustrates the use of multiple imputation to estimate a mean vector, covariance matrix, and a correlation matrix.* The data for this analysis are made up of scores from 480 employees on eight work-related variables: gender, age, job tenure, IQ, psychological well-being, job satisfaction, job performance, and turnover intentions. I generated these data to mimic the correlation structure of published research articles in the management and psychology literature (e.g., Wright & Bonett, 2007; Wright, Cropanzano, & Bonett, 2007). The data have three missing data patterns, each of which is comprised of 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).

The Imputation Phase

First, I used the EM algorithm to estimate the mean vector and the covariance matrix. EM converged in only 20 iterations, which suggests that the data augmentation algorithm should also converge very quickly. Next, I generated an exploratory chain of 5,000 data augmentation cycles and saved the simulated parameter values from each P-step. The purpose of this initial analysis was to assess the convergence of the data augmentation algorithm, and I did so by examining time-series and autocorrelation function plots for each element in the mean vector and the covariance matrix.

To illustrate the convergence diagnostics, Figure 8.2 shows the times-series and autocorrelation function plots for the simulated covariance between well-being and job satisfaction. I paid particularly close attention to the convergence behavior of this parameter because it has the highest percentage of missing data, and thus one of the highest fractions of missing information (only 33% of the cases have data on both variables). The time-series plot in the top panel of Figure 8.2 suggests that the covariances randomly vary, with no discernible long-term trends. In fact, the upward and downward trends in the plot typically last for fewer than 20 iterations. The autocorrelation plot in the bottom panel of the figure also shows very fast convergence, as the autocorrelations drop to chance levels by lag 10 (i.e., the correlation between parameter values separated by 10 iterations is not significantly different from zero). I examined the plots for the remaining parameters, and they were largely consistent with

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

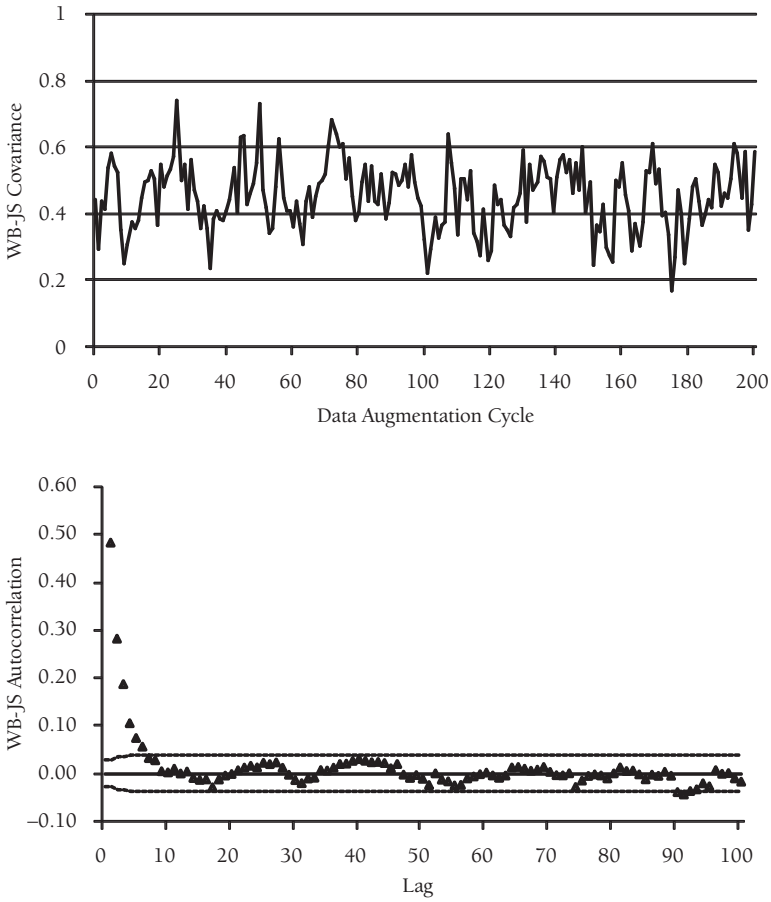

FIGURE 8.2. Time-series plots for the covariance between psychological well-being (WB) and job satisfaction (JS). The top panel shows a time-series plot with no long-term trends. The bottom panel shows autocorrelations that drop to within sampling error of zero after 10 data augmentation cycles.

those in Figure 8.2. Taken together, the graphical diagnostics suggest that the data augmentation algorithm converges very quickly, perhaps in fewer than 20 iterations. The fast convergence may seem somewhat surprising given that such a large proportion of the well-being and job satisfaction scores were missing. However, this example is an ideal situation because the data are MCAR by design.

As a general rule, it is a good idea to assess convergence using a small number of alternate starting values (e.g., bootstrap estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$). However, the graphical displays were so ideal that this additional step did not seem necessary. Consequently, I generated the final imputations using a single data augmentation chain. The graphical diagnostics suggest that the data augmentation algorithm converges in fewer than 20 iterations, but I took a conservative tack of specifying 100 burn-in and 100 between-imputation iterations (i.e., I saved the first imputed data set after an initial burn-in period of 100 cycles and saved subsequent data sets at every 100th I-step thereafter). The exploratory data augmentation chain took just a few seconds to run, so I opted to generate $m = 50$ imputations for the analysis

phase. Estimating means and correlations from 50 data sets takes very little time, so using a large number of imputations posed no practical problems.

The Analysis Phase

I analyzed each of the 50 data sets in the analysis phase. This step produced an estimate of the mean vector, the covariance matrix, and the correlation matrix from each of the 50 filled-in data sets. Although it sounds tedious to repeat the analysis that many times, many software programs automate the process. As an aside, programs that automate the analysis phase have different formatting requirements for the imputed data files. For example, some software packages make it very easy to analyze a data set where the imputations are stacked in a single file, whereas other programs require separate data sets. The companion website has software examples that illustrate both approaches.

The Pooling Phase

In the pooling phase, I used Rubin's (1987) formulas to combine the parameter estimates. Although some of the parameters are unlikely to satisfy the normality requirement (e.g., variances and covariances), I averaged the variable means and the covariance matrix elements without applying any transformations. Fisher's (1915) z transformation is a natural choice for the pooling correlations because it transforms the estimates to a metric that more closely approximates a normal distribution, and it provides a straightforward mechanism for performing significance tests. Equation 8.2 gives the transformation, and the corresponding standard error is as follows:

$$SE_t = \frac{1}{\sqrt{N-3}} \tag{8.36}$$

After pooling the transformed estimates and their standard errors, I used Equation 8.3 to back-transform the average coefficients to the correlation metric.

Table 8.5 shows the pooled point estimates along with the corresponding maximum likelihood estimates from Chapter 4. As seen in the table, the multiple imputation and maximum likelihood estimates are quite similar. The close correspondence of the two sets of estimates is not surprising given that both techniques make the same assumptions (MAR data and multivariate normality) and use the same set of input variables. You might have noticed that maximum likelihood estimates of variances and covariances are slightly smaller than those of multiple imputation, even for the variables that have complete data (e.g., the age variance estimates are 29.968 and 28.908 for multiple imputation and maximum likelihood, respectively). These systematic (albeit small) differences result from the fact that maximum likelihood estimates use N in the denominator rather than $N - 1$.

TABLE 8.5. Mean, Covariance, and Correlation Estimates from Data Analysis Example 1

Variable	1	2	3	4	5	6	7	8
				Multiple imputation				
1: Age	28.968	**0.504**	**–0.010**	**.181**	**0.139**	**–0.049**	**–0.150**	**0.015**
2: Tenure	8.477	9.755	**–0.034**	**.156**	**0.153**	**0.016**	**0.011**	**0.001**
3: Female	–0.028	–0.052	0.249	**.113**	**0.038**	**–0.015**	**0.005**	**0.068**
4: Well-being	1.147	0.576	0.066	1.395	**0.321**	**0.456**	**–0.255**	**0.293**
5: Satisfaction	0.888	0.567	0.023	0.449	1.406	**0.184**	**–0.234**	**0.407**
6: Performance	–0.331	0.061	–0.009	0.675	0.274	1.574	**–.346**	**0.426**
7: Turnover	–0.378	0.016	0.001	–0.141	–0.129	–0.203	0.218	**–0.180**
8: IQ	0.675	0.026	0.285	2.912	4.063	4.505	–0.707	71.040
Means	37.948	10.054	0.542	6.291	5.946	6.021	0.321	100.102
				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

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

8.15 DATA ANALYSIS EXAMPLE 2

The second analysis example applies multiple imputation to 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:

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

I reused the 50 imputations from the previous example for this analysis. Carefully planning the imputation model allows you to use the same imputed data sets for many (if not all) of the subsequent analyses. At a minimum, the imputation phase must include all of the associations that are of interest in the subsequent analysis phase. I imputed the data using all eight variables in the data set, so I can perform any analysis that involves the zero-order associations among the variables. I would only need to generate a new set of imputations if my analysis model included higher-order terms (e.g., interactions) or other variables that I excluded from the imputation phase.

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

The Analysis and Pooling Phases

In the analysis phase, I estimated the regression model parameters separately for each of the 50 filled-in data sets. The imputation phase incorporated a number of extra variables that were not part of the regression analysis (i.e., age, job tenure, gender, IQ), so these additional variables effectively served as auxiliary variables. It is important to reiterate that auxiliary variables play no role in the analysis phase (the filled-in values already contain the auxiliary information), so I did not include the extra variables in the regression model.

In a multiple regression analysis, researchers typically use an omnibus *F* test to determine whether two or more coefficients are statistically different from zero. Of the three multiparameter significance tests outlined previously in the chapter, the D_1 statistic is particularly convenient because it is readily available in multiple imputation software programs. Consequently, I used D_1 to assess whether the two regression slopes were different from zero. This procedure produced a test statistic of $D_1 = 42.87$, and referencing this value against an *F* distribution with $k = 2$ and $\nu_2 = 899.07$ degrees of freedom returned a probability value of $p < .001$. The substantive interpretation of D_1 is identical to that of an omnibus *F* statistic, so rejecting the null hypothesis implies that at least one of the regression coefficients is significantly different from zero.

As an aside, the D_1 statistic assumes that the fractions of missing information are identical across parameters. Multiple imputation software programs generally report these quantities, and the estimates from this analysis are 0.27 and 0.39 for the well-being and job satisfaction slopes, respectively. Recall that missing information is akin to an R^2 statistic, such that a value of 0.27 indicates that 27% of the well-being slope's sampling variance (i.e., squared standard error) is attributable to missing data. Although the fractions of missing information are not identical, the magnitude of this difference is probably not large enough to seriously distort the D_1 statistic (Li, Raghunathan, et al., 1991). I could have also used the D_2 or D_3 statistics to test the regression coefficients, but D_1 is far easier to implement.

Researchers typically follow up a significant omnibus test by examining the partial regression coefficients. Table 8.6 gives the regression model estimates along with the saturated correlates model estimates from Chapter 5. As seen in the table, psychological well-being was a significant predictor of job performance, $\hat{\beta}_1 = 0.470$, $t(231.01) = 8.79$, $p < .001$, but job satisfaction was not, $\hat{\beta}_2 = 0.045$, $t(154.84) = 0.77$, $p = .44$. The interpretation of these regression coefficients is the same as an ordinary least squares analysis. For example, holding job satisfaction constant, a one-point increase in psychological well-being yields a .470 increase in job performance ratings, on average. Note that I used Barnard and Rubin's (1999) degrees of freedom for the *t* tests. This degrees of freedom expression relies, in part, on the degrees of freedom for a complete-data test statistic (e.g., $df_{com} = N - k - 1 = 477$, where k is the number of predictors). I point this out because some multiple imputation software programs require the user to specify the complete-data degrees of freedom value when requesting Barnard and Rubin's formula.

Finally, notice that multiple imputation and maximum likelihood produced very similar parameter estimates and standard errors. In this particular example, the two missing data handling approaches are not exactly comparable because the saturated correlates model in Chapter 5 included only IQ and turnover intentions as auxiliary variables. Nevertheless, the

TABLE 8.6. Regression Model Estimates from Data Analysis Example 2

Parameter	Estimate	*SE*	*t*
	Multiple imputation		
β_0 (intercept)	6.021	0.060	118.096
β_1 (well-being)	0.470	0.053	8.791
β_2 (satisfaction)	0.045	0.058	0.772
R^2	.208		
	Maximum likelihood		
β_0 (intercept)	6.020	0.053	114.642
β_1 (well-being)	0.475	0.054	8.798
β_2 (satisfaction)	0.035	0.058	0.605
R^2	.208		

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

estimates are quite similar, even though the multiple imputation analysis used a larger set of auxiliary variables.

8.16 DATA ANALYSIS EXAMPLE 3

The final data analysis example applies multiple imputation to a confirmatory factor analysis model.* The analyses use artificial data from a questionnaire on eating disorder risk. Briefly, the data contain the responses from 400 college-age 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 eating when I'm hungry") and Food Preoccupation (e.g., "I find myself preoccupied with food"), and mimic the two-factor structure proposed by Doninger, Enders, and Burnett (2005). The 10 questionnaire items combine to measure two constructs. The Drive for Thinness scale consists of seven items (EAT_1, EAT_2, EAT_{10}, EAT_{11}, EAT_{12}, EAT_{14}, and EAT_{24}), and the Food Preoccupation scale has three items (EAT_3, EAT_{18}, and EAT_{21}). Figure 4.2 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},

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

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 where 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.

The Imputation Phase

To get a rough gauge of convergence speed, I first used the EM algorithm to estimate the mean vector and the covariance matrix for the entire set of 13 variables (the 10 EAT items, body mass index, anxiety, and Western standard of beauty). EM converged in only nine iterations, which suggests that data augmentation should also converge very quickly. Next, I generated an exploratory chain of 5,000 data augmentation cycles and saved the simulated parameter estimates from each P-step. The purpose of this initial analysis was to assess the convergence of the data augmentation algorithm, and I did so by examining time-series and autocorrelation function plots for each simulated parameter value in the mean vector and the covariance matrix. For the sake of brevity, I illustrate these plots using the covariance between EAT_1 and EAT_{18}. I chose this parameter because it has one of the highest fractions of missing information (i.e., this pair of variables has one of the highest missing data rates and the lowest correlations). The fraction of missing information largely dictates convergence speed, so this parameter should be among the slowest to converge.

Figure 8.3 shows the time-series and autocorrelation function plots for the covariance between EAT_1 and EAT_{18}. The time-series plot in the top panel of Figure 8.3 suggests that the simulated covariance values randomly vary with no discernible trends whatsoever. The autocorrelation plot in the bottom panel shows that the autocorrelations drop to chance levels by the second lag (i.e., the correlation between parameter values separated by two iterations is not significantly different from zero). Thus, this parameter appears to converge almost immediately. I examined the plots for the remaining parameters, and they were largely consistent with those in Figure 8.3. Taken together, the graphical diagnostics indicate that the data augmentation algorithm converges very quickly, perhaps in fewer than 10 iterations. The rather fast convergence follows from the fact that the fractions of missing information were generally rather low (e.g., values between 2 and 10% were common).

Having established the convergence of the data augmentation algorithm, I used a single data augmentation chain to generate the imputations. Although the data augmentation algorithm appears to converge in fewer than 10 iterations, I took a conservative approach and specified 100 burn-in iterations and 100 between-imputation iterations (i.e., i.e., I saved the first imputed data set after an initial burn-in period of 100 cycles and saved subsequent data sets at every 100th I-step thereafter). For this analysis, I created $m = 100$ imputations for the analysis phase. Because a confirmatory factor analysis model takes very little time to estimate, using a large number of imputations does not pose a computational burden. Analyzing a large number of data sets is also useful for assessing model fit (more on this later). Finally, note that I used the entire set of 13 variables in the imputation phase. The factor analysis model includes the 10 questionnaire items, so the additional variables (body mass index,

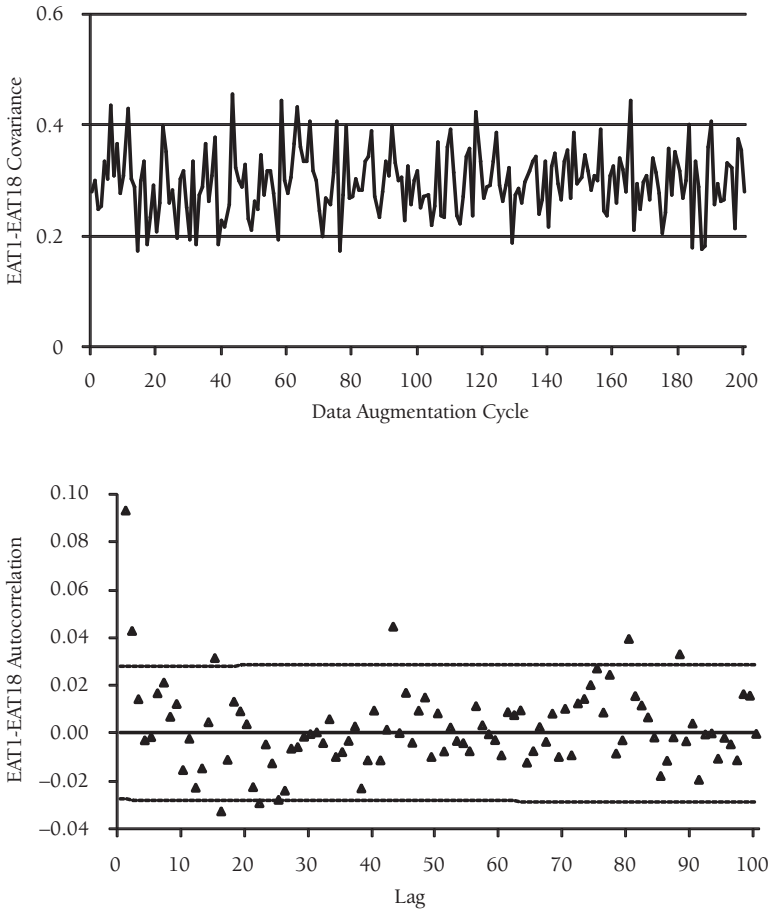

FIGURE 8.3. Time-series plots for the covariance between questions 1 and 18 from the EAT questionnaire (EAT_1 and EAT_{18}, respectively). The top panel shows a time-series plot with no long-term trends. The bottom panel shows autocorrelations that drop to within sampling error of zero by the second data augmentation cycle.

anxiety, and Western standards of beauty) effectively served as auxiliary variables. Again, there is no need to use the auxiliary variables in the subsequent analysis phase.

The Analysis and Pooling Phases

In the analysis phase, I estimated the factor model parameters separately for each of the 100 filled-in data sets. The discrete nature of the questionnaire items violates the multivariate normality assumption, so I used robust (i.e., sandwich estimator) standard errors for each analysis (see Chapter 5). The analysis step produced 100 sets of results, and I subsequently used Rubin's (1987) formulas to combine the parameter estimates and the standard errors (pooling robust standard errors is no different from pooling normal-theory standard errors). Although some of the parameters are unlikely to satisfy the normality requirement (e.g., factor variances, residual variances), I averaged the estimates without applying any transformations.

TABLE 8.7. Confirmatory Factor Analysis Estimates from Data Analysis Example 3

	Loadings		Intercepts		Residuals	
Variable	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*
			Multiple imputation			
EAT_1	0.743	0.049	4.006	0.055	0.606	0.067
EAT_2	0.651	0.050	3.937	0.050	0.536	0.053
EAT_{10}	0.808	0.052	3.955	0.050	0.331	0.038
EAT_{11}	0.765	0.049	3.937	0.047	0.299	0.027
EAT_{12}	0.665	0.054	3.929	0.051	0.540	0.057
EAT_{14}	0.900	0.048	3.962	0.051	0.237	0.028
EAT_{24}	0.625	0.053	3.985	0.051	0.604	0.050
EAT_3	0.774	0.052	3.967	0.050	0.413	0.043
EAT_{18}	0.749	0.055	3.982	0.052	0.456	0.049
EAT_{21}	0.859	0.052	3.950	0.051	0.270	0.043
			Maximum likelihood			
EAT_1	0.741	0.049	4.004	0.055	0.604	0.069
EAT_2	0.649	0.050	3.937	0.050	0.535	0.054
EAT_{10}	0.808	0.052	3.953	0.050	0.328	0.039
EAT_{11}	0.764	0.049	3.938	0.047	0.300	0.027
EAT_{12}	0.662	0.055	3.929	0.051	0.538	0.058
EAT_{14}	0.901	0.047	3.963	0.051	0.234	0.028
EAT_{24}	0.622	0.053	3.986	0.051	0.599	0.049
EAT_3	0.772	0.052	3.967	0.050	0.415	0.042
EAT_{18}	0.751	0.056	3.982	0.052	0.451	0.050
EAT_{21}	0.862	0.053	3.952	0.051	0.264	0.043

Repeating the factor analysis 100 times sounds incredibly tedious, but some structural equation modeling software programs can fully automate the analysis and pooling phases. In fact, estimating the models and combining the results took less than 10 seconds on a laptop computer.

Table 8.7 shows selected parameter estimates and standard errors, along with the corresponding maximum likelihood estimates. To maximize the comparability of the two sets of results, the table gives the saturated correlates estimates from Chapter 5. As seen in the table, multiple imputation and maximum likelihood produced nearly identical estimates and standard errors. Again, this is not a surprise because the two procedures used the same set of variables (i.e., the saturated correlates model included the same 13 variables that I used in the imputation phase). Consistent with the previous analyses, the interpretation of the model parameters is unaffected by the missing data handling procedure. For example, the factor loadings estimate the expected change in the questionnaire items for a one-standard-deviation increase in the latent construct. This interpretation follows from the fact that I fixed the variances of the latent variables to unity in order to identify the model.

Assessing model fit is an important part of a structural equation modeling analysis. Earlier in the chapter, I outlined a D_3 statistic that combines likelihood ratio tests from a mul-

tiple imputation analysis (Meng & Rubin, 1992). This procedure is potentially useful for structural equation modeling analyses because it provides a mechanism for assessing model fit (e.g., by pooling the chi-square test of model fit). In the context of a confirmatory factor analysis, the saturated model serves as the full model, and the hypothesized factor model is the restricted model. The so-called chi-square test of model fit is a likelihood ratio test that compares the relative fit of these two models.

To illustrate the D_3 statistic, I fit the confirmatory factor model and the saturated model to each imputed data set and saved the resulting likelihood ratio tests. This step is straightforward because structural equation modeling programs report the likelihood ratio (i.e., chi-square) test as standard output. Averaging the likelihood ratio tests produced $\overline{LR} = 61.94$. In the next step, I re-estimated the two models after constraining the parameters to their pooled values. For example, I estimated the two-factor model on each imputed data set, but did so by constraining the factor model parameters to the pooled estimates in Table 8.7. I applied the same procedure to the saturated model. Estimating the constrained models produced another set of 100 likelihood ratio tests, the average of which was $\overline{LR}_{Constrained} = 56.93$. The D_3 statistic requires the average relative increase in variance, and substituting the appropriate quantities into Equation 8.33 gives $ARIV_2 = 0.15$. (The factor model has 34 fewer parameters than the saturated model, so $k = 34$.) Finally, substituting the appropriate values into Equation 8.32, a test statistic of $D_3 = 1.456$, and referencing this value to an F distribution with $k = 34$ and $\nu_4 = 197{,}410.74$ degrees of freedom gives a probability value of $p = .04$. Because the substantive interpretation of D_3 is identical to that of the likelihood ratio test, rejecting the null hypothesis implies that the factor model does not fit the data as well as the saturated model. For comparison purposes, the saturated correlates model from Chapter 5 produced a likelihood ratio test of $\chi^2(34) = 49.04$, $p = .05$. Although the two analyses produced very similar conclusions about model fit in this particular example, no studies have examined the performance of D_3 in structural equation modeling applications. Until more research accumulates, it seems prudent to interpret D_3 with some caution.

Researchers generally augment the likelihood ratio test with a number of other fit indices. The methodological literature currently favors the CFI, RMSEA, and the SRMR (Hu & Bentler, 1998, 1999), but there is no established method for pooling these indices. In order to get some sense about model fit, I used the 100 estimates of each index to construct an empirical distribution. The distributions were approximately normal and had means of 0.987 (CFI), 0.041 (RMSEA), and 0.031 (SRMR). I arbitrarily examined the 5th and the 95th percentiles of each index, and these values were as follows: CFI ($P_5 = 0.981$, $P_{95} = 0.993$), RMSEA ($P_5 = 0.032$, $P_{95} = 0.050$), and SRMR ($P_5 = 0.028$, $P_{95} = 0.034$). High CFI values are indicative of good model fit, so the CFI value at the 5th percentile of the distribution should provide a conservative assessment of fit. In contrast, lower values of the RMSEA and SRMR are indicative of good fit, so the values at the 95th percentile of these distributions would be conservative. Considered as a whole, the means and the percentiles of the distributions suggest that the two-factor model fits the data adequately (e.g., the values at the mean and the 5th percentile of the CFI distribution exceed the conventional cutoff of 0.95). The approach outlined here is purely ad hoc and has no theoretical rationale. Until methodologists develop formal pooling rules for popular fit indices, this is probably the best you can do.

8.17 SUMMARY

A multiple imputation analysis consists of three distinct steps: the imputation phase, the analysis phase, and the pooling phase. The product of the imputation phase is a set of filled-in data sets, each of which contains different estimates of the missing values. The purpose of the analysis phase is to analyze the filled-in data sets from the preceding imputation phase. This step consists of *m* statistical analyses, one for each imputed data set. The analysis phase yields several sets of parameter estimates and standard errors, so the goal of the pooling phase is to combine everything into a single set of results. Rubin (1987) outlined relatively straightforward formulas for pooling parameter estimates and standard errors. The pooled parameter estimate is simply the arithmetic average of the estimates from the analysis phase. Combining standard errors is somewhat more complex because it involves two sources of sampling variation. The within-imputation variance is the arithmetic average of the *m* sampling variances (i.e., squared standard errors), and the between-imputation variance quantifies the variability of an estimate across the *m* imputations. The within-imputation variance estimates the sampling fluctuation that would have resulted had there been no missing data, and the between-imputation variance captures the increase in sampling error due to missing data. Together, these two sources of variation combine to form the total sampling variance, the square root of which is the standard error.

The chapter outlined four significance testing procedures. The familiar *t* statistic (the pooled estimate divided by its standard errors) is useful for testing whether a single estimate is different from some hypothesized value. Multiple imputation also offers different mechanisms for testing a set of parameter estimates. The D_1 statistic uses pooled parameter estimates and pooled parameter covariance matrices to construct a test that closely resembles the multivariate Wald statistic. A second approach is to compute a significance test for each imputed data set and pool the resulting test statistics. The D_2 statistic pools Wald tests from the analysis phase, and the D_3 statistic pools likelihood ratio tests. Although these procedures accomplish the same task, they are not equally trustworthy, nor are they equally easy to implement. Relatively little is known about the performance of the multiparameter significance tests, but it is clear that D_1 and D_3 are preferable to D_2.

Chapter 9 outlines a number of practical issues that arise during the imputation phase of a multiple imputation analysis. Specifically, the chapter offers advice on dealing with convergence problems, non-normal data (including nominal and ordinal variables), interaction effects, and large multiple-item questionnaire data sets. The chapter also provides a brief overview of some alternative imputation algorithms that are appropriate for special types of data structures (e.g., mixtures of categorical and continuous variables, multilevel data).

8.18 RECOMMENDED READINGS

Allison, P. D. (2002). *Missing data*. Newbury Park, CA: Sage.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Hoboken, NJ: Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, *91*, 473–489.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York: Chapman.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*, *33*, 545–571.

Sinharay, S., Stern, H. S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods*, *6*, 317–329