9

Practical Issues in Multiple Imputation

9.1 CHAPTER OVERVIEW

Having outlined the technical and the procedural details of multiple imputation in Chapters 7 and 8, I now address a number of practical issues that can arise in a multiple imputation analysis. Chapter 7 outlined a few such practical problems (e.g., assessing convergence, choosing the number of between-imputation iterations, deciding which variables to include in the imputation model), but several others need to be considered. Specifically, this chapter offers advice on dealing with convergence problems, non-normal data (including nominal and ordinal variables), interactive effects, and large multiple-item questionnaire data sets. The chapter also gives 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). As you will see, this chapter is relatively applied in nature and is geared toward practical recommendations rather than toward technical issues. As an aside, many of the issues in this chapter have not been well studied in the methodological literature, so the practical guidelines that I offer are likely to change as additional methodological research accumulates.

9.2 DEALING WITH CONVERGENCE PROBLEMS

The data augmentation algorithm occasionally fails to converge, and it is useful to have some strategies for dealing with the problem. To illustrate a convergence problem, reconsider the small employee data set that I have been using throughout the book. First, I computed a binary employment status variable that denotes whether the company hired each applicant. Table 9.1 shows the resulting data. Next, I used the four variables in the table to generate 5,000 cycles of data augmentation. The fact that a preliminary EM analysis converged in only 25 iterations suggests that data augmentation should also converge very quickly, but graphical diagnostics suggested otherwise.

TABLE 9.1. Employee Selection Data Set

IQ	Psychological well-being	Job performance	Employment status
78	13	—	0
84	9	—	0
84	10	—	0
85	10	—	0
87	—	—	0
91	3	—	0
92	12	—	0
94	3	—	0
94	13	—	0
96	—	—	0
99	6	7	1
105	12	10	1
105	14	11	1
106	10	15	1
108	—	10	1
112	10	10	1
113	14	12	1
115	14	14	1
118	12	16	1
134	11	12	1

Figure 9.1 shows the time-series and autocorrelation function plots for the job performance mean. Two problems are apparent in the time-series plot: systematic trends that last for hundreds of iterations, and implausible parameter values (e.g., many of the simulated means fall outside the 1 to 20 score range). The autocorrelation function plot in the bottom panel of Figure 9.1 is also problematic and shows strong serial dependencies that persist for many cycles. In this example, data augmentation fails to converge because job performance scores are completely missing for the subsample of applicants that the company did not hire. Consequently, there is insufficient data to estimate the association between job performance ratings and the binary employment status variable. At first glance, this seems at odds with the fact that EM converged after only 25 iterations. However, EM's behavior is deceptive because alternate starting values produce a completely different solution. In reality, there is no way to identify a single set of parameter values that are most likely to have produced the observed data.

Convergence problems such as those in Figure 9.1 often occur because there is insufficient data to estimate certain parameters. In some situations, the lack of data results from including too many variables in the imputation phase. For example, when the number of variables exceeds the number of cases, the data contain linear dependencies that cause mathematical difficulties for regression-based imputation. Because missing values reduce the amount of information in a data set, convergence problems can occur even when the number of variables is much smaller than the number of cases. A peculiar missing data pattern can also lead to estimation difficulties and convergence failures. For example, the cohort-sequential

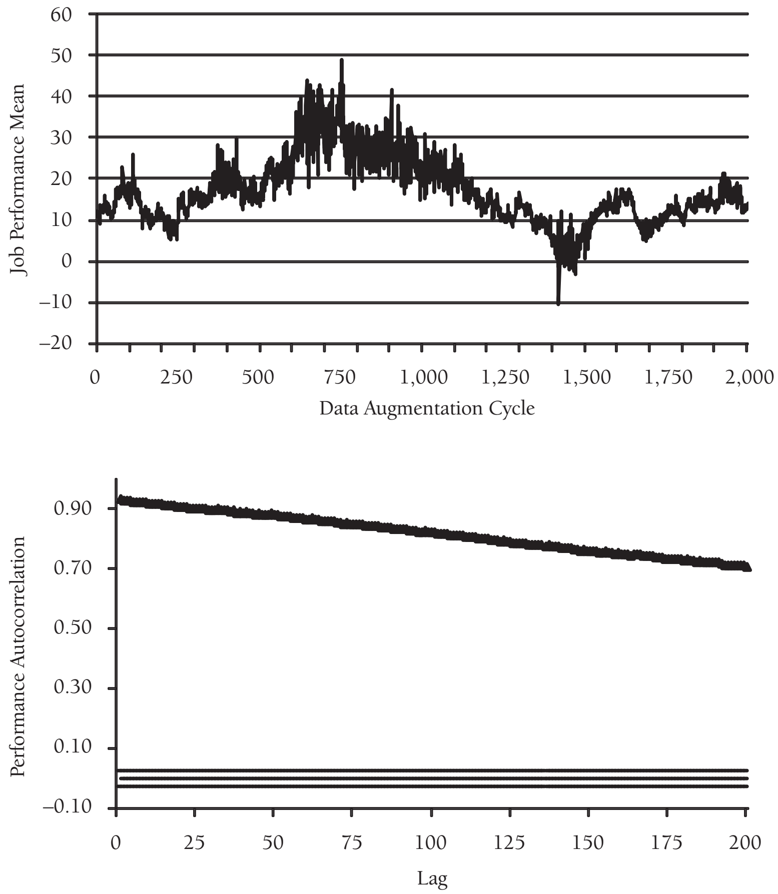

FIGURE 9.1. Time-series and autocorrelation function plot for parameters that do not converge. The top panel shows a time-series plot that exhibits systematic trends that last for hundreds of iterations and simulated parameter values that are outside of the plausible score range of 1 to 20. The bottom panel shows autocorrelations (denoted by a triangle symbol) that are close to $r = 0.70$ at lag-200.

design from Chapter 1 has variable pairs that are concurrently missing, making it impossible to estimate certain elements of the covariance matrix. The same is true for the data in Table 9.1.

The Ridge Prior Distribution

In some situations, reducing the number of variables or eliminating problematic variables is the only way to eliminate convergence problems. An alternate strategy is to use a so-called **ridge prior distribution** for the covariance matrix. The standard practice in a multiple imputation analysis is to adopt a noninformative prior distribution that carries no information about the mean vector and the covariance matrix. Consequently, the data alone define the posterior distributions of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ at each P-step. The ridge prior is a semi-informative distribution that contributes additional information about the covariance matrix. Conceptually, the ridge prior adds a small number of imaginary data records from a hypothetical population where the variables are uncorrelated. These additional data points can stabilize estimation and eliminate convergence problems, but they do so at the cost of introducing a slight bias to the simulated parameter values (and thus the imputations).

To illustrate the ridge prior, consider a hypothetical imputation model that consists of two variables and $N = 100$ cases. Furthermore, suppose the filled-in data from a particular I-step yields the following sample covariance matrix and sum of squares and cross products matrix.

$$\hat{\boldsymbol{\Sigma}}_t = \begin{bmatrix} 1.00 & .50 \\ .50 & 1.00 \end{bmatrix}$$

$$\hat{\boldsymbol{\Lambda}}_t = (N-1)\hat{\boldsymbol{\Sigma}}_t = \begin{bmatrix} 99.00 & 49.50 \\ 49.50 & 99.00 \end{bmatrix}$$

Recall from Chapter 7 that each P-step is a standalone Bayesian analysis that describes the posterior distributions and subsequently draws a new set of estimates of the mean vector and the covariance matrix from their distributions. With the standard noninformative prior, the posterior distribution of the covariance matrix is an inverse Wishart distribution, the shape of which depends on the filled-in data from the preceding I-step (i.e., the sample size and $\hat{\boldsymbol{\Lambda}}_t$).

The ridge prior is also an inverse Wishart distribution, but its shape depends on a degrees of freedom value and an estimate of the sum of squares and cross products matrix. (Collectively, these two parameters are the distribution's hyperparameters.) The sum of squares and cross products matrix for the prior is straightforward because it comes from a population covariance matrix with off-diagonal elements of zero and variances equal to those of the filled-in data. For example, the ridge covariance matrix for the previous bivariate example is as follows.

$$\boldsymbol{\Sigma}_t = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Notice that the variances are identical to those of the filled-in data, but the covariance is zero. Generating the sum of squares and cross products matrix for the prior requires a degrees of freedom value. The degrees of freedom value quantifies the number of "imaginary data points" that you assign to the prior and effectively determines the amount of influence that the prior exerts on the simulated parameter values. For example, assigning two degrees of freedom to the prior is akin to saying that an imaginary sample of two cases generated the previous covariance matrix. Doing so leads to the following sum of squares and cross products matrix:

$$\boldsymbol{\Lambda}_t = (df_{\mathrm{p}})\boldsymbol{\Sigma}_t = 2\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

where df_{p} is the degrees of freedom value for the prior.

After implementing a ridge prior, the pooled degrees of freedom (i.e., the degrees of freedom from the data plus the number of imaginary cases that you assign to the prior) and the pooled sum of squares and cross products matrix (i.e., the sum of $\mathbf{\Lambda}_t$ and $\hat{\mathbf{\Lambda}}_t$) define the shape of the posterior distribution, as follows.

$$p(\mathbf{\Sigma} \mid \hat{\boldsymbol{\mu}}, \mathbf{Y}) \sim W^{-1}(df_{\text{p}} + N - 1, [\mathbf{\Lambda}_t + \hat{\mathbf{\Lambda}}_t]) \tag{9.1}$$

Notice that the shape of the posterior distribution depends on the data and the additional information from the prior (e.g., the usual posterior distribution has $N - 1$ and $\hat{\mathbf{\Lambda}}_t$ as its parameter values). Conceptually, the ridge prior adds df_{p} imaginary data points from a population with uncorrelated variables. Altering the shape of the posterior distribution is the only change that occurs from implementing a ridge prior. Consistent with the description of data augmentation in Chapter 7, the P-step uses Monte Carlo simulation techniques to draw a new covariance matrix from the posterior, and the subsequent I-step uses these simulated parameters to construct a set of imputation regression equations.

The ridge prior eliminates convergence problems by increasing the effective sample size, but it attenuates the associations among the variables in the process. For example, pooling the degrees of freedom values and the sum of squares and cross products matrices from the bivariate example yields the following covariance matrix.

$$\hat{\mathbf{\Sigma}}_t = (df_{\text{p}} + \text{N} - 1)^{-1}(\mathbf{\Lambda}_t + \hat{\mathbf{\Lambda}}_t) = \frac{1}{2 + 99}\left(\begin{bmatrix} 2 & 0 \\ 1 & 2 \end{bmatrix} + \begin{bmatrix} 99.00 & 49.50 \\ 49.50 & 99.00 \end{bmatrix}\right) = \begin{bmatrix} 1.00 & .49 \\ .49 & 1.00 \end{bmatrix}$$

Notice that the covariance matrix has the same diagonal elements (i.e., variances) as the sample covariance matrix, but its off-diagonal elements are slightly smaller in magnitude. This follows from the fact that the prior distribution contributes two cases from a hypothetical population with uncorrelated variables. The imputation regression equations at the subsequent I-step depend on the parameter values from the P-step, so it makes intuitive sense that the imputations will also contain some bias. The magnitude of this bias depends on the number of data points that you assign to the prior, so you should try to minimize the prior distribution's degrees of freedom value. It is impossible to establish good rules of thumb, and identifying an appropriate degrees of freedom value usually requires some experimentation.

To illustrate the effect of the ridge prior, I performed data augmentation on the small employee data set, this time using a ridge prior with two degrees of freedom. The top panel of Figure 9.2 shows the time-series plot for the simulated job performance means. Notice that the long-term trends are gone and that the means stay within a plausible range of values. The bottom panel of the figure shows the time-series plot for the covariance between employment status and job performance ratings. This parameter was previously inestimable, but the simulated parameters now vary around zero (the value specified by the prior). Both plots still display systematic trends, but the ridge prior dramatically reduces the problems that were evident in Figure 9.1.

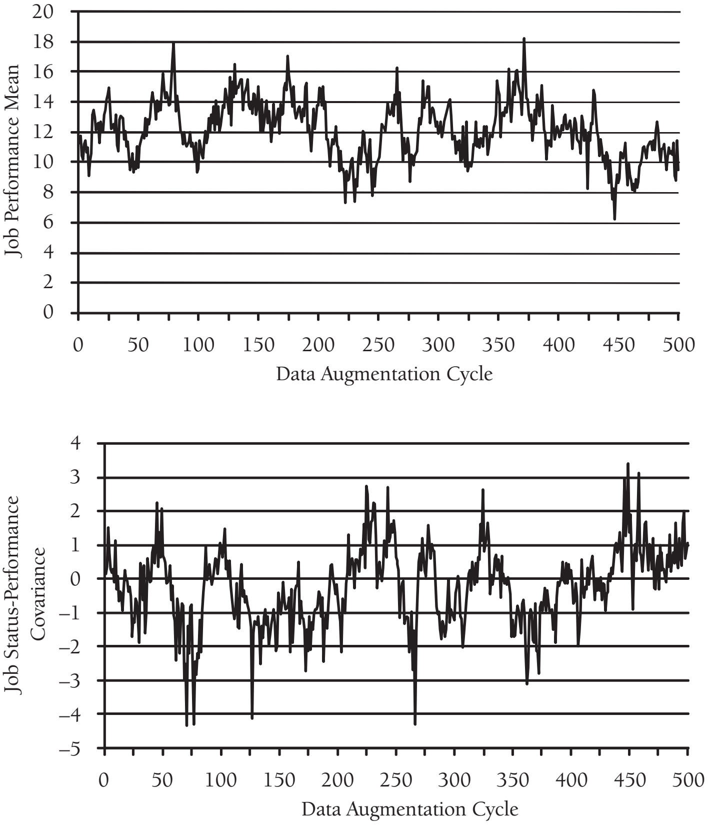

FIGURE 9.2. Time-series plot after specifying a ridge prior with ν = 2 degrees of freedom. The top panel shows a time-series plot of the job performance mean. The ridge prior eliminated the long-term dependencies, and the simulated parameters take on plausible values. The bottom panel shows the covariance between job status and job performance. This parameter was not estimable without the ridge prior, but now varies around a value of zero (the covariance specified by the prior).

9.3 DEALING WITH NON-NORMAL DATA

The data augmentation algorithm assumes multivariate normality, both at the I-step and at the P-step (e.g., the I-step draws residuals from a normal distribution, and the P-step distributions follow from assuming a normal distribution for the population data). However, Schafer and colleagues suggest that normality-based imputation can work for a variety of different distribution types (Bernaards, Belin, & Schafer, 2007; Graham & Schafer, 1999; Schafer, 1997; Schafer & Olsen, 1998). This is an important practical issue because normality is often the exception rather than the rule (Micceri, 1989). The next section describes special issues that arise with discrete data (e.g., nominal and ordinal variables), but for now it is useful to address normality violations in more general terms.

Empirical studies suggest that normality violations may not pose a serious threat to the accuracy of multiple imputation parameter estimates (Demirtas, Freels, & Yucel, 2008; Graham & Schafer, 1999; Leite & Beretvas, 2004; Rubin & Schenker, 1986; Schafer, 1997).

Perhaps not surprisingly, the magnitude of the bias depends on the sample size and the missing data rate. For example, Demirtas et al. (2008) found that the parameter estimates and standard errors from a bivariate data analysis were relatively accurate with a sample size of $N = 400$ but were quite distorted with a sample size of $N = 40$. Other simulation studies have reported accurate estimates and confidence intervals with sample sizes as low as $N = 100$ (Graham & Schafer, 1999; Schafer, 1997). The percentage of missing data also plays a role, such that bias increases as the missing data rate increases. Although it is difficult to establish rules of thumb about the percentage of missing data, Demirtas et al. (2008) reported accurate parameter estimates with missingness rates as high as 25%. Finally, the impact of normality violations varies across different parameter estimates. For example, variance estimates are sensitive to scores in the tails of a distribution, so they are likely to exhibit more bias than means and regression coefficients. Other parameters that depend on the tails of a distribution (e.g., extreme quantiles such as the 90th percentile) can also be quite sensitive to normality violations (Demirtas et al., 2008; Schafer, 1997).

Applying Normalizing Transformations at the Imputation Phase

One way to mitigate the impact of normality violations is to apply normalizing transformations at the imputation phase. Researchers sometimes object to transformations because the metric of the resulting scores is unfamiliar. However, variables can have different scales during the imputation and pooling phases, so it is possible to impute the variable on a transformed metric (e.g., a logarithmic scale) and analyze it on its original metric. Popular multiple-imputation software programs offer a variety of common data transformations, and these programs can automatically back-transform variables to their original metric when outputting the imputed data sets. Analyzing non-normal variables can still cause problems in the subsequent analysis phase, but applying data transformations at the imputation phase can improve the validity of data augmentation.

Despite their intuitive appeal, data transformations pose two potential problems. First, choosing an appropriate transformation is not necessarily straightforward. For example, logarithmic or square root transformations can work well for positively skewed variables, but the magnitude of the skewness and the kurtosis dictates the choice of transformation. Methodologists sometimes recommend experimenting with different transformations until you identify the one that best normalizes the data (Tabachnick & Fidell, 2007). This approach is difficult to implement, however, because there are currently no software programs that estimate skewness and kurtosis with missing data. Unfortunately, using deletion methods to assess the utility of different transformations can produce wildly inaccurate estimates of skewness and kurtosis, particularly if data are systematically missing from a distribution's tails. Data transformations are also problematic because they can alter the covariate structure of the data. Regression-based imputation relies heavily on the associations among the variables, so imputing variables on a transformed metric and back-transforming the scores to the original metric can potentially affect the accuracy of the imputations and the resulting parameter values. This has prompted some methodologists to raise strong concerns over the appropriate use of transformations in the context of multiple imputation (Demirtas et al., 2008, pp. 82–83). Further methodological research is needed to clarify this issue.

Applying Corrective Procedures at the Analysis Phase

Non-normal data can also cause problems at the analysis phase. The methodological literature suggests that normality violations have a limited impact on parameter estimates but can bias standard errors and distort the likelihood ratio test (Finney & DiStefano, 2006; West, Finch, & Curran, 1995). The corrective procedures described in Chapter 5 (e.g., robust standard errors and rescaled test statistics) have long been available for complete-data analyses, and some of these procedures are readily applicable to multiple imputation. For example, it is perfectly appropriate to apply Rubin's (1987) pooling formulas to robust (i.e., sandwich estimator) standard errors. Similarly, the sandwich estimator can generate the within-imputation covariance matrices for the D_1 test statistic from Chapter 8. Unfortunately, it is unclear how to implement corrective procedures for the likelihood ratio test. For example, the methodological literature offers no guidance on whether it is appropriate to use rescaled likelihood ratio tests to compute the D_3 statistic. This is a fruitful area for future methodological research.

9.4 TO ROUND OR NOT TO ROUND?

Discrete measurement scales are exceedingly common in the behavioral and the social sciences, and researchers often incorporate nominal and ordinal variables into the imputation phase. Methodologists have developed specialized imputation algorithms for mixtures of categorical and continuous variables (e.g., the general location model—Schafer, 1997; sequential regression imputation—Raghunathan, Lepkowski, Van Hoewyk, & Solenberger, 2001), some of which I describe later in the chapter. However, these more complex categorical data models do not necessarily produce accurate parameter estimates (Belin, Hu, Young, & Grusky, 1999), so data augmentation may be the best option.

One consequence of applying an imputation model for normal data to discrete variables is that the resulting imputations will have decimals. The traditional advice is to round the imputed values to the nearest integer or to the nearest plausible value (Schafer, 1997; Schafer & Olsen, 1998; Sinharay, Stern, & Russell, 2001). For example, Schafer (1997, p. 148) suggests that "the continuous imputes should be rounded off to the nearest category to preserve the distributional properties as fully as possible and to make them intelligible to the analyst." At an intuitive level, rounding is appealing because it eliminates implausible values and yields imputations that are aesthetically consistent with the observed data. However, recent research suggests that rounding may not be necessary and can actually lead to biased parameter estimates.

Much of the empirical work on rounding has focused on binary variables (Allison, 2005; Bernaards et al., 2007; Horton, Lipsitz, & Parzen, 2003; Yucel, He, & Zaslavsky, 2008). These studies clearly suggest that rounding is something to avoid. At an intuitive level, it is reasonable to expect the effects of rounding to diminish as the number of ordinal response options increases. To date, relatively few studies have systematically examined the impact of rounding multiple-category ordinal variables (e.g., 5-point Likert scales). Computer simulation studies provide some indirect evidence that rounding is not as problematic with 5-category ordinal variables (Van Ginkel, Van der Ark, & Sijtsma, 2007a, 2007b), but analyzing the fractional imputations still appears to be the best option, at least for now (Wu & Enders, 2009). In some situations the analysis model requires rounding (e.g., a binary outcome in a logistic regression, a set of dummy variables). The remainder of this section describes some strategies for dealing with this issue.

Rounding Binary Variables

The impact of rounding seems to be most pronounced with binary variables. Including an incomplete binary variable (e.g., a dummy variable with codes of zero and one) in the imputation phase will produce a range of imputed values, including fractional values between zero and one, values greater than one, and even negative values. One strategy for converting fractional imputations to binary values is to apply a 0.50 rounding threshold to the imputed values (i.e., round imputed values that exceed 0.50 to one, and round imputed values that are less than 0.50 to zero). However, recent research suggests that this so-called **naïve rounding** scheme introduces bias, whereas analyzing the fractional imputations does not (Allison, 2005; Bernaards et al., 2007; Horton et al., 2003; Yucel et al., 2008). Although these studies clearly suggest that rounding a binary variable is a bad idea, some analysis models require a binary outcome variable (e.g., a logistic regression that predicts membership in one of two categories). For these situations, methodologists have proposed rounding rules that appear to work somewhat better than a simple 0.50 threshold. I describe two such strategies next.

Bernaards et al. (2007) describe a so-called **adaptive rounding** procedure that relies on the normal approximation to a binomial distribution. For each imputed data set, adaptive rounding applies the following threshold:

$$c = \hat{\mu}_{\text{UR}} - \Phi^{-1}(\hat{\mu}_{\text{UR}})\sqrt{\hat{\mu}_{\text{UR}}(1 - \hat{\mu}_{\text{UR}})} \tag{9.2}$$

where c is the rounding threshold, $\hat{\mu}_{\text{UR}}$ is the mean of the imputed (i.e., unrounded) binary variable, and $\Phi^{-1}(\hat{\mu}_{\text{UR}})$ is the *z* value from a standard normal distribution, below which the $\hat{\mu}_{\text{UR}}$ proportion of the distribution falls (i.e., the inverse of the standard normal cumulative distribution). To illustrate the adaptive rounding procedure, suppose that the mean of a binary variable is $\hat{\mu}_{\text{UR}} = 0.67$ in a particular imputed data set. In a standard normal distribution, a *z* value of 0.44 separates the lowest 67% of the curve from the rest of the distribution. Consequently, substituting $\hat{\mu}_{\text{UR}} = 0.67$ and $\Phi^{-1}(\hat{\mu}_{\text{UR}}) = 0.44$ into Equation 9.2 yields a rounding threshold of 0.463. Consistent with naïve rounding, imputed values that exceed the threshold are rounded to one, and values that fall below the threshold get rounded to zero.

Yucel et al. (2008) describe an alternate rounding strategy that they refer to as **calibration**. The first step of the calibration procedure is to create a copy of the raw data and delete the observed values of the incomplete binary variable from this file (i.e., make the binary variable completely missing). The second step is to vertically concatenate the original data and the copied data into a single stacked file. The final step is to impute the missing values in the concatenated file. Imputing the stacked data file yields filled-in values for the subsample of cases that actually have complete data on the binary variable. The idea behind calibration is to use these values to identify a rounding threshold that reproduces the frequency of ones and zero in the raw data.

TABLE 9.2. Illustration of Calibration Rounding for a Binary Variable

Stacked data			Imputed data			Rounded data		
ID	*X*	*Y*	*ID*	*X*	*Y*	*ID*	*X*	*Y*
				Original data				
1	7	0	1	7	0	1	7	0
2	10	1	2	10	1	2	10	1
3	3	1	3	3	1	3	3	1
4	5	0	4	5	0	4	5	0
5	5	1	5	5	1	5	5	1
6	8	0	6	8	0	6	8	0
7	1	—	7	1	0.596	7	1	1
8	2	—	8	2	0.172	8	2	0
9	4	—	9	4	0.857	9	4	1
10	8	—	10	8	0.961	10	8	1
				Duplicate data				
1	7	—	6	8	0.270	N/A	N/A	N/A
2	10	—	4	5	0.311	N/A	N/A	N/A
3	3	—	5	5	0.315	N/A	N/A	N/A
4	5	—	1	7	0.500	N/A	N/A	N/A
5	5	—	3	3	0.733	N/A	N/A	N/A
6	8	—	2	10	0.737	N/A	N/A	N/A
7	1	—	7	1	0.451	N/A	N/A	N/A
8	2	—	8	2	0.421	N/A	N/A	N/A
9	4	—	9	4	0.535	N/A	N/A	N/A
10	8	—	10	8	0.953	N/A	N/A	N/A

To illustrate the calibration procedure, Table 9.2 shows a hypothetical sample of $N = 10$ cases, 60% of which have data on a binary variable, *Y*. Furthermore, among the subsample of cases that have data, 50% have a code of one. The left-most set of columns shows the original data and the duplicate data file where *Y* is completely missing. The middle set of columns shows the data that result from imputing the entire set of $N = 20$ data records. Notice that the subsample of complete cases (i.e., the calibration subsample) has imputed values that range between 0.270 and 0.737. The goal of calibration is to use this subset of imputations to identify a rounding threshold that reproduces the frequency of ones and zeros in the observed data (i.e., a 50/50 split). For clarity, Table 9.2 orders the calibration subsample (shown in a shaded box) by their imputed values. As you can see, applying a rounding threshold of 0.32 to the calibration subsample yields a 50/50 split of ones and zeros. I applied this threshold to the four incomplete cases from the original sample, and the right-most column of the table shows the resulting *Y* values.

To date, no research has compared adaptive rounding to calibration, but both approaches appear to be superior to naïve rounding (Bernaards et al., 2007; Yucel et al., 2008). Calibration is likely to exhibit some bias with missing at random (MAR) data (Yucel et al., 2008), but simulation studies suggest that adaptive rounding does not suffer from this problem

(Bernaards et al., 2007). Adaptive rounding also has the advantage of being easier to implement, so until more research accumulates, it seems prudent to recommend this approach over calibration and naïve rounding.

Rounding a Set of Dummy Variables

A second situation in which rounding may be necessary occurs with incomplete nominal variables that have more than two categories. The appropriate way to impute a nominal variable is to recast it as a set of $g - 1$ dummy variables prior to imputation. With complete data, cases that belong to the reference group (e.g., a control group or some other normative group) have a value of zero on the entire set of dummy variables, and the remaining cases have zeros on all but one of the code variables (Cohen, Cohen, West, & Aiken, 2003). However, applying naive rounding to a set of imputed dummy variables can produce illogical values where a case has a code of one on multiple dummy variables. Consequently, it is necessary to apply rounding rules that produce a logical set of dummy codes.

Allison (2002) proposed straightforward rules for rounding a set of dummy variables. The cases with missing data on the nominal variable have imputed values for each of the $g - 1$ dummy codes. The first step of Allison's procedure is to compute a new variable that subtracts the sum of the imputed values from a value of one. This new variable serves as a pseudo-imputation for membership in the reference category (i.e., the group coded all zeros). Next, if the pseudo-imputation variable has the highest numeric value, you round the $g - 1$ dummy codes to zero, thereby assigning the case to the reference group. Otherwise, if the highest imputed value corresponds to one of the $g - 1$ dummy variables, you assign a value of one to the appropriate code variable and set the remaining dummy codes to zero. To illustrate Allison's rounding rules, Table 9.3 shows a small set of hypothetical imputations for a set of two dummy codes, D_1 and D_2 (i.e., a nominal variable with three categories). The first two columns contain the imputed values for D_1 and D_2 and the middle column is the pseudo-imputation for membership in the reference category (i.e., $1 - D_1 - D_2$). As you can see, the highest value in the first three columns determines each case's group membership. It is important to note that Allison's rounding rules have not been evaluated in the literature. Nevertheless, his rules provide a convenient solution for an imputation model that includes a number of multiple-category nominal variables.

TABLE 9.3. Illustration of Dummy Code Rounding Rules

Imputed codes			Rounded codes	
D_1	D_2	$1 - D_1 - D_2$	D_1	D_2
0.65	0.23	0.12	1	0
–0.12	0.55	0.57	0	0
0.77	–0.02	0.25	1	0
0.37	0.82	–0.19	0	1
0.05	1.08	–0.13	0	1
0.42	–0.02	0.60	0	0

Out-of-Range Imputations

In addition to producing fractional values, data augmentation will often produce imputations that fall outside of the plausible score range (e.g., a 5-point Likert variable that has an imputed value of 5.23). There are essentially three options for dealing with out-of-range values: (1) analyze the imputed values as they are, (2) round to the nearest plausible score value, or (3) generate new imputations for cases that have out-of-range values (e.g., by adding a new random residual to each predicted score). Multiple-imputation software packages make the latter two options easy to implement, but analyzing the out-of-range values may be a fine option, particularly if they are relatively few in number. At an intuitive level, out-of-range imputations can inflate variance estimates, but this bias is probably trivial if the number of implausible values is relatively small.

A large proportion of out-of-range imputations can be symptomatic of a normality violation, so transforming the data at the imputation phase may reduce or eliminate out-of-range values. However, transformations are unlikely to eliminate implausible imputations that occur when an ordinal variable has an asymmetric distribution (e.g., responses are isolated to small number of categories). Rounding the imputed values to the nearest plausible value is one solution, but an alternate strategy is to recast the ordinal variable as a set of dummy codes and apply Allison's (2001) rounding rules following imputation.

9.5 PRESERVING INTERACTION EFFECTS

Researchers in the behavioral and the social sciences are often interested in estimating interaction (i.e., moderation) effects where the magnitude of the association between two variables depends on a third variable. In some situations, the interaction effect appears as an explicit term in the analysis model. For example, if it was of interest to determine whether the association between psychological well-being and job performance is different for males and females, including a product term in a multiple regression model could address this question (i.e., moderated multiple regression; Aiken & West, 1991). Many other analyses model implicit interaction effects. For example, multiple-group structural equation models do not contain explicit interaction terms, yet they allow for group differences in the mean structure, the covariance structure, or both. A multilevel model with random intercepts and slopes is another analysis that involves implicit interaction effects.

When using multiple imputation to treat missing data, it is important to specify an imputation model that preserves any interaction effects that are of interest in the subsequent analysis model because failing to do so will attenuate the magnitude of these effects, even if the data are missing completely at random (MCAR). For example, if gender moderates the association between psychological well-being and job performance, failing to build this complex association into the imputation model is likely to produce an analysis that masks the gender difference. Similarly, an imputation model that fails to preserve group differences in the mean or the covariance structure could lead to the conclusion that the parameters of a multiple group structural equation model are invariant (i.e., the same) across groups when they are truly different in the population. This section outlines different imputation strategies for dealing with interactive effects. The appropriate strategy depends largely on whether the interaction involves a categorical or a continuous moderator variable.

Interactions That Involve Quantitative Variables

If the analysis model includes an interaction effect between two quantitative variables, then the imputation phase should include a variable that is the product of the two interacting variables. This is effectively the only way to preserve the interaction. For example, suppose that it is of interest to determine whether the number of years on the job moderates the relationship between psychological well-being and job performance. A standard approach for addressing this question is to estimate a multiple regression model that includes main effects and a product term as predictor variables (e.g., years on the job, psychological well-being, and the product of years on the job and well-being). The imputation phase also employs a multiple regression model, so it too should include the same set of variables. The product variable is particularly important because it preserves the complex associations among the variables. It is important to point out that including a product variable in the imputation phase does not create an interaction effect where none exists. Rather, it simply preserves the natural structure of the data. Finally, note that the product term strategy also applies to nonlinear associations. For example, if the analysis model includes a quadratic effect, then the imputation phase should include main effects and a squared term.

When an analysis model includes an interaction effect between two or more quantitative variables, it is important to center predictor variables at their means (i.e., subtract the mean from each score) prior to analyzing the data (Aiken & West, 1991). However, centering becomes difficult when one of the variables in the product term has missing data. One option is to center the variables prior to imputation, compute the necessary product term, and fill in the missing variables (including the product term) on their centered metrics. This approach requires estimates of the variable means, so maximum likelihood estimates (e.g., from an initial EM analysis) are a logical choice. A second strategy is to fill in the missing variables (including the product term) on their original metrics and subsequently perform the centering procedure on each of the complete data sets. Because the product of two uncentered variables has a larger mean and a larger variance than the product of two centered variables (Bohrnstedt & Goldberger, 1969), this method requires a complete rescaling of the imputed product variable. Neither of these approaches has been evaluated in the literature, but centering the variables prior to imputation is far easier and tends to yield estimates that are similar to those of a maximum likelihood analysis. Until further research suggests otherwise, this is probably the best strategy.

Interactions That Involve a Categorical Variable

When it is of interest to examine an interaction effect that involves a categorical variable, imputing the data separately for each subgroup is often more accurate than including product terms in the imputation model (Enders & Gottschall, in press). To understand why, suppose that it is of interest to determine whether a binary categorical variable *D* moderates the association between *X* and *Y* (e.g., gender moderates the association between psychological well-being and job performance). Furthermore, suppose that some individuals have missing *Y* values. Using a product term to preserve the interaction effect yields the following imputation model:

$$y_i^* = \hat{\beta}_0 + \hat{\beta}_1(X_i) + \hat{\beta}_2(D_i) + \hat{\beta}_2(X_i)(D_i) + z_i \tag{9.3}$$

where y_i^* is the imputed value for case *i*, X_i and D_i are the observed scores for that case, and z_i is a normally distributed residual term. Including the dummy code variable in the imputation model preserves group mean differences on *Y*, and the product term preserves group differences in the covariance between *X* and *Y*. It may not be immediately obvious, but using a single normal distribution to generate the residual terms effectively assumes that both groups have the same *Y* variance (i.e., data augmentation generates imputations that are **homoscedastic**). This subtle assumption may have a relatively minor impact on many analyses, but in a number of situations the substantive goal is to determine whether the covariance structure is the same across qualitatively different subpopulations (e.g., measurement invariance analyses, multiple-group structural equation models). If the subgroups have different population variances, then the product term approach will generate imputations that mask these group differences (Enders & Gottschall, in press).

A simple solution to the previous problem is to impute the data separately for each subgroup (i.e., **separate-group imputation**). Because this approach uses a unique imputation equation and a unique residual distribution for each subpopulation, every element in the mean vector and the covariance matrix will freely vary across groups. The downsides of separate-group imputation are that it (1) is limited to situations that involve categorical moderator variables, (2) requires adequate group sizes, and (3) necessitates additional effort to assess convergence (e.g., by examining the time-series and autocorrelation function plots for each subgroup). Despite these potential limitations, the approach is very easy to implement in multiple imputation software programs and has performed well in computer simulation studies, even with a smple size as low as $n = 50$ per group (Enders & Gottschall, in press).

Models with Implicit Interaction Effects

Many common statistical analyses involve implicit interaction effects. Multiple-group structural equation models are one such example. To illustrate, consider a measurement invariance analysis in which it is of interest to determine whether the factor model parameters (e.g., the loadings, measurement intercepts) are the same across qualitatively different subpopulations (e.g., males and females, Caucasians, and Hispanics). A typical measurement invariance analysis begins with separate factor models for each subgroup. Subsequent analysis steps constrain sets of parameter estimates (e.g., the factor loadings) to be equal across groups. If the constrained model fits the data as well as the unconstrained model, then there is evidence that the subgroups have the same population mean vector or covariance matrix. In contrast, a constrained model that shows worse fit suggests that the subgroups have a different mean vector or covariance matrix.

Multiple-group structural equation models do not contain explicit interaction effects, but they allow for group differences in the mean structure, covariance structure, or both.

Consequently, it is necessary to specify an imputation model that preserves these group differences. Incorporating product terms into the imputation phase is problematic because it generates imputations from a model that assumes equal variances across groups. As a result, the subsequent analyses are likely to suggest that certain parameters are invariant (i.e., the same) across groups, when they are actually different in the population. In contrast, separate-group imputation naturally preserves group differences in the mean vector and the covariance matrix and will lead to more accurate assessments of subgroup differences. Computer simulations suggest that the separate-group imputation approach produces accurate parameter estimates in a variety of multiple-group structural equation models (e.g., moderated mediation, multiple-group confirmatory factor analysis, multiple-group growth curves), with sample sizes as low as $n = 50$ per group (Enders & Gottschall, in press).

A multilevel model with random intercepts and slopes is another analysis that contains implicit interaction effects. To illustrate, consider an educational study in which students (i.e., level-1 units) are nested within schools (i.e., level-2 units). Furthermore, suppose that it is of interest to examine the influence of student socioeconomic status on academic achievement. A random intercept model is one in which the mean achievement level differs across schools, and a random slope model allows the association between socioeconomic status and achievement to vary across schools. These group differences in the mean and the covariance structure show up as variance estimates rather than as regression coefficients, but they are interaction effects, nevertheless.

The data augmentation algorithm from Chapter 7 is not designed for multilevel data structures where the associations among variables potentially vary across clusters. In principle, separate-group imputation is appropriate for imputing missing values at the lowest level of the data hierarchy (e.g., by imputing individual-level variables separately for each cluster), but this approach requires a relatively large number of cases within each cluster. Many (if not most) common applications of multilevel modeling (e.g., dyadic data, longitudinal data, children nested within classrooms) do not have adequate group sizes to support this method. A better strategy is to use a specialized imputation algorithm for multilevel data (Schafer, 2001; Schafer & Yucel, 2002; Yucel, 2008). I describe one such algorithm later in the chapter.

A Cautionary Note on Latent Categorical Variables

A number of popular statistical models treat group membership as a latent categorical variable. Finite mixture models (McLachlan & Peel, 2000; Muthén, 2001, 2004) and latent class models (McCutcheon, 1987) are two common examples. Consistent with a multiple group structural equation model, it is often of interest to determine whether the latent classes have different mean and covariance structures. For example, a growth mixture model is characterized by a number of latent subgroups, each of which can have a different growth trajectory (i.e., different mean structures) and varying degrees of individual heterogeneity in the growth trajectories (i.e., different covariance structures). These models are important to consider because they are becoming increasingly common in the social sciences.

Because group membership is inferred from the data during the analysis, there is no way to use product terms or separate-group imputation to preserve the implicit interaction effects that are present in the data. Consequently, multiple imputation can produce biased estimates of the model parameters, even when the data are MCAR (Enders & Gottschall, in press). Fortunately, maximum likelihood missing data routines are readily available for many popular latent class models (e.g., growth mixture models, factor mixture models), so there is no need to rely on multiple imputation. Methodologists are also beginning to develop imputation algorithms for latent categorical variables (Vermunt, Van Ginkel, Van der Ark, & Sijtsma, 2008). As a result, these procedures are likely to become increasingly common in the near future.

9.6 IMPUTING MULTIPLE-ITEM QUESTIONNAIRES

Researchers in the behavioral and the social sciences routinely use multiple-item questionnaires to measure complex constructs. For example, psychologists typically use several questionnaire items to measure depression, each of which taps into a different depressive symptom (e.g., sadness, lack of energy, sleep difficulties, feelings of hopelessness). With multiple-item questionnaires, respondents often omit one or more of the items within a given scale. Multiple imputation is advantageous because it provides a mechanism for dealing with item-level missingness (maximum likelihood can be less flexible in this regard). However, imputation can be challenging or even impossible when the data contain a large number of questionnaire items. This is an important practical issue because it is not uncommon for researchers to administer a dozen or more questionnaires in a single study, each of which may contain 20 or more items. The number of variables can quickly multiply in a longitudinal study that has several questionnaires administered on multiple occasions.

Ideally, the imputation phase should include all of the individual questionnaire items because this maximizes the information that goes into creating the imputations. However, item-level imputation may not be feasible when the number of questionnaire items is very large. As an upper limit, the number of variables in the imputation model cannot exceed the number of cases because the input data contain linear dependencies that cause mathematical difficulties for regression-based imputation. Because missing data exacerbate these mathematical difficulties, the allowable number of variables tends to be much lower than the number of cases. One possible solution for imputing large data sets is to use a ridge prior described earlier in the chapter. Conceptually, the ridge prior adds a number of imaginary data records (i.e., degrees of freedom) to the estimation process, but it does so at the cost of attenuating the associations among the variables. A complex imputation model can require a relatively large number of additional degrees of freedom, in which case the ridge prior might be a poor solution. An alternative approach is to perform separate data augmentation runs for different subsets of variables. However, this strategy effectively assumes that variables from different subsets are uncorrelated, and it is viable only if variables from different subsets are not part of the same analysis model. This section outlines three alternative approaches for imputing large questionnaire data sets: scale-level, duplicate scale, and a three-step imputation approach.

Scale-Level Imputation

When collecting data with multiple-item questionnaires, researchers are often interested in analyzing scale scores based on a sum or an average of the item responses. When the analysis

TABLE 9.4. Input Data for Item-Level, Scale-Level, and Duplicate-Scale Imputation

Item-level imputation							Scale-level imputation			Duplicate-scale imputation				
X_1	X_2	X_3	Y_1	Y_2	Y_3	Z	S_X	S_Y	Z	S_X	S_Y	Z	A_X	A_Y
5	4	5	3	—	4	20	4.67	—	20	4.67	—	20	4.67	3.50
2	—	1	3	2	3	17	—	2.67	17	—	2.67	17	1.50	2.67
4	3	5	5	5	4	24	4.00	4.67	24	4.00	4.67	24	4.00	4.67
—	3	2	—	—	4	13	—	—	13	—	—	13	2.50	4.00
1	1	3	2	2	1	9	1.67	1.67	9	1.67	1.67	9	1.67	1.67

Note. S_X and S_Y are scale scores that average the individual questionnaire items ($X_1 - X_3$ and $Y_1 - Y_3$). The scale scores are missing if one or more of the items are missing. A_X and A_Y are averages of the available items within each scale, and Z is an auxiliary variable.

model involves scale scores, ignoring the item-level data and imputing the scale scores themselves can dramatically reduce the number of imputation model variables (Graham, 2009). Under this **scale-level imputation** approach, the cases that have complete data on a particular subset of items (e.g., a set of depression items) also have complete data on the scale score, whereas the individuals who fail to answer one or more of the questionnaire items have missing data. To illustrate, Table 9.4 shows a small data set with a single auxiliary variable and six questionnaire items (X_1 to X_3 and Y_1 to Y_3) that combine to form two subscales, S_X and S_Y. The scale-level imputation procedure for these data would include just three variables: S_X, S_Y, and the auxiliary variable, Z.

Scale-level imputation can dramatically reduce the number of variables in the imputation model and can eliminate the mathematical difficulties associated with imputing a large number of individual items. However, it does so at the cost of reducing statistical power. In my experience, scale-level imputation can increase standard errors by up to 10% relative to an ideal analysis that uses scale scores from an item-level imputation procedure. This decrease in statistical power becomes increasingly evident as the number of items within a scale increases. The failure of scale-level imputation stems from the fact that questionnaire items within a scale tend to have stronger correlations than items from different scales. Consequently, the imputation phase effectively discards the strongest predictors of the missing scale scores (i.e., the items within the scale) in favor of weaker correlates (i.e., items from different scales).

One way to mitigate the power loss from scale-level imputation is to incorporate the item-level information back into the imputation model. A simple way to do this is to compute a second set of scale scores by averaging the available items within each questionnaire. For example, if a respondent answered 8 out of 10 items on a particular questionnaire, the scale score for that individual would be the average of the eight observed items. Incorporating these additional scales into the imputation phase as auxiliary variables can recapture much of the item-level information that scale-level imputation ignores. For lack of a better term, I henceforth refer to this approach as **duplicate-scale imputation**. The right-most section of Table 9.4 illustrates the input data for this method. Notice that the complete cases have identical scores on both sets of scales (e.g., S_X and A_X are the same), whereas the incomplete cases only have data on the duplicate scales. The duplicate-scale imputation approach requires twice as many variables as scale-level imputation, but it can dramatically reduce the complexity of the imputation model. For example, suppose that the two questionnaires in Table 9.4 had 20 items each. Duplicate-scale imputation would still only require five variables: S_X, S_Y, A_X, A_Y, and Z.

In my experience, duplicate-scale imputation tends to yield parameter estimates and standard errors that are nearly identical to those of an ideal analysis that uses scale scores from an item-level imputation procedure. However, getting duplicate-scale imputation to work properly requires an additional nuance. Because the cases with complete data have identical scores on both sets of scales, the data contain linear dependencies that cause estimation problems for data augmentation. Using a ridge prior distribution to add imaginary data records to the imputation process can solve this problem. Fortunately, adding a small number of additional degrees of freedom usually eliminates the linear dependencies, so any bias that results from use of a ridge prior is negligible. For example, later in the chapter I present an analysis example in which adding a single imaginary data record (i.e., a ridge prior with a single degree of freedom) eliminates the linear dependencies in the imputation model and produces parameter estimates and standard errors that are virtually identical to those of item-level imputation.

A Three-Step Approach for Item-Level Imputation

The duplicate-scale approach can work well for analyses that involve scale scores, but many analysis models require item-level data (e.g., internal consistency reliability analyses, confirmatory factor analyses). In situations where the number of items is prohibitively large, Little, McConnell, Howard, and Stump (2008) outline a three-step approach for item-level imputation. The idea behind their procedure is to separately impute different subsets of questionnaire items. This strategy is usually undesirable because it assumes that variables from different item subsets are uncorrelated. However, Little et al. solve this problem by using scale scores to preserve the between-subset associations.

The Little et al. procedure requires a complete set of scale scores. The authors use scale-level imputation to generate these scores, but averaging the available items within a scale is another option. These initial scale scores are simply temporary auxiliary variables, so the method that you use to generate them probably makes little difference. The second step involves an iterative imputation process that repeatedly fills in the item scores from one subset while using the scale scores from the remaining subsets as auxiliary variables. As an example, consider a study that collects data on 10 multiple-item questionnaires (i.e., Q_1 to Q_{10}), each of which has 20 items. The first imputation phase might consist of the 20 items from Q_1 and the scale scores for Q_2 through Q_{10}. Similarly, the second imputation model could include the Q_2 items and the scale scores for the nine remaining questionnaires (i.e., Q_1, Q_3 through Q_{10}). Depending on the sample size, it may be possible to perform fewer data augmentation runs with larger item subsets (e.g., impute the items from Q_1 through Q_5 while using the scale scores for Q_6 through Q_{10} as auxiliary variables). After completing the imputation process for each item subset, the temporary placeholder scales from the first step are no longer necessary. Consequently, the final step is to discard the initial scales and compute a new set of composite scores from the filled-in item responses.

To date, no studies have evaluated the duplicate-scale approach or the three-step imputation approach. Because a single scale score preserves the between-subset associations, these procedures probably work best when the items within one subset have relatively uniform correlations with the items from another subset. Fortunately, this is a fairly realistic condition for many scales in the behavioral and the social sciences, so these procedures probably work well in a variety of settings. The imputation of large item-level data sets is an important practical topic that warrants future methodological research.

9.7 ALTERNATE IMPUTATION ALGORITHMS

Although the data augmentation algorithm in Chapter 7 is probably the most popular imputation strategy, methodologists have developed a number of alternative imputation routines. Some of these algorithms are applicable to specialized situations that are relatively uncommon in the behavioral and the social sciences (e.g., data comprised entirely of categorical variables; Schafer, 1997; randomized trials with monotone missing data patterns; Lavori, Dawson, & Shera, 1995), while others are suitable replacements for data augmentation (King, Honaker, Joseph, & Scheve, 2001). A thorough review of different imputation options is beyond the scope of this chapter, but it is useful to briefly describe some of these alternative models. This section begins with a description of an EM-based imputation algorithm that is statistically equivalent to data augmentation. Next, the section outlines two algorithms for imputing data sets that contain a mixture of categorical and continuous variables. The final section describes an imputation algorithm for multilevel data structures. Note that the algorithms in this section simply replace data augmentation in the imputation phase and do not require changes to the analysis and pooling phases.

EM-Based Algorithms for Multivariate Normal Data

Generating unique sets of imputations from multivariate normal data requires several alternate estimates of the mean vector and the covariance matrix. The P-step of data augmentation generates these estimates by simulating random draws from a posterior distribution. King et al. (2001) describe two approaches that use the EM algorithm from Chapter 4 to generate alternate estimates of the mean vector and the covariance matrix. These EM-based approaches also simulate random draws from a posterior distribution, but they do so in a very different fashion. The EM with an importance sampling algorithm is particularly interesting because it is statistically equivalent to data augmentation, yet it does not require the same complicated definition of convergence.

EM with sampling (EMS) begins by using the EM algorithm to estimate the mean vector and the covariance matrix. These maximum likelihood estimates describe the central tendency of the posterior distributions from which the algorithm will draw alternate parameter estimates. Next, the algorithm computes the parameter covariance matrix for the EM estimates and uses this matrix to define the spread of the posterior distributions. Having characterized the shape of the posterior distribution, the EMS algorithm uses Monte Carlo simulation techniques to draw *m* new estimates of the mean vector and the covariance matrix from their respective posteriors. This process does not require a long iterative chain. Rather, the algorithm simply generates the desired number of alternate estimates. Finally, EMS uses each set of parameter values to construct regression equations that impute the missing values. The final imputation stage is identical to stochastic regression imputation (or alternatively, the I-step of data augmentation).

Using the parameter covariance matrix to estimate the spread of the posterior distribution is only appropriate in very large samples and can produce biased parameter estimates in small to moderate samples (King et al., 2001). To correct this problem, King et al. proposed a modified algorithm that they call **EM with importance sampling** (EMIS). EMIS also uses maximum likelihood estimates and the parameter covariance matrix to approximate the posterior distributions, but it uses the likelihood function to fine-tune the shape of the distributions. Rather than retaining every set of simulated the parameter values, the algorithm selectively discards the estimates that are inconsistent with the data (i.e., estimates that have a low likelihood of producing the sample data).

More specifically, the EMIS algorithm works as follows. First, the algorithm uses Monte Carlo simulation techniques to draw a set of alternate parameter values from a multivariate normal posterior distribution, the shape of which is defined by the EM estimates and the corresponding parameter covariance matrix. With small to moderate samples, the true posterior distribution may quite skewed, in which case the simulated parameters are not always accurate. Then, to remedy this problem, EMIS uses the likelihood function to weed out implausible parameter values. (Assuming a noninformative prior distribution, the likelihood function has the same shape as the correct posterior distribution.) Specifically, the algorithm generates an **importance ratio** by substituting the simulated parameters into the likelihood function and converting the resulting likelihood value into a probability. Simulated parameter values that have a high likelihood of producing the sample data also have a high importance ratio (i.e., probability), whereas parameters that are unlikely to have produced the sample data have a low importance ratio. To decide whether to retain a particular set of parameters, the algorithm generates a uniform random number between zero and one and compares this number to the importance ratio. EMIS retains the estimates if the uniform random number is less than the importance ratio. Otherwise, the algorithm discards the estimates and generates a new set. This so-called **acceptance-rejection algorithm** repeatedly screens simulated parameter values until it retains *m* sets of plausible estimates. The resulting estimates more closely approximate random draws from the true posterior distribution, the shape of which may not resemble a normal distribution. Finally, EMIS uses the retained parameter values to construct regression equations that impute the missing values. The final imputation stage is identical to stochastic regression imputation.

The EMIS algorithm is statistically equivalent to data augmentation (i.e., it will yield the same analysis results, on average) but offers some potential advantages. One advantage is that EMIS can be easier to implement. Because the simulated parameter values do not depend on the imputed values from a preceding iteration, the *m* sets of imputations are automatically independent samples from the distribution of missing values. This simplifies the imputation process considerably because it eliminates the need for graphical convergence diagnostics. By extension, there is no need to worry about the number of between-imputation iterations or other convergence-related issues that make data augmentation challenging to implement. Speed is a second advantage. Data augmentation often requires thousands of iterations to generate a relatively small number of data sets. With large data files, this can take a considerable amount of time. Because EMIS does not continually iterate between draws, it can generate the same number of data sets in a much shorter period (e.g., problems that take data augmentation several minutes to run take EMIS just a few seconds). Although data augmentation is the predominant method for generating imputations with multivariate normal data, the EMIS algorithm is certainly worth considering. At the time of this writing, Amelia is the only software program that implements EMIS.

Algorithms for Categorical and Continuous Variables

One shortcoming of the data augmentation is that it assumes a common distribution for every variable in the data set (i.e., the multivariate normal distribution). This is an unrealistic assumption because data sets often contain a mixture of categorical and continuous variables. Schafer (1997) and colleagues suggest that normality-based imputation can often work well with categorical (e.g., nominal and ordinal) variables, but it is worth considering imputation algorithms that do not assume a common distribution. This section describes two such approaches: the general location model and sequential regression imputation. These methods are similar in the sense that they apply different imputation models to categorical and continuous variables, but their procedural details are quite different. Of the two, sequential regression is particularly promising because it is conceptually straightforward and has performed well in empirical studies.

Schafer (1997, Chapter 9) describes an imputation approach for categorical and continuous variables based on the so-called **general location model** (Little & Schluchter, 1985; Olkin & Tate, 1961). The general location model uses a fully crossed contingency table to represent the categorical variables, and it assumes that the continuous variables follow a normal distribution within each cell of the table. The model for the continuous variables resembles a factorial multivariate analysis of variance (MANOVA) in the sense that the cells share a common covariance matrix but can have different means. To illustrate the general location model, consider a data set with two continuous variables and two categorical variables, both of which have three levels. (The categories can be ordered, but the model treats them as nominal.) The saturated general location model for this example has 29 parameters. The contingency table is comprised of nine cells, so the categorical variables contribute eight parameters to the model (if the sample size is fixed, the frequency for the ninth cell is determined by the other eight). The continuous variable means vary across cells, adding another 18 parameters, and the covariance matrix of the continuous variables has three unique elements.

Schafer (1997) outlined a data augmentation algorithm for the general location model that consists of an I-step and a P-step. The procedure follows the same basic logic as data augmentation for multivariate normal data, but it uses different distribution families (e.g., the categorical variables follow a multinomial distribution, and the continuous variables are normally distributed within cells of the contingency table). The I-step imputes the incomplete categorical variables by assigning each missing observation to a cell in the contingency table, and it then uses a stochastic regression procedure to impute the missing continuous variables. The continuous variables inform the categorical imputations and vice versa. Conditional on the imputations from the preceding I-step, the P-step draws new cell probabilities for the contingency table and subsequently generates a new covariance matrix and a new set of cell means. Schafer describes the data augmentation algorithm in considerable detail, and both Schafer and Belin et al. (1999) illustrate applications of the general location model.

The general location model is seemingly well-suited for many realistic missing data problems, but it may not be the best option for imputing mixtures of categorical and continuous variables. One problem is that the model becomes exceedingly complex as the number of variables increases. For example, Belin et al. (1999) applied the general location model to a data set with 16 binary variables and 18 continuous variables. Although this data set is not unusually large, the saturated model has more than one million parameters! The staggering number of parameters is attributable to the fact that the model includes main effects as well as every possible higher-order interaction among the categorical variables. In practice, it is usually necessary to perform a series of preliminary analyses to simplify the model prior to data augmentation, but doing so adds a layer of complexity to the imputation process. Complexity issues aside, Belin et al. (1999) raise concerns about the accuracy of the general location model, particularly for categorical imputations. Until further research is done, it may be best to view the general location model with some caution.

Sequential regression imputation is a second approach for imputing data sets that contain mixtures of categorical and continuous variables. (The literature also refers to this method as **chained equations** and **fully conditional specification**.) Unlike the general location model, sequential regression imputation fills in the data on a variable-by-variable basis, each time matching the imputation model to a variable's distributional form. For example, the algorithm can use a linear regression to impute continuous variables, a logistic regression to impute binary variables, a Poisson regression to impute count variables, and so on. The remainder of this section gives a brief overview of the algorithm, and a number of sources provide more detailed descriptions of this approach (Raghunathan et al., 2001; van Buuren, 2007; van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006).

Like data augmentation, the sequential regression approach uses regression equations to generate draws from the conditional distribution of the missing values, given the observed data. However, the mechanics of imputation are quite different. For one, the algorithm imputes variables in a sequence, one at a time. The imputation order is determined by the rates of missingness, where the variable with the fewest missing values gets imputed first, the variable with the next lowest missing data rate gets imputed second, and so on. Each step in the imputation sequence can apply a regression model that is appropriate for the scale of the incomplete variable (e.g., a logistic regression imputes incomplete binary variables, a linear regression imputes normally distributed variables, and so on). Unlike data augmentation, each regression model uses the filled-in values from one sequence to generate imputations for subsequent sequences. For example, suppose that Y_3 gets imputed in the first regression sequence, Y_1 gets imputed in the second sequence, and Y_4 in the final sequence. After the initial sequence, the algorithm treats Y_3 as a complete variable and uses the observed and the imputed values as predictors of the missing Y_1 scores. Similarly, the next sequence uses the filled-in values of Y_3 and Y_1 to impute Y_4.

After filling in the entire data set, the algorithm uses a Bayesian procedure that is akin to the P-step of data augmentation to sample a new set of regression parameters, and the process begins anew. The second and subsequent rounds of imputation also fill in the data on a variable-by-variable basis, but they do so using all variables in the imputation model, including the filled-in variables from the preceding iteration. For example, the filled-in values of Y_1 and Y_4 from the first imputation cycle serve as predictors of Y_3 in the first sequence of the second imputation cycle. The sequential regression algorithm iterates for a specified number of cycles, and the imputed values from the final iteration serve as data for a subsequent analysis. Repeating the imputation chain m times generates unique sets of imputed values.

The sequential regression approach has a number of advantages over data augmentation. Most importantly, it is unnecessary to assume that the variables share a common distribution because the algorithm tailors the imputation model to each incomplete variable. In addition, formulating a separate imputation model for each variable makes it easy to specify constraints that preserve special characteristics of the data. For example, to avoid logical inconsistencies between two variables, the range of imputed values for one variable can depend on the responses to another variable. Similarly, it is straightforward to accommodate survey skip patterns by restricting imputation to the subsample of cases that endorse a screener question. Despite its advantages, using separate regression models for imputation also introduces difficulties. For one, implementing the procedure is more cumbersome because it requires additional programming that is not necessary with data augmentation. Second, the use of diverse regression models can produce a situation where the algorithm fails to converge to a stable distribution (Raghunathan et al., 2001). In addition, assessing convergence is typically more difficult with sequential regression than it is with data augmentation (see Van Buuuren, 2007, for an illustration). Despite these potentially serious difficulties, simulation studies suggest that sequential regression performs well and can produce unbiased parameter estimates and standard errors (Raguhunathan et al., 2001; van Buuren et al., 2006). Although additional methodological research is needed, the sequential regression method may become a viable alternative to data augmentation when the data contain mixtures of categorical and continuous variables. A number of specialized software packages implement the sequential regression approach (e.g., MICE, ICE, IVEWARE), and the SPSS Missing Values add-on (available in version 17 and higher) also offers this imputation option.

An Algorithm for Multilevel Data

Multilevel data structures are characterized by observations that are nested within higher-level units or clusters (e.g., children nested within schools, employees nested within workgroups, repeated measures nested within individuals). Multilevel analysis techniques are well-suited for these data structures because they appropriately account for the nesting and allow researchers to investigate associations at different levels of the data hierarchy (Raudenbush & Bryk, 2002). The data augmentation algorithm from Chapter 7 is inappropriate for multilevel data sets because it fails to preserve between-cluster differences in the mean structure and the covariance structure. For example, in an education study, the association between socioeconomic status and student achievement might differ across schools, but data augmentation imputes missing values from a model where this association is constant for all schools in the sample. Not surprisingly, this can seriously distort the subsequent parameter estimates.

Methodologists have developed specialized imputation algorithms for multilevel data (Schafer, 2001; Schafer & Yucel, 2002; Yucel, 2008). These routines may require software packages that you are not familiar with (e.g., the PAN library for the S-Plus program), but taking the time to learn one of these programs can provide an advantage over using maximum likelihood to estimate a multilevel model with missing data. At this time, multilevel software packages generally allow for missing data on outcome variables, but they eliminate cases with missing predictor variables. Although there is often little reason to prefer multiple imputation over maximum likelihood (or vice versa), the ability to retain cases with missing predictor variables gives multiple imputation a clear advantage in this situation. Analyzing multiply imputed data sets is also very straightforward because a number of multilevel software packages have facilities for automating the analysis and pooling phases.

Before describing the multilevel imputation algorithm, it is useful to review the multilevel model. As an illustration, consider a study of school achievement where children (i.e., level-1 units) are nested within a number of different schools (i.e., level-2 units). Furthermore, suppose that it is of interest to predict student achievement based on socioeconomic status and school size. The multilevel regression model for this analysis is

$$Y_{ij} = \gamma_{00} + \gamma_{10}(SES_{ij}) + \gamma_{01}(Size_j) + \gamma_{11}(SES_{ij})(Size_j) + u_{oj} + u_{1j}(SES_{ij}) + r_{ij} \tag{9.4}$$

where Y_{ij} is the achievement score for child i in school j, the γ terms are regression coefficients, u_{0j} is a level-2 residual that allows the achievement means to differ across schools, u_{1j} is a level-2 residual that allows the association between socioeconomic status and achievement to vary across schools, and r_{ij} is a level-1 residual that captures individual differences within a particular school. The level-2 residuals (i.e., the u terms) in the equation are essentially latent variables, the values of which differ across clusters (e.g., schools). Finally, it is worth noting that the multilevel model estimates a level-1 and a level-2 covariance matrix as opposed to the residuals themselves.

Multilevel imputation uses an iterative algorithm called the **Gibbs sampler** (Casella & George, 1992; Gelfand & Smith, 1990), which closely resembles data augmentation. The Gibbs sampler consists of a series of steps where the values at one step depend on the quantities from the previous step. In the context of multiple imputation, each iteration of the Gibbs sampler consists of three steps: (1) draw level-2 residuals from a distribution of plausible values, (2) draw new parameter values (i.e., regression coefficients, the level-2 covariance matrix, and the level-1 covariance matrix) from their respective posterior distributions, and (3) impute the missing values. I give a brief sketch of the imputation algorithm in the remainder of this section; interested readers can find additional details in Schafer (2001) and Schafer and Yucel (2002).

To begin, the Gibbs sampler draws a set of level-2 residuals from a normal distribution. The exact shape of this distribution depends on the filled-in data and the parameter values (i.e., the regression coefficients and the covariance matrices) from the previous iteration. The level-2 residuals are an important starting point because they define the shape of the posterior distributions in the second step and because they facilitate the computation of the multilevel model parameters. Next, the Gibbs sampler uses Monte Carlo simulation to draw new parameter values from their respective posterior distributions. Similar to the P-step of data augmentation, the algorithm draws the level-1 and level-2 covariance matrices from an inverse Wishart distribution, and it uses a multivariate normal distribution to generate a new set of regression coefficients. The exact shape of these distributions depends on the level-2 residuals from the first step and on the imputed values from the preceding iteration. The final step of the Gibbs sampler generates predicted scores for each case by substituting the observed variables and the level-2 residuals into a multilevel regression model similar to that in Equation 9.5. Consistent with the I-step of data augmentation, the algorithm restores variability to the imputed data by augmenting each predicted score with a normally distributed residual term.

Implementing a multilevel imputation model involves additional nuances that are not relevant to standard data augmentation. For example, deciding what to include in the imputation model becomes more complex. Following standard procedure, the imputation phase should include analysis model variables and auxiliary variables. However, you also need to decide which level-2 residual terms (i.e., random effects) to include in the imputation regression model. These residuals determine whether the association between two variables varies across clusters, so omitting an important residual term can bias the subsequent parameter estimates. Although it may seem like a good idea to include every possible residual term, doing so can lead to estimation problems and convergence failures. In addition to specifying which residual terms get included in the model, it is necessary to specify a covariance structure for the residuals. For example, a saturated covariance matrix allows the residuals for different variables to freely correlate, but it is also possible to specify a matrix that restricts the between-variable associations to zero. The first option will better preserve the associations among the variables, but the complexity of the resulting imputation model can cause estimation problems. Schafer (2001) and Schafer and Yucel (2002) describe model specification issues in more detail and give an analysis example that applies a multilevel imputation model.

9.8 MULTIPLE-IMPUTATION SOFTWARE OPTIONS

A number of software packages generate multiply imputed data sets, some of which are commercially available, while others are freely available on the Internet. Software programs tend to change at a rapid pace, so a detailed description of these packages would quickly become out of date. Rather, this section provides a very general overview of multiple imputation computing options, and I discuss a small handful of software options in more detail in Chapter 11. A variety of resources are available for readers interested in the details of specific software programs (e.g., Allison, 2000; Honaker, King, & Blackwell, 2009; Horton & Lipsitz, 2001; Raghunathan, Solenberger, & Van Hoewyk, 2002; Royston, 2005; Schafer & Olsen, 1998; Yuan, 2000), and there are also useful websites that provide information about individual software packages (e.g., *www.multiple-imputation.com*).

Multiple-imputation software packages generally fall into one of three categories: programs that (1) generate multiply imputed data sets, (2) analyze multiply imputed data sets created by other programs, and (3) generate and analyze multiply imputed data sets. The programs that generate multiple imputations tend to offer the same set of features, some of which are described earlier in this chapter and in previous chapters (e.g., data transformations, rounding options, ridge prior distributions). Although there is considerable overlap in features, software programs differ in the type and the number of algorithms that they implement. For example, the SAS MI procedure implements the data augmentation algorithm, whereas SPSS offers the sequential regression approach (also known as chained equations and fully conditional specification) in its Missing Values add-on. SAS and SPSS are arguably the most popular statistical software packages in the social and the behavioral sciences, but a number of specialized imputation programs are also available (e.g., NORM, Amelia, MICE), as are open-source programs that offer a variety of user-written modules (e.g., the S-Plus and R statistical packages). Finally, software programs differ in their overall ease of use; some programs have point-and-click interfaces (e.g., the NORM program), but most are syntax driven (e.g., the SAS MI procedure and various R modules).

Regardless of which program you use to generate the multiple imputations, you have a number of options for analyzing the data and combining the resulting estimates. For example, many popular software packages offer built-in routines for analyzing multiply imputed data sets (e.g., SAS, Mplus, HLM, to name just a few). Some of these programs require considerable programming to combine the *m* sets of estimates and standard errors, whereas others are so easy to use that the pooling process is virtually transparent to the user. Software programs also differ in the amount of summary information that they provide, so this is an additional consideration when choosing an analysis platform. For example, some programs output detailed diagnostic information (e.g., fraction of missing information, relative increase in variance, between- and within-imputation variance), whereas others simply report the pooled estimates and standard errors. In my experience, it is often convenient to use one program to generate the imputations and use a different program to analyze the data, but this choice is largely one of personal preference.

9.9 DATA ANALYSIS EXAMPLE 1

The first analysis example uses multiple imputation to estimate a regression model with an interaction term.* The data for this analysis consist of scores from 480 employees on eight work-related variables: gender, age, job tenure, IQ, psychological well-being, job satisfaction, job performance, and turnover 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 accounts for 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 goal of the analysis is to determine whether gender moderates the association between psychological well-being and job performance. The multiple regression equation is as follows:

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

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

Using maximum likelihood to estimate a model with an interaction term is straightforward and follows the same procedure as any multiple regression analysis (e.g., see the analysis example in Chapter 4). However, dealing with interactive effects is more complex in a multiple imputation analysis because the imputation phase must account for group differences in the mean and the covariance structure. With a nominal moderator variable such as gender, the best way to preserve an interactive effect is to impute the data separately for each group.

The Imputation Phase

To implement separate-group imputation, I sorted the data by gender and performed data augmentation separately for males and females. Note that the imputation model included every variable except gender, which was constant in each group. As I explained previously, separate-group imputation naturally preserves interaction effects because it allows the mean and the covariance structure to freely vary across subpopulations. The graphical diagnostics for males and females suggested fast convergence, so I specified 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). Consistent with the analysis example from Chapter 8, I opted to use $m = 50$ imputations for the analysis phase.

The Analysis and Pooling Phases

The standard advice in the regression literature is to center continuous predictor variables at the grand mean (Aiken & West, 1991; Cohen et al., 2003). To do so, I merged the male and female files and computed the mean well-being score within each of the 50 imputed data sets. Next, I centered the psychological well-being scores by subtracting the appropriate mean from each score, and I then computed a product variable (i.e., interaction term) by multiplying gender and the centered well-being scores. Finally, I estimated a multiple regression model with job performance scores as the outcome variable and gender, psychological well-being, and the product term as predictors. The analysis phase produced 50 sets of regression coefficients and standard errors that I subsequently pooled into a single set of results.

Researchers often begin a regression analysis with an omnibus F test, and the D_1 statistic from Chapter 8 is ideally suited for this purpose. This analysis produced a test statistic of $D_1 = 40.34$. Referencing this value to an F distribution with 3 numerator and 3802.40 denominator degrees of freedom returned a probability value of $p < .001$. Consistent with the omnibus F test from an ordinary least squares regression analysis, a significant test statistic indicates that at least one of the regression slopes is statistically different from zero.

Table 9.5 shows the pooled estimates and standard errors, along with the corresponding maximum likelihood estimates from Chapter 4. Although the imputation and analysis models are not congenial (the imputation model is more complex than the analysis model), the two analysis procedure produced nearly identical parameter estimates and standard errors. Turning to the individual parameter estimates note that, males and females do not differ with respect to their mean job performance ratings, $\hat{\beta}_2 = -0.175$, $t = -1.66$, $p = .10$, but the interaction term indicates that the association between well-being and performance is differ-

TABLE 9.5. Regression Model Estimates from Data Analysis Example 1

Parameter	Estimate	*SE*	*t*
	Multiple imputation		
β_0 (intercept)	6.092	0.076	79.828
β_1 (well-being)	0.332	0.065	5.107
β_2 (gender)	–0.173	0.105	–1.644
β_3 (interaction)	0.355	0.100	3.566
$\hat{\sigma}_e^2$ (Residual)	1.193	0.083	14.403
R^2	0.240		
	Maximum likelihood estimation		
β_0 (intercept)	6.091	0.076	79.755
β_1 (well-being)	0.337	0.071	4.723
β_2 (gender)	–0.167	0.105	–1.587
β_3 (interaction)	0.362	0.106	3.426
$\hat{\sigma}_e^2$ (residual)	1.234	0.084	14.650
R^2	0.214		

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

ent for males and females, $\hat{\beta}_3 = 0.355$, $t = 3.57$, $p < .001$. Because the gender variable is coded such that female = 1 and male = 0, the sign of the interaction coefficient indicates that the relationship is stronger for females. Notice that the interpretation of the regression coefficients is identical to what it would have been had the data been complete. In addition, the computation of simple slopes is identical to that of a complete-data analysis. For example, the regression equation for the subsample of males (the group coded 0) is $\hat{Y}_M = \hat{\beta}_0 + \hat{\beta}_1(WB)$, and the corresponding equation for females (the group coded 1) is $\hat{Y}_F = (\hat{\beta}_0 + \hat{\beta}_2) + (\hat{\beta}_1 + \hat{\beta}_3)(WB)$.

9.10 DATA ANALYSIS EXAMPLE 2

The second data analysis example illustrates the difference between scale-level imputation and duplicate-scale imputation.* The analyses use artificial data from a questionnaire on eating disorder risk. Briefly, the data contain the responses from 400 college-aged women on 10 questions from the Eating Attitudes Test (EAT; Garner, Olmsted, Bohr, & Garfinkel, 1982), a widely used measure of eating disorder risk. The 10 questions measure two constructs, Drive for Thinness (e.g., "I avoid 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 data set also contains an anxiety scale score, a variable that

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

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}). Expecting a relationship between body weight and missingness, I created MAR data on five variables (EAT_1, EAT_{10}, EAT_{12}, EAT_{18}, and EAT_{24}) by deleting the EAT scores for a subset of cases in both tails of the BMI distribution. These same EAT questions were also missing for individuals with elevated anxiety scores. Finally, I introduced a small amount of MNAR data by deleting a number of the high body mass index scores (e.g., to mimic a situation 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

For the imputation phase, I generated three sets of $m = 20$ imputations by (1) imputing the individual questionnaire items (i.e., item-level imputation), (2) imputing the scale scores directly (i.e., scale-level imputation), and (3) imputing the scale scores using averages of the available items as auxiliary variables (i.e., duplicate-scale imputation). The number of variables in this data set is not nearly large enough to pose a problem for item-level imputation (the ideal procedure). Nevertheless, imputing the data using three approaches is useful for illustrating the differences that can result from using scale-level and duplicate-scale imputation. For each strategy, I used a single sequential data augmentation chain with 100 burn-in and 100 between-imputation iterations. The data augmentation algorithm converged very quickly and without problems, so there is no need to present the graphical diagnostics from the exploratory data augmentation chain.

The item-level imputation model included all 13 variables in the data set (i.e., the 10 EAT questionnaire items, anxiety scores, Western standards of beauty scores, and body mass index values). For scale imputation, I began by computing scale scores by averaging the two sets of questionnaire items. 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}). Consequently, the Drive for Thinness scale score was an average of seven Likert items, and the Food Preoccupation scale was an average of three items. I restricted the scale score computations to the cases with complete data, so respondents who were missing one or more of the item responses within a particular scale were also missing the scale score. This produced 291 cases with Drive for Thinness scores, 352 cases with Food Preoccupation scores, and 276 individuals with complete data on both scale scores. The subsequent scale-level imputation model had five variables: the two EAT scale scores, anxiety scores, Western standards of beauty scores, and body mass index values.

The duplicate-scale imputation procedure was identical to that of scale imputation, but it also included two additional variables that I computed by averaging the available items within each scale. Again, the purpose of the duplicate scales is to recapture the important item-level information that scale imputation discards. An important nuance of the duplicate-score approach is that the input data contain linear dependencies. Although the graphical diagnostics looked ideal, the software package issued a warning message that the initial covariance matrix (i.e., the Bayesian estimate of $\boldsymbol{\Sigma}$ that generates the regressions for the first I-step) was singular. I eliminated this problem by specifying a ridge prior distribution with a single degree of freedom. This effectively added one imaginary data record to the data augmentation procedure.

The Analysis and Pooling Phases

To keep the analysis model simple, I estimated the mean vector and the covariance matrix for five variables: the Drive for Thinness and Food Preoccupation scale scores, the anxiety scores, the Western standards of beauty scores, and the body mass index values. The analysis phase produced a mean vector and a covariance matrix for each of the 20 imputed data sets. I subsequently used the pooling formulas from Chapter 8 to combine the estimates and the standard errors; Table 9.6 shows the results for three imputation approaches. Even when the ultimate goal is to analyze scale scores, imputing the individual questionnaire items and computing scale scores from the filled-in item responses should provide better results than imputing the scale scores directly. Consequently, the item-level imputation is the "gold standard" against which to compare the other methods. Focusing on the covariance matrix elements for the EAT scale scores, notice that scale-level imputation produced larger standard errors than item-level imputation. Also, notice that the standard error inflation tends to be somewhat larger for the seven-item Drive for Thinness scale. This suggests that the power loss may increase as the number of scale items increases. In contrast, duplicate-scale imputation produced estimates and standard errors that are quite similar to those of item-level imputation. Scale-level imputation performed poorly because questionnaire items within a scale tend to have stronger correlations than items from different scales. Consequently, the imputation phase effectively discards the strongest predictors of the missing scale scores (i.e., the items within the scale) in favor of weaker correlates (i.e., items from different scales). Although it is not possible to draw firm conclusions from a single artificial data set, the standard error differences in Table 9.6 are consistent with what you might expect to see in real data sets.

9.11 SUMMARY

This chapter addressed a number of practical issues that arise during the imputation phase. The chapter began with a discussion of convergence problems. Convergence issues often occur because there is insufficient data to estimate certain parameters. This lack of data can result from including too many variables in the imputation phase or from a peculiar missing data pattern. In some situations, reducing the number of variables or eliminating the problematic variables can solve convergence problems. An alternate strategy is to specify a ridge prior distribution for the covariance matrix. Conceptually, the ridge prior adds a small number of imaginary data records (i.e., degrees of freedom) from a hypothetical population where the variables are uncorrelated. These additional data points can stabilize estimation and eliminate convergence problems, but they do so at the cost of introducing a slight bias to the simulated parameter values (and thus the imputations). The biasing effect of the ridge prior

TABLE 9.6. Mean Vector and Covariance Matrix Estimates from Data Analysis Example 2

Variable	1	2	3	4	5
			Item-level imputation		
1: DFT	0.612 (0.044)				
2: FP	0.349 (0.039)	0.759 (0.054)			
3: ANX	1.227 (0.135)	1.254 (0.149)	9.078 (0.655)		
4: WSB	0.549 (0.084)	0.462 (0.089)	0.997 (0.307)	3.667 (0.270)	
5: BMI	0.846 (0.115)	0.664 (0.125)	1.164 (0.422)	1.109 (0.275)	7.343 (0.521)
Means	3.959 (0.039)	3.966 (0.044)	11.979 (0.152)	8.964 (0.099)	22.405 (0.136)
			Scale-level imputation		
1: DFT	0.599 (0.047)				
2: FP	0.349 (0.042)	0.739 (0.521)			
3: ANX	1.213 (0.151)	1.236 (0.151)	9.035 (0.663)		
4: WSB	0.577 (0.083)	0.443 (0.093)	1.007 (0.306)	8.964 (0.097)	
5: BMI	0.815 (0.126)	0.617 (0.129)	1.137 (0.418)	1.087 (0.274)	7.347 (0.521)
Means	3.957 (0.047)	3.971 (0.044)	11.979 (0.153)	8.964 (0.097)	22.401 (0.136)
			Duplicate-scale imputation		
1: DFT	0.616 (0.044)				
2: FP	0.353 (0.039)	0.768 (0.054)			
3: ANX	1.223 (0.135)	1.227 (0.149)	9.009 (0.649)		
4: WSB	0.547 (0.082)	0.452 (0.089)	1.050 (0.308)	3.637 (0.266)	
5: BMI	0.838 (0.115)	0.668 (0.125)	1.134 (0.419)	1.111 (0.278)	7.369 (0.524)
Means	3.959 (0.039)	3.967 (0.044)	11.965 (0.152)	8.968 (0.098)	22.402 (0.137)

Note. DFT = drive for thinness; FP = food preoccupation; ANX = anxiety; WSB = Western standards of beauty; BMI = body mass index. Values in parentheses are standard errors.

depends on the number of degrees of freedom that you assign to the prior, so it is generally a good idea to select a value that is as small as possible.

Like maximum likelihood estimation, the data augmentation algorithm in Chapter 7 assumes multivariate normality, both at the I-step and at the P-step. However, methodologists suggest that normality-based imputation can work for a variety of different distribution types. Empirical studies suggest that normality violations may not pose a serious threat to the accuracy of multiple imputation parameter estimates, particularly if the sample size is not too small and the missing data rate is not too large. One way to mitigate the impact of normality violations is to apply normalizing transformations to the data prior to performing data augmentation. Variables can have different scales during the imputation and pooling phases, so you can impute a variable on a transformed metric (e.g., a logarithmic scale) and subsequently analyze it on its original metric.

Nominal and ordinal variables are a special case of non-normal data that arises frequently in the behavioral and the social sciences. One consequence of applying an imputation model for normal data to discrete variables is that the resulting imputations will have decimals. The traditional advice is to round imputed values to the nearest integer or to the nearest plausible value in order to produce imputations that are aesthetically consistent with the observed data. However, recent research suggests that rounding may not be necessary and can actually lead to biased parameter estimates. Aesthetics aside, there appear to be no negative consequences associated with analyzing fractional imputations, so analyzing the data without rounding seems to be the safest strategy, at least for now. However, in some cases the analysis model requires rounding (e.g., a binary outcome in a logistic regression, a set of dummy variables), and the chapter described some rounding strategies for these situations.

Researchers in the behavioral and the social sciences are often interested in estimating interaction (i.e., moderation) effects where the magnitude of the association between two variables depends on a third variable. When using multiple imputation to treat missing data, it is important to specify an imputation model that preserves any interaction effects that are of interest in the subsequent analysis model. Failing to do so will attenuate the magnitude of these effects, even if the data are MCAR or MAR. The best strategy for preserving interaction effects depends largely on whether the interaction involves a continuous or a categorical moderator variable. If the analysis model includes an interaction effect between two quantitative variables, the only way to preserve the interaction effect is to include a product variable in the imputation phase. The downside of this approach is that the imputation regression model generates filled-in values that are homoscedastic. This subtlety may have a relatively minor impact on many analyses, but in a number of situations the substantive goal is to determine whether the covariance structure is the same across qualitatively different subpopulations (e.g., measurement invariance analyses, multiple-group structural equation models). If the subgroups have different population variances, then the product term approach will generate imputations that mask these group differences. Consequently, when an interactive effect involves a categorical moderator variable, imputing the data separately for each subgroup is often more accurate than including product terms in the imputation model.

Researchers in the behavioral and social sciences routinely use multiple-item questionnaires to measure complex constructs. Multiple imputation is advantageous for dealing with item-level missingness, but imputation can be challenging when a data set contains a large number of variables. Ideally, the imputation phase should include all of the individual questionnaire items in order to maximize the information that goes into creating the imputations. However, this may not be feasible when the number of questionnaire items is very large. When the analysis model involves scale scores, ignoring the item-level data and imputing the scale scores themselves can dramatically reduce the number of imputation model variables. This approach tends to lack power, but using the average of the available items as auxiliary variables (i.e., duplicate-scale imputation) can yield estimates and standard errors that are quite similar to those of an item-level imputation procedure. For situations that require item-level data, I outlined a three-step approach for item-level imputation. The basic idea behind this procedure is to separately impute different subsets of questionnaire items, each time using scale scores to preserve the between-subset associations among the items.

The chapter concluded with a description of some alternate imputation algorithms. Although the data augmentation algorithm in Chapter 7 is probably the most popular imputation strategy, methodologists have developed a number of alternative imputation algorithms. The chapter described an EM-based imputation algorithm that is statistically equivalent to data augmentation. This EMIS algorithm is appealing because it automatically yields independent imputations and does so more quickly than data augmentation. The chapter also described two algorithms (the general location model and sequential regression) appropriate for data sets that contain a mixture of categorical and continuous variables. Of these two, the sequential regression approach appears particularly promising. Unlike data augmentation, which assumes a common distribution for every variable in the data set, sequential regression imputation fills in the data on a variable-by-variable basis, each time matching the imputation model to a variable's distributional form. Preliminary simulation studies suggest that this procedure works well. Finally, I described an imputation algorithm for multilevel data structures. This algorithm is important because standard data augmentation fails to preserve any differences in the mean and the covariance structure that might exist across clusters.

The majority of this book is devoted to two so-called modern missing data techniques: maximum likelihood and multiple imputation. These methods use quite different approaches, but both assume MAR data. Although MAR-based methods are a substantial improvement over traditional methods that require the MCAR mechanism, they will produce bias when the data are missing not at random (MNAR). Chapter 10 outlines models that are designed specifically for MNAR data. As you will see, these MNAR methods are far from perfect and require assumptions that can be just as tenuous as MAR. In fact, when the model assumptions are violated, MNAR approaches can yield estimates that are worse than what you would have obtained from an MAR analysis. Nevertheless, MNAR models are useful for sensitivity analysis and are an important area of ongoing methodological research.

9.12 RECOMMENDED READINGS

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

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549–576.

Graham, J. W., & Schafer, J. L. (1999). On the performance of multiple imputation for multivariate data with small sample size. In R. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 1–29). Thousand Oaks, CA: Sage.

Horton, N. J., & Lipsitz, S. R. (2001). Multiple imputation in practice: Comparison of software packages for regression models with missing data. *The American Statistician*, *55*, 244–254.

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.