5

Improving the Accuracy of Maximum Likelihood Analyses

5.1 CHAPTER OVERVIEW

Now that the basic mechanics of maximum likelihood estimation have been established, this chapter outlines a collection of procedures that can improve the accuracy of a maximum likelihood analysis. The first half of the chapter focuses on the use of auxiliary variables—potential correlates of missingness or correlates of the incomplete analysis model variables. These variables are not of substantive interest; the sole purpose of including them in an analysis is to increase power or reduce bias. The chapter describes a number of issues related to the use of auxiliary variables, including their potential benefits, the process of identifying auxiliary variables, and approaches to incorporating these variables into a maximum likelihood analysis. Much of this discussion is very general and also applies to the use of auxiliary variables in multiple imputation analyses. However, the procedures for including auxiliary variables in a maximum likelihood analysis are unique and require a special model setup.

The second half of the chapter addresses problems related to non-normal data. Chapters 3 and 4 described the important role of normal distribution in a maximum likelihood analysis, but they did not discuss the ramifications of violating the normality assumption. The methodological literature shows that non-normal data tend to have a minimal impact on the parameter estimates themselves but can bias standard errors and distort the likelihood ratio test. Fortunately, methodologists have developed a number of corrective procedures for non-normal data (e.g., rescaled test statistics, the bootstrap, and robust standard errors), several of which are available for missing data analyses. The limited research to date suggests that these methods work quite well.

5.2 THE RATIONALE FOR AN INCLUSIVE ANALYSIS STRATEGY

Methodologists regard maximum likelihood estimation as a state-of-the-art missing data technique (Schafer & Graham, 2002) because it yields unbiased parameter estimates under

the missing at random (MAR) mechanism. The MAR mechanism holds when the probability of missing data on a variable *Y* relates to some other measured variable (or variables) but not to the values of *Y* itself. This definition seems to imply that MAR is automatically satisfied when a correlate of missingness is a variable in the data set, but it is the variables in the analysis model that dictate whether the MAR assumption is met.

To illustrate the subtleties of the MAR mechanism, consider a study that examines a number of health-related behaviors (e.g., smoking, drinking, and sexual activity) in a teenage population. Because of its sensitive nature, researchers decide to administer the sexual behavior questionnaire to participants who are above the age of 15. At first glance, this example appears to satisfy the MAR assumption because a measured variable (i.e., age) determines whether data are missing. However, this is not necessarily true because the MAR assumption is only satisfied if the researchers incorporate age into their analysis model. For example, suppose that the researchers use maximum likelihood missing data handling to estimate a simple regression model where self-esteem predicts risky sexual behavior. Because the age variable is not in the model, this analysis is actually consistent with the missing not at random (MNAR) mechanism and is likely to produce biased parameter estimates, particularly if age and sexual activity are correlated.

Understanding why the regression model is biased requires a brief review of Rubin's (1976) missing data theory. In Rubin's theory, a binary variable *R* denotes whether a variable is observed or missing (e.g., $r = 1$ if the sexual activity score is observed and $r = 0$ if it is missing). The MAR mechanism allows for an association between *R* and other measured variables such as age, but it stipulates that *R* is unrelated to sexual activity. Omitting age from the regression model is likely to introduce bias because it induces a spurious association between *R* and the missing sexual activity scores. The magnitude of the bias may not be problematic and depends on the correlation between age and sexual activity, but the analysis is nevertheless consistent with an MNAR mechanism. Had the researchers incorporated age into the regression model, the spurious association between *R* and the sexual activity scores would disappear because the age variable fully explains the relationship (i.e., after controlling for age, there is no residual association between *R* and sexual activity).

The previous scenario illustrates that the variables in the analysis model dictate whether the MAR assumption is met. For this reason, methodologists recommend a so-called **inclusive analysis strategy** that incorporates a number of auxiliary variables into the missing data handling procedure (Collins, Schafer, & Kam, 2001; Graham, 2003; Rubin, 1996; Schafer & Graham, 2002). An **auxiliary variable** is one that is ancillary to the substantive research questions but is a potential correlate of missingness or a correlate of the missing variable. Incorporating these variables into the missing data handling procedure can mitigate (or eliminate) bias and can improve power. For example, had the researchers in the health study included age in their model, they would have converted the analysis from MNAR to MAR and would have completely eliminated bias. Omitting an important correlate of missingness from an analysis can produce an MNAR mechanism, but MNAR data can also result from a direct relationship between missingness and the scores on the incomplete variable (e.g., teenagers who are engaging in risky sexual behavior skip the questionnaire items that address this topic). In this situation, auxiliary variables can still reduce bias, but they cannot completely eliminate it. Finally, even if the analyses are consistent with an MCAR or MAR mechanism,

auxiliary variables can improve power by recapturing some of the lost information in the missing variable. Consequently, it is nearly always beneficial to include auxiliary variables in the missing data handling procedure, and there appears to be no downside to an inclusive analysis strategy (Collins et al., 2001).

5.3 AN ILLUSTRATIVE COMPUTER SIMULATION STUDY

To illustrate the impact of auxiliary variables, I conducted a series of Monte Carlo computer simulations. The simulations mimicked a simple research scenario in which the goal is to estimate the mean vector and the covariance matrix for two variables, *X* and *Y*. The artificial data sets also included a third variable, *A*, that served as an auxiliary variable. The first simulation program generated 1,000 samples of $N = 100$ and subsequently produced a 30% missing data rate by deleting *Y* values for cases in the lower tail of an auxiliary variable's distribution. This simulation mimics the previous health study example because a measured variable completely determines missingness. The impact of omitting a correlate of missingness from an analysis depends on its association with the incomplete variables, so the population correlation between *Y* and the auxiliary variable varied between 0.10 and 0.80. For simplicity, the population correlation between the two analysis variables was always $\rho = .30$, as was the correlation between *X* and the auxiliary variable. After generating each data set, the simulation program used maximum likelihood to estimate the mean vector and the covariance matrix for *X* and *Y*, but it did so by omitting the auxiliary variable (i.e., the "cause" of missingness) from the analysis.

The top panel of Table 5.1 shows the average parameter estimates across the 1,000 replications. The purpose of this simulation is to illustrate the impact of excluding a correlate of missingness from an analysis (i.e., performing an MNAR analysis when it would have been possible to perform an MAR analysis). The left-most column gives the population correlation between the auxiliary variable (i.e., the omitted cause of missingness) and the incomplete analysis model variable, *Y*. As you can see, omitting a correlate of missingness was not that detrimental when the correlation was weak (e.g., $r \leq .30$), but the estimates became increasingly biased as the correlation between the auxiliary variable and the incomplete analysis variable increased in magnitude. Because the auxiliary variable completely determined missingness, incorporating this variable into the maximum likelihood analysis would eliminate bias, regardless of the magnitude of the correlation between the auxiliary variable and *Y*.

An MNAR mechanism can also result from a direct relationship between missingness and the scores on the incomplete variable (e.g., teenagers who are engaging in risky sexual behavior skip the questionnaire items that address this topic). In this situation, auxiliary variables can reduce, but not eliminate, bias. To illustrate the impact of auxiliary variables in this context, I performed a second simulation study that produced a 30% missing data rate by deleting *Y* values for cases in the lower tail of the *Y* distribution. As before, I varied the magnitude of the correlation between *Y* and the auxiliary variable, but this time I included the auxiliary variable in the analysis.

The bottom panel of Table 5.1 shows the average parameter estimates from the second simulation. As you can see, the overall magnitude of the bias was greater than in the first

TABLE 5.1. Simulation Results Showing the Impact of an Auxiliary Variable on Parameter Estimate Bias

	Average parameter estimates		
$\rho_{A,Y}$	$\rho_{X,Y}$	μ_Y	σ^2_Y
MNAR due to omitted auxiliary variable			
0.10	0.300	0.002	0.990
0.20	0.296	0.026	0.985
0.30	0.291	0.054	0.977
0.40	0.286	0.079	0.967
0.50	0.281	0.106	0.951
0.60	0.276	0.132	0.934
0.70	0.271	0.160	0.913
0.80	0.266	0.187	0.889
MNAR due to Y			
0.10	0.257	0.243	0.835
0.20	0.258	0.240	0.835
0.30	0.259	0.232	0.837
0.40	0.262	0.219	0.840
0.50	0.265	0.200	0.845
0.60	0.270	0.175	0.854
0.70	0.275	0.144	0.869
0.80	0.282	0.106	0.893
True values	0.300	0	1.000

Note. $\rho_{A,Y}$ is the population correlation between the auxiliary variable and the missing analysis variable, Y.

simulation, and the auxiliary variable did not eliminate bias. Although the average parameter estimates did get closer to the true population values as the correlation increased, the bias was still noticeable, even when the correlation between the auxiliary variable and the missing analysis variable was 0.80. Finally, the table shows that the largest reductions in bias occurred when the auxiliary variable's correlation with Y exceeded .50.

Auxiliary variables can also improve the power of maximum likelihood significance tests, regardless of their impact on bias (Collins et al., 2001). For example, I performed a third simulation study in which 30% of the Y values were MCAR. With MCAR data, an auxiliary variable has no impact on bias, but it can improve power. Consistent with the previous simulation, I varied the magnitude of the correlation between Y and the auxiliary variable and included the auxiliary variable in the analysis model. Because I generated the data from a population with a nonzero correlation, the proportion of the 1,000 replications that produced a statistically significant parameter estimate serves as an empirical estimate of statistical power.

Figure 5.1 shows the power estimates for the correlation between X and Y expressed relative to the power values from a maximum likelihood analysis that omits the auxiliary variable.

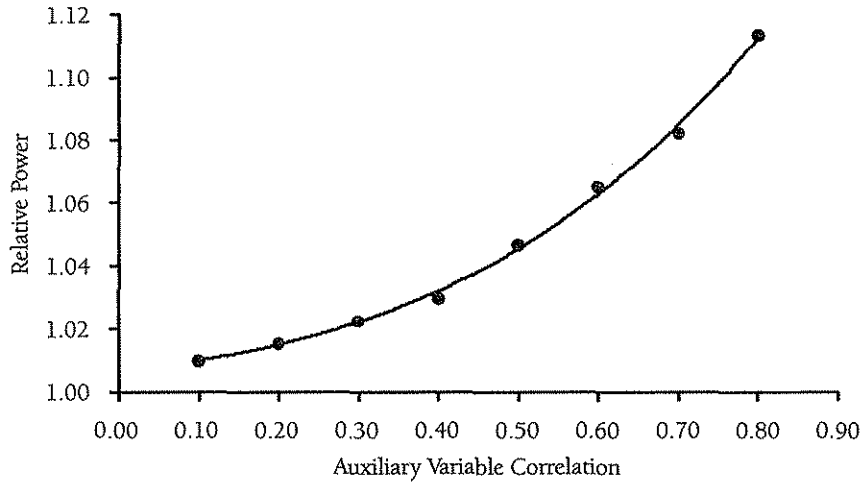

FIGURE 5.1. The figure shows how the correlation between an auxiliary variable and a missing analysis variable affects power. The largest gains in power occur after the auxiliary variable's correlation exceeds 0.40.

For example, a correlation of 0.30 between Y and the auxiliary variable produced a relative power value of 1.023, which means that the auxiliary variable increased power by approximately 2.3%. As you can see, there is a nonlinear relationship between the auxiliary variable correlation and power, such that the largest gains occur when the correlation exceeds 0.40. Incorporating an auxiliary variable always produces some benefit, but stronger correlations are clearly desirable.

Taken as a whole, the simulation results suggest that an auxiliary variable is most useful when it has a relatively strong correlation (e.g., $r > .40$) with the missing analysis variable. Conversely, omitting a correlate of missingness from the analysis has a minimal impact when the correlation is low (e.g., $r < .40$). Although there is no harm in using auxiliary variables with weak (or even zero) correlations (Collins et al., 2001), the benefits of an inclusive analysis strategy become more evident as the correlations become greater. This suggests that selecting auxiliary variables based on their correlation with the missing analysis variables is a useful strategy. The next section outlines additional strategies for selecting auxiliary variables.

5.4 IDENTIFYING A SET OF AUXILIARY VARIABLES

Given the benefits of incorporating auxiliary variables into a maximum likelihood analysis, it is useful to consider how to go about choosing these variables. As a rule, a useful auxiliary variable is a potential cause or correlate of missingness or a correlate of the incomplete variables in the analysis model (Collins et al., 2001; Schafer, 1997). For example, age is an important auxiliary variable in the health study because it directly influences missingness (i.e., researchers only administer the sexual behavior questionnaire to participants who are above the age of 15). Other useful auxiliary variables are correlates of the analysis model variables, regardless of whether they are also related to missingness. For example, a survey question that asks teenagers to report whether they have a steady boyfriend or girlfriend is a good auxiliary variable because it is likely correlated with the missing sexual activity scores.

The health example is straightforward because the researchers know why the data are missing. In most situations, identifying correlates of missingness involves some educated guesswork. To illustrate the process, consider an educational study that examines the development of self-report behavioral problems throughout the course of middle school and high school. Student mobility is a common cause of attrition in school-based research studies (Enders, Dietz, Montague, & Dixon, 2006; Graham, Hofer, Donaldson, MacKinnon, & Schafer, 1997), so the researchers could administer a survey question that asks parents to report how likely they are to move during the course of the study. Socioeconomic status is another factor that can influence attrition in school- and community-based samples, so it would be useful to collect a measure of socioeconomic status or a suitable proxy (e.g., participation in a lunch assistance program). Finally, in states where students are required to pass a statewide assessment in order to matriculate, students who fail to achieve the minimum passing score are at risk for dropping out of school. Consequently, the scores from a standardized achievement test might relate to subsequent attrition.

Identifying auxiliary variables that are correlates of the missing behavior reports is relatively straightforward, and a literature review can facilitate this process, if necessary. For example, objective measures of behavior such as disciplinary referrals, absenteeism, and incidents with the juvenile justice system are variables that should be readily available from the school district's database. In addition, measures of parental supervision and stability of the home environment might also serve as useful auxiliary variables because the educational literature suggests that these factors influence problem behavior. The previous examples are just a few ideas for auxiliary variables in an educational study, and it is easy to generate additional examples.

In the absence of (or perhaps in addition to) other information, the MCAR tests from Chapter 1 can also identify potential auxiliary variables. Univariate *t* tests are particularly useful in this regard because they can identify variables that are inconsistent with the MCAR mechanism. Recall that the *t*-test procedure separates the missing and complete cases on a variable and examines group mean differences on other variables in the data set. Variables that yield large mean differences are inconsistent with an MCAR mechanism and are potential correlates of missingness. The problem with *t* tests is that they do a poor job of pinpointing the true cause of missingness and can produce a number of spurious associations. Although there is ultimately no harm in choosing the wrong auxiliary variable, focusing on the *t* tests that produce the largest mean differences can help limit the pool of candidate variables.

Finally, it is a good idea to be proactive about satisfying the MAR assumption by collecting variables that are correlates of the analysis variables or correlates of missingness. For example, Graham, Taylor, Olchowski, and Cumsille (2006) suggest that variables such as reading speed and conscientiousness might explain why some respondents leave questionnaire items blank. In a longitudinal study, Schafer and Graham (2002) recommend a survey question that asks respondents to report their likelihood of dropping out of the study prior to the next measurement occasion. Schafer and Graham (2002, p. 173) suggest that collecting data on potential causes of missingness "may effectively convert an MNAR situation to MAR." You should therefore strongly consider this strategy when designing a study.

Practical Considerations

In the context of a multiple imputation analysis, methodologists generally recommend using an extensive set of auxiliary variables. For example, Rubin (1996, p. 479) stated that "the advice has always been to include as many variables as possible when doing multiple imputation." This suggestion is more difficult to implement in a maximum likelihood analysis because the auxiliary variables require a slightly awkward model specification. From a practical standpoint, this means that you may have to limit the number of auxiliary variables in an analysis. It is difficult to establish a rule of thumb for the number of auxiliary variables, but the previous simulation results clearly show that the correlation between an auxiliary variable and the incomplete analysis model variables largely determines the influence of an auxiliary variable. Consequently, a reasonable goal is to maximize the squared multiple correlation between the auxiliary variables and the analysis model variables using as few auxiliary variables as possible. Although there is no harm in using auxiliary variables with low (or zero) correlations (Collins et al., 2001), the most useful auxiliary variables are those that have correlations greater than ± 0.40 with the incomplete analysis variables.

5.5 INCORPORATING AUXILIARY VARIABLES INTO A MAXIMUM LIKELIHOOD ANALYSIS

Returning to the previous health study example, suppose that the researchers want to include age as an auxiliary variable in their regression model. One option is to add age as an additional predictor variable, but this is a bad solution because it accommodates the auxiliary variable by changing the substantive interpretation of the parameter estimates (i.e., the effect of self-esteem becomes a partial regression coefficient if age is a predictor in the model). Instead, the researchers need to incorporate the auxiliary variables in such a way that the interpretation of the parameter estimates is the same as it would have been had the data been complete. They can do this using a structural equation model (Graham, 2003) or a two-stage analysis approach (Savalei & Bentler, 2007; Yuan & Bentler, 2000).

Graham (2003) outlined two structural equation modeling strategies for incorporating auxiliary variables into a maximum likelihood analysis, the **extra dependent variable model** and the **saturated correlates model**. The basic goal of both approaches is to use a series of correlations to work the auxiliary variables into the analysis without altering the substantive interpretation of the parameters. Graham's simulation results favored the saturated correlates model, and this approach is generally easier to implement than the extra dependent variable model. Consequently, I focus on the saturated correlates model in the next section. Interested readers can consult Graham (2003) for details on the extra dependent variable model.

The **two-stage approach** is an alternative method for incorporating auxiliary variables into a maximum likelihood analysis (Savalei & Bentler, 2007; Yuan & Bentler, 2000). As its name implies, the two-stage approach deals with missing data in two steps: the first stage uses maximum likelihood missing data handling to estimate the mean vector and the covariance matrix, and the second stage uses the resulting estimates as input data for a subsequent

analysis (an approach I used to perform an exploratory factor analysis in Chapter 4). The advantage of the two-stage approach is that it can readily incorporate any number of auxiliary variables into the first step of the procedure. Because the mean vector and the covariance matrix reflect the information from the auxiliary variables, there is no need to include the extra variables in the subsequent analysis stage.

Unfortunately, the two-stage approach has a serious drawback that limits its use. Using summary statistics as input data requires a sample size value. However, no single value of N accurately describes the precision of the estimates in the mean vector and the covariance matrix. Therefore, specifying a particular sample size value is likely to bias the standard errors from the analysis stage (Enders & Peugh, 2004). Yuan and Bentler (2000) and Savalei and Bentler (2007) describe a corrective procedure that combines the information matrices from both stages of the analysis, but software programs have yet to implement their approach. Because the two-stage standard errors currently require custom computer programming, I do not discuss the technique in this chapter. However, the two-stage method may become a viable alternative to Graham's (2003) structural equation approach in the near future.

5.6 THE SATURATED CORRELATES MODEL

The saturated correlates model incorporates auxiliary variables via a series of correlations between the auxiliary variables and the analysis model variables (or their residual terms). As you will see, the name "saturated correlates" follows from the fact that the model includes all possible associations among the auxiliary variables as well as all possible associations between the auxiliary variables and the manifest analysis model variables (i.e., the auxiliary variable portion of the model is saturated). Because the rules for incorporating auxiliary variables vary slightly depending on whether the analysis model includes latent variables, I describe these two situations separately.

Manifest Variable Models

To begin, consider an analysis that involves a set of manifest variables (i.e., a statistical model with no latent variables). Graham's (2003) rules for specifying a saturated correlates model are as follows: correlate an auxiliary variable with (1) explanatory variables, (2) other auxiliary variables, and (3) the residual terms of the outcome variables. As an example, consider a multiple regression analysis in which X_1 and X_2 predict Y. Furthermore, suppose that it was of interest to incorporate two auxiliary variables, AV_1 and AV_2, into the regression model. Figure 5.2 shows a path model diagram of the saturated correlates model. Path diagrams use single-headed straight arrows to denote regression coefficients and double-headed curved arrows to represent correlations, and they differentiate manifest variables and latent variables using rectangles and ellipses, respectively (Bollen, 1989; Kline, 2005). The model in Figure 5.2 illustrates all three of Graham's (2003) rules. Specifically, the curved arrows that connect the *AV*s and the *X*s are correlations between the auxiliary variables and the predictors, the curved arrow between AV_1 and AV_2 is the correlation between the auxiliary variables; and the

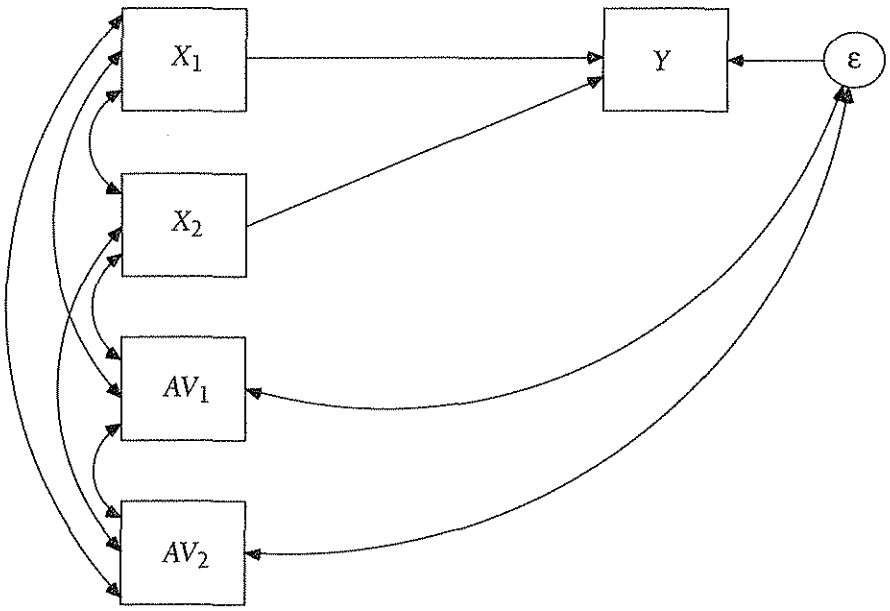

FIGURE 5.2. The saturated correlates version of a manifest variable regression model. The saturated correlates model requires that the auxiliary variables (i.e., AV_1 and AV_2) correlate with (1) one another, (2) the predictor variables, and (3) the residual term.

curved arrows that connect the *AV*s to the ellipse labeled ε are the correlations between the auxiliary variables and the residual term.

The important aspect of the saturated correlates model is that it transmits the information from the auxiliary variables to the analysis model variables without affecting the interpretation of the parameter estimates. Consequently, the interpretation of the regression coefficients is the same as it would have been had the data been complete. For example, the straight arrow that connects X_1 to Y is a partial regression coefficient that quantifies the expected change in Y for a unit increase in X_1 after holding X_2 constant. Adding the auxiliary variables to the model can change the estimated value of this coefficient (e.g., by removing bias or reducing random error), but the interpretation of the slope is unaffected because X_2 is still the only variable being partialled out of Y.

Latent Variable Models

The rules for specifying a latent variable model with auxiliary variables are as follows: correlate an auxiliary variable with (1) manifest predictor variables, (2) other auxiliary variables, and (3) the residual terms of the manifest indicator variables. Note that Graham's (2003) rules describe the linkages between the auxiliary variables and the manifest variables in the model, so that the auxiliary variables never correlate with latent variables or with latent disturbance (i.e., residual) terms. This means that rule 1 applies strictly to models with manifest predictor variables (e.g., so-called multiple indicators and multiple causes, or MIMIC models), and rule 3 applies only to manifest indicators of the latent variables.

To illustrate Graham's (2003) rules, consider a latent variable regression analysis in which LX_1 and LX_2 predict LY. Furthermore, assume that the latent variables each have two

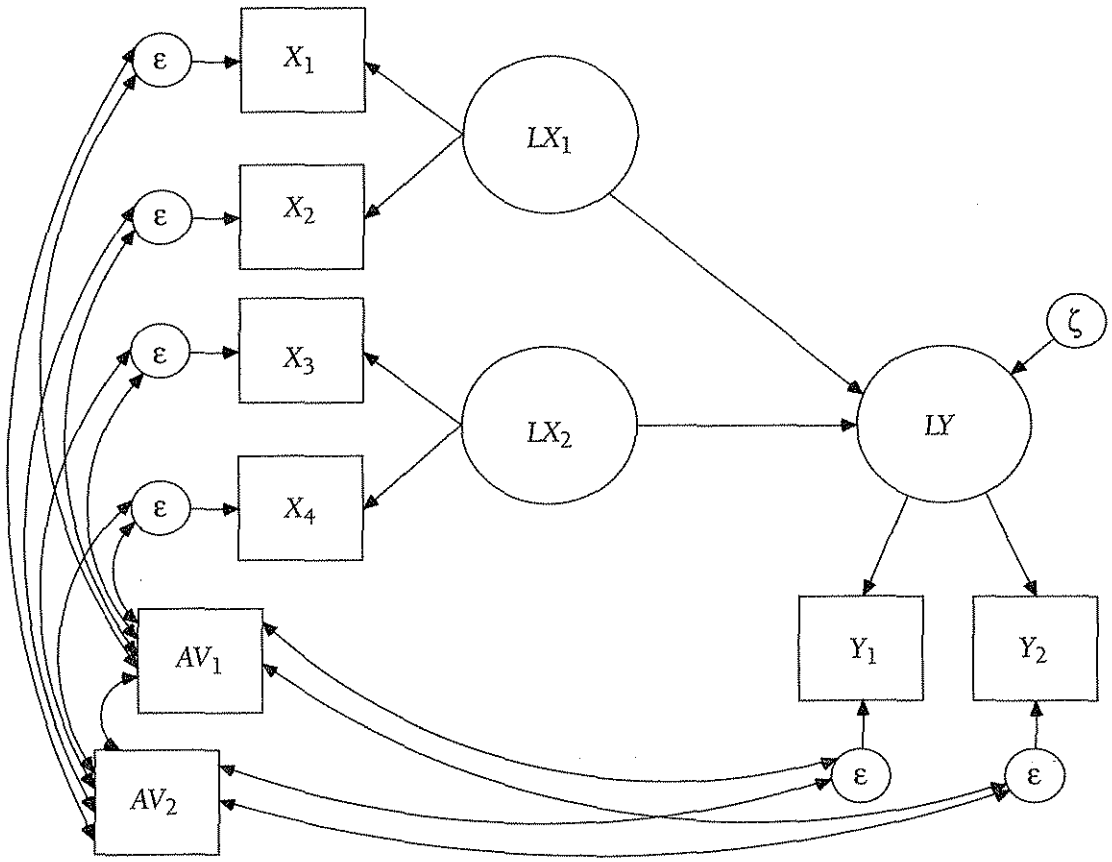

FIGURE 5.3. The saturated correlates version of a latent variable regression model. The saturated correlates model requires that the auxiliary variables (i.e., AV_1 and AV_2) correlate with (1) one another, (2) manifest predictor variables, and (3) the residual terms of the manifest variables. The auxiliary variables never correlate with a latent variable or with a latent variable residual term.

manifest indicators (e.g., *LY* is measured by Y_1 and Y_2) and it is of interest to incorporate two auxiliary variables, AV_1 and AV_2, into the model. Figure 5.3 shows a path model diagram of the saturated correlates model for this analysis. To begin, notice that there are no associations (i.e., curved arrows) between the auxiliary variables and the latent factors. Instead, the curved arrows link the auxiliary variables to the residuals of the six manifest indicators. Like the manifest variable model in Figure 5.2, the curved arrow between AV_1 and AV_2 denotes the correlation between the auxiliary variables. Finally, because there are no manifest predictor variables, rule 1 does not apply to this analysis.

Two additional points are worth noting about the latent variable saturated correlates model. First, like its manifest variable counterpart, the inclusion of the auxiliary variables does not alter the interpretation of the latent variable model parameters. For example, the straight arrow that connects LX_1 with *LY* is a partial regression coefficient that quantifies the expected change in *LY* for a unit increase in LX_1 after holding LX_2 constant. Although the auxiliary variables can change the estimated value of this coefficient, the substantive interpretation of the path is the same as it would have been in a complete-data analysis with no auxiliary variables. Second, the auxiliary variable portion of the path model includes all possible associations between the auxiliary variables and the manifest variables as well as all possible associations among the auxiliary variables (i.e., the auxiliary variable portion of

the model is saturated). This means that the auxiliary variables do not affect the degrees of freedom or the fit of the model.

What If the Auxiliary Variables Have Missing Data?

Including auxiliary variables in an analysis can improve the missing data handling procedure, either by reducing bias (i.e., better approximating the MAR assumption) or by increasing power (i.e., recapturing some of the missing information). Ideally, the auxiliary variables have no missing values, but this need not be the case. From a procedural standpoint, there is nothing special about applying Graham's (2003) specification rules to incomplete auxiliary variables. However, it is reasonable to ask whether it is beneficial to include these variables in the analysis. Although little research has been done on this topic, but the answer appears to be yes.

Enders (2008) used Monte Carlo simulations to examine the impact of including an incomplete auxiliary variable in regression models similar to those in Figures 5.2 and 5.3. Because the auxiliary variable in this study determined missingness and had a strong correlation with the incomplete outcome variable, excluding it from the model produced biased estimates. The simulation results indicated that including the auxiliary variable in the analysis dramatically reduced bias, even when 50% of its scores were missing. Interestingly, the reduction in bias was virtually the same when the missing auxiliary variable was MCAR or MNAR. When the auxiliary variable was MNAR, the auxiliary variable portion of the model (e.g., the correlation between an auxiliary variable and a predictor) was severely biased, but the regression model parameter estimates were quite accurate. Fortunately, bias in the auxiliary variable portion of the model is no problem because these parameters are not of substantive interest.

When deciding whether to use an incomplete auxiliary variable, it is important to examine the proportion of cases that have missing data on both the auxiliary variable and the analysis model variables. When this proportion is high, the amount of information that the auxiliary variable can contribute to the estimation process becomes limited. Establishing definitive guidelines is difficult, but including an auxiliary variable appears to be of little benefit when more than 10% of its observations are concurrently missing with one of the analysis model variables (Enders, 2008).

Computing Incremental Fit Indices

Assessing model fit is an important part of a structural equation modeling analysis. The fact that the saturated correlates model does not change the degrees of freedom implies that the likelihood ratio statistic (i.e., the chi-square test of model fit) and the RMSEA are unaffected by the auxiliary variables. However, the same is not true for incremental (i.e., comparative) fit indices, and it is currently necessary to compute these indices by hand. The idea behind an **incremental fit index** is to compare the relative fit of a hypothesized model (e.g., the latent variable regression model in Figure 5.3) to that of a baseline model. The most common choice of baseline model is a so-called **null model** or **independence model** that estimates the means and the variances of the manifest variables (Bollen, 1989; Kline, 2005).

To illustrate how auxiliary variables affect incremental fit indices, consider the Comparative Fit Index (CFI; Bentler, 1990). The CFI is

$$CFI = \frac{(LR_I - df_I) - (LR_M - df_M)}{(LR_I - df_I)} \quad (5.1)$$

where LR_M and LR_I are the likelihood ratio tests from the hypothesized model and the independence model, respectively, and df_M and df_I are the degrees of freedom for these two models. As I explained previously, the saturated correlates approach does not affect the likelihood ratio test for the hypothesized model because the degrees of freedom are the same with or without the auxiliary variables. However, the standard independence model constrains the auxiliary variable correlations to zero during estimation, which increases the values of the likelihood ratio test and its degrees of freedom. This effectively penalizes the independence model by making its fit worse than it would have been without the auxiliary variables. Consequently, the saturated correlates model artificially inflates the CFI and makes the hypothesized model appear to fit better than it actually does (the same is true for other incremental fit indices).

Fortunately, it is straightforward to compute the correct CFI value after fitting a special independence model that estimates the auxiliary variable correlations and constrains the correlations among the analysis model variables to zero. To illustrate, reconsider the saturated correlates model in Figure 5.3. The top panel of Figure 5.4 shows the standard independence model, and the bottom panel of the figure shows the correct independence model for this analysis. Notice that the correct independence model includes the same number of auxiliary variable correlations as the latent variable model in Figure 5.3, so the auxiliary variables exert a constant influence on the fit of both models. Substituting the likelihood ratio test and the degrees of freedom from the modified independence model into Equation 5.1 yields the correct CFI value. Structural equation modeling programs provide these quantities as part of their standard output; these computations are illustrated in an analysis example later in the chapter.

In general, applying the following steps to the manifest variables in the analysis yields an appropriate independence model: (1) estimate the variance of all variables, (2) estimate the correlations among the manifest predictors, (3) fix the correlations between manifest predictors and the outcomes to zero, (4) fix the correlations among the outcome variables to zero, (5) estimate the correlations between the auxiliary variables and all other variables, and (6) estimate the correlations among the auxiliary variables. These rules are applicable to situations in which the standard independence model would have been appropriate, had there been no auxiliary variables. However, the standard independence model is not appropriate for all circumstances (Widaman & Thompson, 2003), so you may need to modify these rules accordingly.

Limitations of the Saturated Correlates Model

The saturated correlates model has a number of practical limitations. In my experience, using a large set of auxiliary variables can lead to estimation problems and convergence failures.

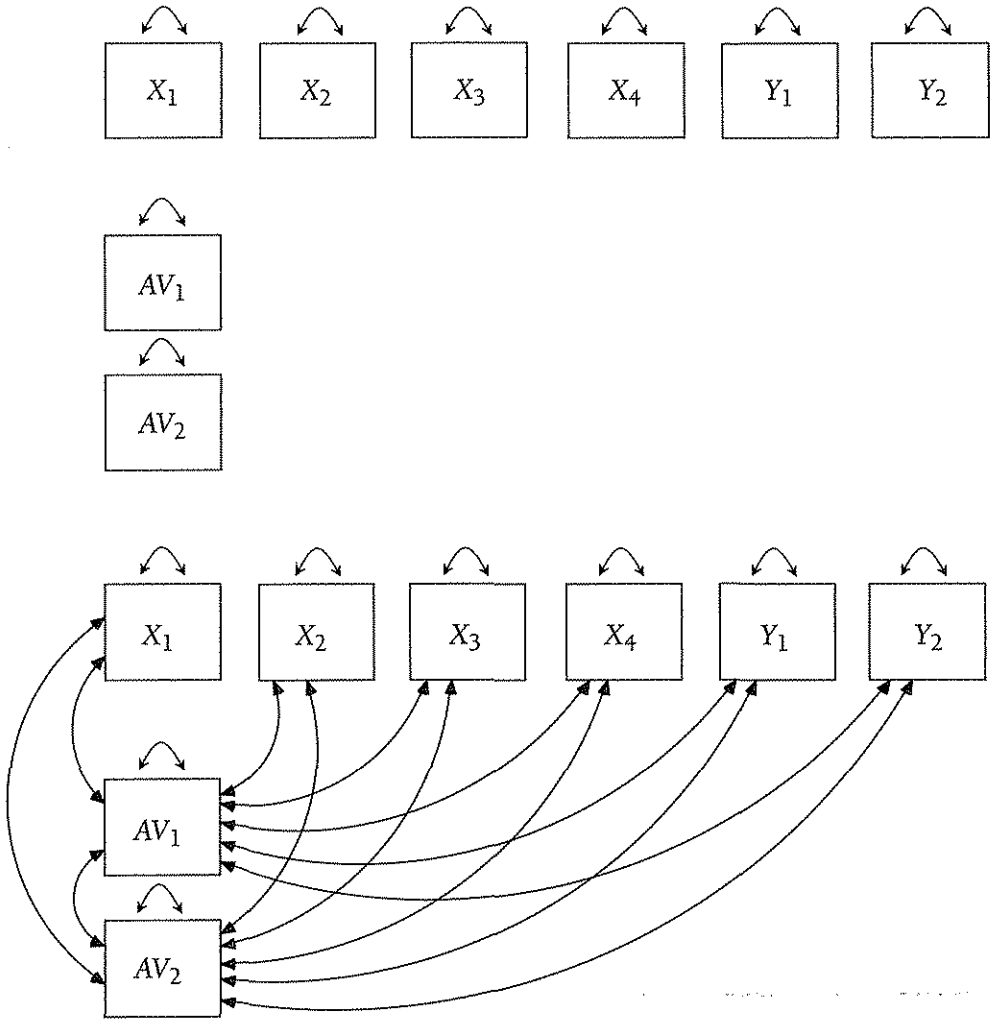

FIGURE 5.4. The top panel of the figure shows the standard independence model where all of the variables are uncorrelated. A double-headed curved arrow that connects a variable to itself denotes a variance. The bottom panel of the figure shows the modified independence model that adds covariances between the auxiliary variables and the analysis model variables.

Although it is not necessary for the auxiliary variables to have complete data, incomplete auxiliary variables can exacerbate these problems. Savalei and Bentler (2007) note that convergence problems are also related to small residual variance terms. In some situations, rescaling the variables to have a similar metric (e.g., by multiplying or dividing a variable by a constant) can alleviate these estimation problems, but reducing the number of auxiliary variables is often the best option. When you do have to reduce the number of auxiliary variables, it is a good idea to retain the variables that have the highest correlations with the incomplete variables in the analysis model.

Even when the saturated correlates model converges to a proper solution, some software programs issue a warning message indicating that the solution is invalid. Structural equation modeling programs use a set of parameter matrices to represent the analysis model, and the pattern of associations required by the saturated correlates approach can produce nonpositive definite matrices. In particular, the correlations between the auxiliary variables and the

residual terms often lead to dire warning messages about the residual covariance matrix (i.e., the so-called psi matrix). Although computer programs tend to issue warning messages when a solution produces a nonpositive definite matrix, these messages are usually benign when they are caused by an auxiliary variable. If the model produces valid parameter estimates (e.g., parameter estimates that seem reasonable, no negative variance estimates) and converges properly without the auxiliary variables, then you can generally ignore these warning messages and interpret your results.

5.7 THE IMPACT OF NON-NORMAL DATA

The multivariate normal distribution plays an integral role in every phase of a maximum likelihood analysis (e.g., the log-likelihood provides a basis for identifying the most likely population parameter values, and the second derivatives of the normal distribution are the building blocks for standard errors). In practice, non-normal data are relatively common, and some authors argue that normality is the exception rather than the rule (Micceri, 1989). Given the important role that the normal distribution plays in the estimation process, it is reasonable to ask whether normality violations are problematic for maximum likelihood analyses.

A good deal of research has examined the impact of non-normality on complete-data estimation (Chou, Benter, & Satorra, 1991; Curran, West, & Finch, 1996; Finch, West, & MacKinnon, 1997; Hu, Bentler, & Kano, 1992; Yuan, Bentler, & Zhang, 2005), and an increasing number of studies have investigated the issue of non-normal missing data (Enders, 2001, 2002; Gold & Bentler, 2000; Graham, Hofer, & MacKinnon, 1996; Savalei, 2008; Savalei & Bentler, 2005, 2007; Yuan, 2007; Yuan & Bentler, 2000). This literature suggests that non-normal data tend to have a minimal impact on the parameter estimates themselves but can bias standard errors and distort the likelihood ratio test. Yuan et al. (2005) showed that kurtosis largely dictates this bias, whereby leptokurtic data can attenuate standard errors and inflate the likelihood ratio test, and platykurtic data can inflate standard errors and reduce the magnitude of the likelihood ratio statistic. Simulation studies suggest that skewness can also exert a negative impact, particularly with sample sizes that are common in the behavioral and the social sciences. Interested readers can consult Finney and DiStefano (2006) and West, Finch, and Curran (1995) for a detailed review of the non-normality literature.

Corrective procedures for complete data have been available for some time (Bentler, 1983; Bollen & Stine, 1992; Browne, 1984; Satorra & Bentler, 1988, 1994, 2001), and methodologists have extended many of these procedures to missing data analyses (Arminger & Sobel, 1990; Enders, 2002; Savalei & Bentler, 2007; Yuan & Bentler, 2000). The subsequent sections illustrate some of these corrective procedures. In particular, I describe two approaches for estimating standard errors ("robust" standard errors and bootstrap resampling) and two methods for correcting the bias in the likelihood ratio test (rescaling and bootstrap resampling). I chose these procedures because they are readily available in software packages, but other techniques will likely become available in the near future (e.g., distribution-free test statistics that do not rely on the multivariate normality assumption; Yuan & Bentler, 2000; Savalei & Bentler, 2007).

5.8 ROBUST STANDARD ERRORS

Recall from Chapter 3 that the curvature of the log-likelihood function dictates the magnitude of the maximum likelihood standard errors. In particular, the matrix of second derivatives (i.e., the Hessian matrix) plays an important role in the standard error computations. The second derivative formulas from the normal distribution are missing terms that depend on the skewness and the kurtosis of the population distribution. The absence of these terms can overestimate or underestimate the standard errors, depending on whether the population data are leptokurtic or platykurtic. The robust standard error formula combines information from the first and the second derivatives into a single estimate of sampling error, such that the information from the first derivatives effectively serves as an adjustment term that corrects for normality violations.

To illustrate robust standard errors, I use a univariate example that involves the standard error of the mean. This is not an ideal example because normality violations only affect the standard errors for variance and covariance parameters (White, 1982; Yuan et al., 2005). However, an algebraic relationship between the first and second derivative formulas plays an important role in the formulation of the robust standard errors, and illustrating this relationship is more tedious with covariance matrix parameters. Nevertheless, the computations in the subsequent sections readily extend to covariance matrix parameters as well as to multivariate estimation problems. For simplicity, I use complete-data formulas throughout this section, but the underlying logic is the same with missing data. Some of the information in the subsequent sections is relatively technical, so readers who are not interested in the mathematical underpinnings of robust standard errors may want to skim this section.

First and Second Derivatives Revisited

Understanding the robust standard error computations requires some additional background information on derivatives. From Chapter 3, the equations for the first and second derivatives of the sample log-likelihood function with respect to the mean are as follows:

$$\frac{\partial \log L}{\partial \mu} = \frac{1}{\sigma^2}\left(-N\mu + \sum_{i=1}^{N} y_i\right) \tag{5.2}$$

$$\frac{\partial^2 \log L}{\partial^2 \mu} = \frac{-N}{\sigma^2} \tag{5.3}$$

The first derivative equation was useful for identifying the maximum of the log-likelihood function (e.g., by setting the formula to zero and solving for the population mean), and the second derivative equation quantified the curvature of the function.

Recall from Chapters 3 and 4 that the sample log-likelihood is the sum of N individual log-likelihood equations. The individual log-likelihood equation (e.g., Equation 4.1) also has first and second derivatives, and Equations 5.2 and 5.3 are actually sums of these casewise

derivative equations. Rewriting the previous equations as sums shows each case's contribution to the derivative formulas, as follows:

$$\frac{\partial \log L}{\partial \mu} = \frac{1}{\sigma^2}\left(-N\mu + \sum_{i=1}^{N} y_i\right) = \sum_{i=1}^{N}\left[\frac{1}{\sigma^2}(y_i - \mu)\right] = \sum_{i=1}^{N}\left[\frac{\partial \log L_i}{\partial \mu}\right] \tag{5.4}$$

$$\frac{\partial^2 \log L}{\partial^2 \mu} = \frac{-N}{\sigma^2} = \sum_{i=1}^{N}\left[\frac{-1}{\sigma^2}\right] = \sum_{i=1}^{N}\left[\frac{\partial^2 \log L_i}{\partial^2 \mu}\right] \tag{5.5}$$

where the bracketed terms contain the derivative formulas for the individual log-likelihood equation. In words, the right-most terms in the equations say that the derivatives of the sample log-likelihood function equal the sum of the derivatives of the individual log-likelihood equations. (In calculus, the derivative of a sum equals the sum of the derivatives.) The individual contributions to the derivative equations play an important role in the formulation of robust standard errors.

Two Formulations of Information

Recall that the information matrix contains values that quantify the curvature of the log-likelihood function (e.g., peaked functions produce large information values and small standard errors, whereas flat functions produce small information values and large standard errors). When the population data are multivariate normal, there are two equivalent methods for computing information. The method from previous chapters is based on second derivatives, whereas the alternate approach uses first derivatives. As an illustration, consider each case's contribution to the first derivative equation, $(y_i - \mu)/\sigma^2$. This collection of terms is itself a variable, the value of which varies across cases. Applying expectation rules for variables, note that the variance of the casewise first derivative is as follows.

$$\text{var}\left(\frac{\partial \log L_i}{\partial \mu}\right) = \text{var}\left[\frac{1}{\sigma^2}(y_i - \mu)\right] = \left(\frac{1}{\sigma^2}\right)^2 \text{var}(y_i - \mu) = \frac{1}{\sigma^4}(\sigma^2) = \frac{1}{\sigma^2} \tag{5.6}$$

Notice that the result of Equation 5.6 is virtually identical to the second derivative of the individual log-likelihood equation (i.e., the bracketed terms in Equation 5.5), but differs by a multiplicative constant of negative 1. Stated differently, computing the variance of the individual first derivatives gives each case's contribution to the second derivative equation, as follows:

$$\text{var}\left(\frac{\partial \log L_i}{\partial \mu}\right) = -\left(\frac{\partial^2 \log L_i}{\partial^2 \mu}\right) \tag{5.7}$$

The equality in Equation 5.7 leads to two equivalent expressions for information, one of which is based on first derivatives (i.e., I_F) and the other on second derivatives (i.e., I_S).

$$I_F = N\left[\text{var}\left(\frac{\partial \log L_i}{\partial \mu}\right)\right] = N\left(\frac{1}{\sigma^2}\right) \tag{5.8}$$

$$I_S = -N\left[\frac{\partial^2 \log L_i}{\partial^2 \mu}\right] = -\frac{\partial^2 \log L}{\partial^2 \mu} = \frac{N}{\sigma^2} \tag{5.9}$$

The information values provide the building blocks for maximum likelihood standard errors, such that the inverse (i.e., reciprocal) of information is the sampling variance, and the square root of the sampling variance is the standard error. When the population data are normally distributed, the two formulations of information are equivalent and should produce very similar standard errors.

The Sandwich Estimator

When the population data are non-normal, Equations 5.8 and 5.9 are no longer equivalent, and both formulas yield inaccurate standard errors. The problem is that the density function for the normal distribution generates the derivative equations, when some other non-normal distribution should have generated these formulas. From a practical standpoint, the derivative formulas are missing terms that depend on the skewness and kurtosis of the population distribution (White, 1982; Yuan et al., 2005). The absence of these terms can overestimate or underestimate the standard errors, depending on whether the population data are leptokurtic or platykurtic (Yuan et al., 2005).

The robust standard error combines the two information equations into a single estimate of sampling error (Freedman, 2006; Huber, 1967; White, 1982). This so-called **sandwich estimator** of the sampling variance is

$$\text{var}(\hat{\theta}) = I_S^{-1} I_F I_S^{-1} \tag{5.10}$$

where I_F and I_S are information estimates based on the first and second derivatives, respectively, and $\text{var}(\hat{\theta})$ is the sampling variance. The methodological literature refers to Equation 5.10 as the sandwich estimator because it resembles a piece of meat (i.e., I_F) sitting between two slices of bread (i.e., I_S). Consistent with previous chapters, taking the square root of the sampling variance gives the standard error.

How Does the Robust Standard Error Work?

The "meat" of the sandwich estimator is important because it effectively serves as a correction factor that increases or decreases the standard error, depending on the kurtosis of the data. Returning to Equation 5.4, notice that the first derivative of the individual log-likelihood equation involves the deviation between a score and the mean (i.e., $y_i - \mu$). Because a leptokurtic distribution has thicker tails (i.e., a higher proportion of large deviation values) than a normal curve, the presence of outliers increases the value of I_F relative to that of a normal distribution. Consequently, multiplying by I_F increases the magnitude of the standard error

and counteracts the negative bias that results from leptokurtic data. In contrast, a platykurtic distribution has fewer extreme scores than a normal curve and thus produces a smaller value of I_F. In this situation, using I_F as multiplier produces a downward adjustment that decreases the inflation in the normal-theory standard errors. Finally, when the data are normally distributed, the sandwich estimator reduces to the usual maximum likelihood standard error because the first two terms of Equation 5.10 cancel out when $I_F = I_S$. Although the first derivative formulas for the covariance matrix parameters are more complex than those of the mean parameters, they too contain deviation scores. Consequently, the basic operation of the sandwich estimator is the same for any parameter. Finally, note that the basic form of Equation 5.10 is the same in multivariate estimation problems. In the multivariate context, I_F and I_S are information matrices and $\text{var}(\hat{\theta})$ is the parameter covariance matrix, the diagonal of which contains the squared standard errors.

A Bivariate Example

To further illustrate the robust standard errors, I generated a single artificial data set with two variables and $N = 500$ cases. The purpose of this example is to illustrate the impact of nonnormal data, so I generated X to have a platykurtic distribution with kurtosis of –1.00 and Y to have a leptokurtic distribution with kurtosis of 4.00 (a normal distribution has kurtosis of zero). I used maximum likelihood to estimate the mean vector and the covariance matrix and obtained the normal-theory and robust standard errors for each parameter estimate.

Table 5.2 shows the parameter covariance matrices from the bivariate analysis. The parameter covariance matrix is a 5 by 5 symmetric matrix, where each row and column corresponds to one of the estimated parameters (there are two means and three unique covariance matrix elements). The diagonals of the two matrices are particularly important because they contain the sampling variances (i.e., squared standard errors). The mean parameters are un-

TABLE 5.2. Parameter Covariance Matrices for the Kurtotic Data Analysis Example

Parameter	1	2	3	4	5
	Parameter covariance matrix (normality assumed)				
1: μ_X	0.002119				
2: μ_Y	0.000928	0.001864			
3: σ^2_X	0	0	0.004490		
4: $\sigma_{X,Y}$	0	0	0.001967	0.002406	
5: σ^2_Y	0	0	0.000862	0.001730	0.003474
	Parameter covariance matrix (robust)				
1: μ_X	0.002119				
2: μ_Y	0.000928	0.001864			
3: σ^2_X	0.000036	–0.000096	0.002117		
4: $\sigma_{X,Y}$	–0.000096	–0.000091	0.001094	0.002388	
5: σ^2_Y	–0.000091	–0.000111	0.000841	0.003494	0.010259

Note. X has a platykurtic distribution ($K = 2$) and Y has a leptokurtic distribution ($K = 7$). Bold typeface denotes the sampling variance (i.e., squared standard error) of each parameter estimate.

affected by nonnormal data (White, 1982; Yuan et al., 2005), so these elements are identical in both matrices. However, the covariance matrix elements are quite different. For example, notice that the normal-theory standard error for σ_X^2 is larger than the corresponding robust standard error (i.e., $\sqrt{.004490} = 0.067$ versus $\sqrt{.002117} = 0.046$, respectively), whereas the standard error for σ_Y^2 is smaller than that of the robust estimator (i.e., $\sqrt{.003474} = 0.059$ versus $\sqrt{.010259} = 0.101$, respectively). This pattern of differences is consistent with the notion that leptokurtic data can attenuate standard errors and platykurtic data can inflate standard errors (Yuan et al., 1995). Of course, analyzing a single sample provides very little evidence about the relative accuracy of the two estimators, but published computer simulation studies clearly favor robust standard errors (Chou & Bentler, 1995; Chou et al., 1991; DiStefano, 2002; Yuan & Bentler, 1997).

Robust Standard Errors for Missing Data

The formulation of the sandwich estimator is identical with missing data (Arminger & Sobel, 1990; Yuan & Bentler, 2000). However, missing data introduce one small nuance that is not necessarily an issue with complete data. You might recall from Chapter 4 that the observed information matrix produces standard errors that are valid with MAR data (Kenward & Molenberghs, 1998), whereas the expected information matrix requires MCAR data. This implies that the observed information matrix is the appropriate "bread" for the sandwich estimator in Equation 5.10. At the time of this writing, most software programs that implement robust standard errors for missing data use the observed information matrix for this purpose, and simulation studies suggest that this approach provides a substantial improvement over normal-theory standard errors (Enders, 2001).

5.9 BOOTSTRAP STANDARD ERRORS

Bootstrap resampling is a second approach to generating standard errors with nonnormal data. The bootstrap is quite different from the sandwich estimator because it uses Monte Carlo simulation techniques to generate an empirical sampling distribution for each parameter, the standard deviation of which is the standard error. Because the bootstrap makes no distributional assumptions, the accuracy of the procedure is unaffected by normality violations. This section describes a so-called **naïve bootstrap** that is strictly limited to estimating standard errors. Later in the chapter, I describe an alternate bootstrap procedure (the Bollen-Stine bootstrap) that can correct for bias in the likelihood ratio test. A number of detailed overviews of the bootstrap are available in the literature for readers interested in more details (e.g., Bollen & Stine, 1992; Efron & Tibshirani, 1993; Enders, 2002; Stine, 1989). The vast majority of the bootstrap literature deals with complete-data applications, but the procedural details are essentially the same with or without missing data.

The basic idea behind **bootstrap resampling** is to repeatedly draw samples of size *N with replacement* from a data set. In effect, the sample data serve as a miniature population for the Monte Carlo sampling procedure. Because the samples are drawn with replacement, some data records will appear more than once in a given sample, whereas others will not appear at

TABLE 5.3. Bootstrap Sample from an Employee Selection Data Set

Sample data				Bootstrap sample			
ID	*IQ*	*JP*	*WB*	*ID*	*IQ*	*JP*	*WB*
1	78	—	13	18	115	14	14
2	84	—	9	14	106	15	10
3	84	—	10	17	113	12	14
4	85	—	10	3	84	—	10
5	87	—	—	16	112	10	10
6	91	—	3	7	92	—	12
7	92	—	12	3	84	—	10
8	94	—	3	15	108	10	—
9	94	—	13	14	106	15	10
10	96	—	—	10	96	—	—
11	99	7	6	8	94	—	3
12	105	10	12	16	112	10	10
13	105	11	14	14	106	15	10
14	106	15	10	10	96	—	—
15	108	10	—	9	94	—	13
16	112	10	10	1	78	—	13
17	113	12	14	4	85	—	10
18	115	14	14	16	112	10	10
19	118	16	12	10	96	—	—
20	134	12	11	18	115	14	14

Note. *JP* = job performance; *WB* = well-being.

all. Table 5.3 shows a single bootstrap sample from the small employee selection data set that I have been using throughout the book. Notice that case 14 appears three times in the bootstrap sample, whereas case 2 does not appear at all.

The ultimate goal of the bootstrap is to construct an empirical sampling distribution for each parameter estimate. Drawing a large number of bootstrap samples (e.g., $B = 2000$) and fitting the analysis model to each sample yields a set of estimates for each parameter. The collection of B parameter estimates forms an empirical sampling distribution, the standard deviation of which is the bootstrap standard error

$$SE_{\text{Bootstrap}} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b - \bar{\theta})^2} \tag{5.11}$$

where B is the number of bootstrap samples, $\hat{\theta}_b$ is the parameter estimate from one of the bootstrap samples, and $\bar{\theta}$ is the mean of the B parameter estimates. Notice that Equation 5.11 is the usual formula for the sample standard deviation, where the B parameter estimates serve as data points. Although the process of repeatedly drawing samples and analyzing the data sounds tedious, software packages that implement the bootstrap completely automate the procedure. It is also relatively straightforward to implement the bootstrap in software packages that do not have built-in routines (Enders, 2005).

A Bivariate Analysis Example

To illustrate the bootstrap, reconsider the artificial bivariate data set from the previous section. Recall that X has platykurtic distribution with kurtosis equal to –1.00 and Y has a leptokurtic distribution with kurtosis of 4.00. To begin, I drew 2,000 samples of $N = 500$ (the size of the original sample) with replacement from the data and subsequently used maximum likelihood to estimate the mean vector and the covariance matrix from each sample. This procedure produced 2,000 estimates of each mean and covariance parameter. Next, I generated standard errors by computing the standard deviation of each parameter across the 2,000 bootstrap samples. The bootstrap produced standard errors that are nearly identical to those of the sandwich estimator. For example, the bootstrap standard errors for σ_X^2 and σ_Y^2 are 0.046 and 0.099, respectively, and the corresponding robust standard errors are 0.046 and 0.101. It is not unusual for the two procedures to produce similar estimates, particularly when the sample size is relatively large. Consequently, convenience is often the only reason to prefer one approach to another.

Bootstrap Confidence Intervals

There are two methods for constructing bootstrap confidence intervals. Following the usual complete-data procedure, the first approach is to multiply the bootstrap standard error by the appropriate critical value from the unit normal table:

$$CI_{Bootstrap} = \hat{\theta} + (z_{1-\alpha/2})(SE_{Bootstrap}) \tag{5.12}$$

where $\hat{\theta}$ is the parameter estimate from the initial maximum likelihood analysis, and $z_{1-\alpha/2}$ is the two-tailed critical value (e.g., $z = 1.96$ for an alpha level of .05). Alternatively, the parameter estimates that correspond to the 95th and the 5th percentiles of the bootstrap sampling distribution can define the upper and lower confidence interval limits, respectively. Little and Rubin (2002) suggest that the former approach is appropriate when the empirical sampling distribution is approximately normal, but they prefer the second method when the distribution is non-normal. You can readily ascertain the shape of the empirical sampling distribution and determine the appropriate percentiles by examining a frequency distribution of the B parameter estimates.

How Many Bootstrap Samples Should I Use?

It is difficult to establish a good rule of thumb for the number of bootstrap samples because any such recommendation is a bit arbitrary. In part, this decision depends on the shape of the empirical sampling distribution. When the sampling distribution approximates a normal curve, Little and Rubin (2002, p. 197) suggest that a relatively small number of bootstrap samples will suffice. In contrast, they recommend using a large number of samples (e.g., $B >$ 2000) when the empirical sampling distribution is non-normal. Of course, the problem with these recommendations is that you cannot determine the shape of the empirical sampling

distribution without first running the bootstrap procedure. With many analysis models, implementing the bootstrap takes very little time, so there is often no practical reason to avoid using a very large number of bootstrap samples.

Limitations of the Bootstrap

The bootstrap procedure is advantageous because it requires no distributional assumptions. However, treating the sample data as a miniature population effectively assumes that the sample is a representative surrogate for the entire population. This assumption is tenuous in its own right, particularly in small samples. Another issue to be aware of is that a subset of the bootstrap samples may produce analyses that fail to converge. Small samples and misspecified models are common causes of convergence failures, and missing data only exacerbate the problem. Discarding the failed replicates is a common way to deal with convergence failures (Yung & Bentler, 1996), but methodologists have proposed other options (Yuan & Hayashi, 2003).

5.10 THE RESCALED LIKELIHOOD RATIO TEST

When the multivariate normality assumption is violated, the sampling distribution of the likelihood ratio test no longer follows the appropriate central chi-square distribution. With univariate population data, the likelihood ratio test is proportional to kurtosis, such that leptokurtic data inflate the test statistic, and platykurtic data attenuate its value (Yuan et al., 2005). Consequently, the likelihood ratio test can yield excessive type I or type II error rates, depending on the population kurtosis. The nature of the bias becomes more complex in multivariate analyses, but the underlying problem remains the same—the likelihood ratio test does not follow its theoretical sampling distribution. One solution to this problem is to rescale the likelihood ratio test so that it more closely approximates the appropriate chi-square distribution. This correction has been available for some time (Satorra & Bentler, 1988, 1994), although its application to missing data analyses is more recent (Yuan & Bentler, 2000). The limited research to date suggests that the rescaling procedure for missing data effectively controls the error rates of the likelihood ratio test (Enders, 2001; Savalei & Bentler, 2005). Because the logic of the rescaling process is the same with or without missing data, this section gives a generic description of the procedure. Yuan and Bentler (2000) give additional technical details on the rescaling procedure for missing data.

The Satorra–Bentler Chi-Square

Readers who use structural equation modeling techniques may already be familiar with the rescaled likelihood ratio statistic. In this context, the so-called **Satorra–Bentler chi-square** (Satorra & Bentler, 1988, 1994) uses a correction factor to rescale the likelihood ratio test, as follows:

$$LR_{RS} = cLR \tag{5.13}$$

In a structural equation modeling analysis, LR_{RS} is the rescaled (i.e., Satorra–Bentler) test statistic, LR is a likelihood ratio test that compares the relative fit of the hypothesized model (e.g., a confirmatory factor analysis model) to that of a saturated model (e.g., a model that estimates the sample covariance matrix), and c is a scaling factor that depends on the distribution shape. With univariate data, Yuan et al. (2005) show that the scaling factor is related to kurtosis, such that c decreases the value of LR when the distribution is leptokurtic and increases the test statistic when the distribution is platykurtic. When the population data are normally distributed, the scaling factor equals one, and the rescaled statistic is identical to the usual likelihood ratio test.

A General Rescaling Procedure

Structural equation modeling applications of the rescaled test statistic are particularly straightforward to implement because software packages automatically perform the rescaling. The rescaling procedure is applicable to any likelihood ratio test, but implementing it requires special procedures (Satorra & Bentler, 2001). Recall from Chapter 3 that the likelihood ratio test is

$$LR = -2(\log L_{Restricted} - \log L_{Full}) \tag{5.14}$$

where $\log L_{Full}$ and $\log L_{Restricted}$ are the log-likelihood values from the full and restricted models, respectively. The scaling factor for the likelihood ratio test incorporates information from both models, as follows:

$$c_{LR} = \frac{(q_{Restricted})(c_{Restricted}) - (q_{Full})(c_{Full})}{(q_{Restricted} - q_{Full})} \tag{5.15}$$

where $q_{Restricted}$ is the number of parameter estimates from the restricted model, $c_{Restricted}$ is the scaling factor for the restricted model, q_{Full} is the number of estimated parameters in the full model, and c_{Full} is the scaling factor for the full model. Computing c_{LR} is straightforward because software packages that implement the rescaling procedure report all of the necessary terms. Finally, the rescaled test statistic divides the likelihood ratio test by the scaling factor, as follows:

$$LR_{RS} = \frac{-2(\log L_{Restricted} - \log L_{Full})}{c_{LR}} = \frac{LR}{c_{LR}} \tag{5.16}$$

I illustrate the rescaled likelihood ratio test in one of the analysis examples presented later in the chapter.

5.11 BOOTSTRAPPING THE LIKELIHOOD RATIO STATISTIC

Bootstrap resampling is a second option for correcting bias in the likelihood ratio test statistic. Whereas the rescaling procedure attempts to adjust the value of the likelihood ratio test so that it more closely approximates its theoretical sampling distribution, the bootstrap leaves the test statistic intact and uses Monte Carlo simulation techniques to generate a new sampling distribution. Rather than correcting the test statistic itself, the bootstrap corrects the probability value for the test by referencing the likelihood ratio statistic to the empirical sampling distribution. Although the actual resampling procedure is identical to that of the naïve bootstrap, obtaining the correct empirical distribution requires a transformation of the data prior to drawing the samples.

The Problem with the Naïve Bootstrap

Using the naïve bootstrap to construct a sampling distribution for the likelihood ratio test is inappropriate because the resulting samples are inconsistent with the null hypothesis. For example, suppose that it was of interest to use the likelihood ratio to test the slope coefficient from a simple regression model. This test involves a comparison of the regression model (i.e., the full model) to a restricted model that constrains the regression coefficient to zero during estimation. The null hypothesis for this test states that the population regression coefficient is equal to zero (i.e., the restricted model is true), and the *p*-value quantifies the probability of observing a likelihood ratio test that is equal to or greater than that of the sample data, given that the null hypothesis is true.

Fitting the two regression models to a large number of bootstrap samples and computing the likelihood ratio test for each sample would not produce an appropriate sampling distribution because the collection of test statistics is inconsistent with the null hypothesis. Even if the regression slope is truly zero in the population, the sample estimate is unlikely to exactly equal zero. Consequently, drawing bootstrap samples from data yields a distribution that reflects natural sampling fluctuation as well as model misfit (i.e., the discrepancy between the data and the null hypothesis). The appropriate sampling distribution should reflect sampling fluctuation only.

The Bollen–Stine Bootstrap

Beran and Srivastava (1985) and Bollen and Stine (1992) modified the bootstrap procedure by applying an algebraic transformation to the data prior to drawing samples. This transformation aligns the mean and the covariance structure of the data to the null hypothesis and produces a distribution that reflects only the sampling fluctuation of the likelihood ratio statistic. Because the transformation does not affect distribution shape, the bootstrap procedure effectively incorporates the influence of nonnormal data. Consequently, referencing the likelihood ratio test to the empirical sampling distribution of likelihood ratio statistics can generate an accurate probability value, even when the data are nonnormal. I refer to the modified bootstrap procedure as the **Bollen–Stine bootstrap** throughout the remainder of the chapter

because the Bollen and Stine (1992) manuscript was largely responsible for popularizing the technique, particularly in structural equation modeling applications.

To align the data with the null hypotheses, Bollen and Stine transform the sample data to have the same mean and covariance structure as the restricted model. This transformation requires the mean vector and the covariance matrix from the sample data as well as the mean vector and the covariance matrix that would result if the null hypothesis were true (i.e., the model-implied mean vector and covariance matrix from the restricted model). Both sets of estimates are readily available from structural equation modeling programs.

The Bollen–Stine transformation is as follows:

$$\mathbf{Z}_i = (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_S)^T \hat{\boldsymbol{\Sigma}}_S^{-1/2} \hat{\boldsymbol{\Sigma}}_R^{1/2} + \hat{\boldsymbol{\mu}}_R^T \tag{5.17}$$

where $\mathbf{Z}_i$ is the transformed data vector for case i, $\mathbf{Y}_i$ is the raw data vector for case i, $\hat{\boldsymbol{\mu}}_S$ is the sample mean vector, $\hat{\boldsymbol{\Sigma}}_S$ is the sample covariance matrix, $\hat{\boldsymbol{\Sigma}}_R$ is the implied covariance matrix from the restricted model, and $\hat{\boldsymbol{\mu}}_R$ is the implied mean vector from restricted model. In words, the $(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_S)^T \hat{\boldsymbol{\Sigma}}_S^{-1/2}$ portion of the formula essentially "erases" the mean and the covariance structure of the sample data and converts the variables to uncorrelated z scores. Next, multiplying the z scores by $\hat{\boldsymbol{\Sigma}}_R^{1/2}$ transforms the data to have the same covariance matrix as the restricted model. Finally, adding $\hat{\boldsymbol{\mu}}_R$ equates the sample means to the predicted means from the restricted model.

The Bollen–Stine transformation produces a data set that is exactly consistent with the null hypothesis. After applying the transformation, the bootstrap procedure is the same as before. The specific steps are as follows: (1) draw B bootstrap samples with replacement from the transformed data set, (2) fit the full model and the restricted model to each bootstrap sample, (3) compute the likelihood ratio statistic for each bootstrap sample, and (4) construct a frequency distribution of the B likelihood ratio statistics. The proportion of bootstrap test statistics that exceed the value of the original likelihood ratio test serves as the corrected probability value.

A Bivariate Example

To illustrate the Bollen–Stine bootstrap, reconsider the bivariate data set from the previous sections. Recall that X has platykurtic distribution with kurtosis of –1.00 and Y has a leptokurtic distribution with kurtosis of 4.00. Furthermore, suppose that it is of interest to use the likelihood ratio statistic to test the slope from the regression of Y on X. As I explained previously, this test involves a comparison of the regression model (i.e., the full model) to a restricted model that constrains the regression coefficient to zero during estimation. The first step is to estimate the two regression models and compute the likelihood ratio test. Doing so yields a likelihood ratio statistic of LR = 123.05. Normally, a central chi-square distribution with one degree of freedom would generate the probability value for the test. However, the theoretical chi-square distribution is likely to produce an inaccurate probability value because the data are nonnormal. The purpose of the bootstrap is to generate an empirical sampling distribution that reflects the influence of the nonnormal data.

TABLE 5.4. Mean Vectors and Covariance Matrices for the Bollen–Stine Transformation

Variable	1	2	1	2	1	2
	Sample data		Restricted model		Transformed data	
1: X	1.060		1.060		1.060	
2: Y	0.464	0.932	0	0.932	0	0.931
Means	−0.030	0.045	−0.030	0.045	−0.030	0.046

The null hypothesis for the likelihood ratio test states that the population regression coefficient is equal to zero (i.e., the restricted model is true). The left-most section of Table 5.4 shows the mean vector and the covariance matrix for the sample data (i.e., $\hat{\boldsymbol{\mu}}_S$ and $\hat{\boldsymbol{\Sigma}}_S$, respectively). Notice that the two variables have a positive covariance, so the data are not perfectly consistent with the null hypothesis. (If the population regression coefficient is zero, this covariance should also equal zero.) Transforming the data to have a perfect fit to the null hypothesis requires the predicted mean vector and the predicted covariance matrix from the restricted model. I used a structural equation program to estimate the restricted model, and the middle section of Table 5.4 shows model-implied parameter estimates, $\hat{\boldsymbol{\mu}}_R$ and $\hat{\boldsymbol{\Sigma}}_R$. Notice that the covariance is zero because the restricted model implies that there is no association between the variables. Next, I transformed the sample data by substituting the parameter estimates from the table into Equation 5.17. After applying the transformation, I estimated the mean vector and the covariance matrix of the transformed data, and the right-most section of Table 5.4 shows the resulting estimates. As you can see, the transformed data have the same mean and covariance structure as the restricted model (within sampling error), so it is now appropriate to draw bootstrap samples from the data.

Having applied the Bollen–Stine transformation, I drew 2,000 bootstrap samples with replacement from the transformed data and subsequently fit the two regression models to each bootstrap sample. Next, I computed the likelihood ratio test from each bootstrap sample and constructed a frequency distribution of the test statistics. Figure 5.5 shows the empirical sampling distribution of the likelihood ratio statistic as well as the theoretical chi-square distribution with one degree of freedom. For illustration purposes, I also drew 2,000 naïve bootstrap samples from the untransformed data and computed the likelihood ratio test from each of those samples. I previously explained that the naïve bootstrap is inappropriate because it yields a distribution that reflects natural sampling fluctuation as well as model misfit. The fact that the naïve sampling distribution is shifted far to the right clearly illustrates this effect. In contrast, the Bollen–Stine sampling distribution is similar in shape to the theoretical chi-square distribution, but has a thicker tail. Leptokurtic data tend to inflate the likelihood ratio test, so the larger-than-expected proportion of statistics in the tail of the distribution makes intuitive sense.

Finally, I performed a significance test by referencing the likelihood ratio statistic to the empirical sampling distribution rather than to the theoretical chi-square distribution. Recall that the initial analysis produced a likelihood ratio statistic of LR = 123.05. The bootstrap sampling distribution did not include any values that were this large, so it is impossible to

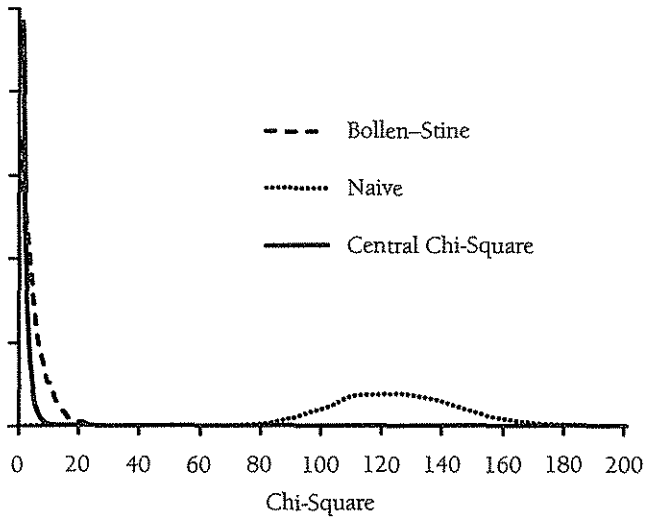

FIGURE 5.5. The theoretical central chi-square distribution and empirical sampling distributions of the likelihood ratio test generated from the Bollen–Stine bootstrap and naïve bootstrap. The naïve sampling distribution is centered on an inappropriately large chi-square value because it reflects natural sampling fluctuation as well as model misfit. In contrast, the Bollen–Stine sampling distribution is unaffected by model misfit and reflects the sampling fluctuation of the likelihood ratio test with nonnormal data. The Bollen–Stine distribution has a slightly thicker tail than the theoretical chi-square distribution, which is a result of the leptokurtic data.

compute an exact probability. However, the 99th percentile of the distribution corresponds to a test statistic of 17.97, so it is conservative to describe the probability value as $p < .01$.

Applying the Bollen–Stine Bootstrap to Missing Data

The matrix computations in Equation 5.17 require complete data, but Enders (2002) proposed the following transformation for missing data:

$$Z_i = (Y_i - \hat{\boldsymbol{\mu}}_{Si})^T(\hat{\boldsymbol{\Sigma}}_{Si})^{-1/2}(\hat{\boldsymbol{\Sigma}}_{Ri})^{1/2} + \hat{\boldsymbol{\mu}}_{Ri}^T \tag{5.18}$$

The missing data transformation is nearly identical to that of the complete data, but the parameter matrices now have an *i* subscript. The basic idea behind Equation 5.18 is to apply the transformation using only those estimates for which a case has complete data. Consequently, the *i* subscript denotes the fact that the size and the contents of the matrices can vary across individuals.

To illustrate the missing data transformation, reconsider the previous simple regression example. Furthermore, suppose that a subset of cases has missing *Y* values. Transforming the data effectively requires a unique transformation formula for each missing data pattern. For example, the transformation for the complete cases is

$$\mathbf{Z}_i = \left(\begin{bmatrix} X_i \\ Y_i \end{bmatrix} - \begin{bmatrix} \hat{\mu}_{X_S} \\ \hat{\mu}_{Y_S} \end{bmatrix}\right)^T \begin{bmatrix} \hat{\sigma}^2_{X_S} & \hat{\sigma}_{XY_S} \\ \hat{\sigma}_{YX_S} & \hat{\sigma}^2_{Y_S} \end{bmatrix}^{-1/2} \begin{bmatrix} \hat{\sigma}^2_{X_R} & \hat{\sigma}_{XY_R} \\ \hat{\sigma}_{YX_R} & \hat{\sigma}^2_{Y_R} \end{bmatrix}^{1/2} + \begin{bmatrix} \hat{\mu}_{X_R} \\ \hat{\mu}_{Y_R} \end{bmatrix}^T$$

For the cases with missing Y values, the transformation eliminates the parameters that correspond to Y and uses only those estimates that correspond to the complete variable, X. This transformation is as follows:

$$Z_i = (X_i - \hat{\mu}_{X_S})^T[\hat{\sigma}^2_{X_S}]^{-1/2}[\hat{\sigma}^2_{X_R}]^{1/2} + [\hat{\mu}_{X_R}]^T$$

Applying the transformation to each case's observed data yields a transformed data matrix that has the same missing data patterns as the sample data. At that point, the bootstrap procedure follows the same steps as before: (1) draw a large number of samples with replacement from the transformed data, (2) use maximum likelihood missing data handling to estimate the full model and the restricted model, (3) compute the likelihood ratio statistic for each bootstrap sample, and (4) generate a probability value by computing the proportion of bootstrap test statistics that exceed the value of the likelihood ratio test from the original analysis.

Some structural equation modeling programs (e.g., Mplus and EQS) implement the missing data bootstrap, and it is also relatively straightforward to implement the bootstrap in software packages that do not have built-in routines (Enders, 2005). The limited research to date suggests that the bootstrap yields relatively accurate type I error rates (Enders, 2001, 2002). For example, computer simulation studies report type I error rates between .05 and .07, even with kurtosis values as high as $K = 20$. The bootstrap appears to yield conservative error rates in small samples (e.g., type I error rates of 1% with $N = 100$), which suggests that the likelihood ratio test may lack power. However, studies have yet to examine this issue.

5.12 DATA ANALYSIS EXAMPLE 1

Recall from Chapter 4 that the EM algorithm does not require the computation of first or second derivatives, so many software packages do not report standard errors from an EM analysis. The first analysis example illustrates how to use bootstrap resampling to generate standard errors for EM estimates of a mean vector, covariance matrix, and a correlation matrix.* The data for this analysis are comprised of scores from 480 employees on eight work-related variables: gender, age, job tenure, IQ, psychological well-being, job satisfaction, job performance, and turnover intentions. I generated these data to mimic the correlation structure of published research articles in the management and psychology literature (e.g., Wright & Bonett, 2007; Wright, Cropanzano, & Bonett, 2007). The data have three missing data patterns, each of which is consists of one-third of the sample. The first pattern comprises 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).

I previously reported the EM parameter estimates in Table 4.12. I implemented the bootstrap procedure by drawing 2,000 samples of $N = 480$ with replacement from the job performance data set. Next, I performed an EM analysis on each bootstrap sample and saved the parameter estimates to a data file for further analysis. Finally, I generated standard errors

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

TABLE 5.5. EM Covariance Matrix and Bootstrap Standard Errors from Data Analysis Example 1

Variable	1	2	3	4	5	6	7	8
				EM covariance matrix				
1: Age	28.908							
2: Tenure	8.459	9.735						
3: Female	–0.028	–0.052	0.248					
4: Well-being	1.148	0.569	0.067	1.382				
5: Satisfaction	0.861	0.565	0.028	0.446	1.386			
6: Performance	–0.330	0.061	–0.009	0.671	0.271	1.570		
7: Turnover	–0.377	0.016	0.001	–0.141	–0.129	–0.203	0.218	
8: IQ	0.674	0.026	0.284	2.876	4.074	4.496	–0.706	70.892
				Bootstrap standard errors				
1: Age	1.828							
2: Tenure	0.823	0.630						
3: Female	0.121	0.070	0.002					
4: Well-being	0.320	0.192	0.032	0.108				
5: Satisfaction	0.307	0.198	0.032	0.092	0.099			
6: Performance	0.302	0.170	0.028	0.084	0.079	0.100		
7: Turnover	0.113	0.068	0.011	0.029	0.028	0.026	0.008	
8: IQ	2.166	1.278	0.195	0.565	0.597	0.534	0.177	5.255

Note. Elements in **bold** typeface are statistically significant at the .05 level because the estimated divided by its standard error exceeds ± 1.96.

by computing the standard deviation of each parameter estimate across the 2,000 samples. This process sounds tedious and time consuming, but it took less than two minutes on a laptop computer. Table 5.5 shows the EM covariance matrix estimates from Chapter 4 along with the bootstrap standard errors. I computed test statistics for each covariance term by dividing the parameter estimate by its bootstrap standard error. If the empirical sampling distributions are relatively normal, it is reasonable to compare these tests to a critical z value from the unit normal table. The table denotes the statistically significant estimates (i.e., the estimates for which the z test exceeded the two-tailed critical value of ± 1.96) in **bold** typeface.

5.13 DATA ANALYSIS EXAMPLE 2

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

$$JP_i = \hat{\beta}_0 + \hat{\beta}_1(WB_i) + \hat{\beta}_2(SAT_i) + \varepsilon$$

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

TABLE 5.6. Regression Model Estimates from Data Analysis Example 2

Parameter	Estimate	*SE*	*z*
	Maximum likelihood (auxiliary variables)		
β_0 (Intercept)	6.020	0.053	114.642
β_1 (Well-being)	0.475	0.054	8.798
β_2 (Satisfaction)	0.035	0.058	0.605
$\hat{\sigma}^2_\varepsilon$ (Residual)	1.241	0.086	14.369
R^2	.210		
	Maximum likelihood (no auxiliary variables)		
β_0 (Intercept)	6.021	0.053	113.123
β_1 (Well-being)	0.476	0.055	8.664
β_2 (Satisfaction)	0.027	0.060	0.445
$\hat{\sigma}^2_\varepsilon$ (Residual)	1.243	0.087	14.356
R^2	.208		

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

I used Graham's (2003) saturated correlates approach to incorporate IQ and turnover intentions (a binary variable) as auxiliary variables in the regression model. The path diagram for this analysis is identical to that in Figure 5.2. I chose IQ and turnover intentions as auxiliary variables because they had the strongest correlations with the regression model variables (correlations ranged between 0.30 and 0.40). In practice, stronger correlations would be better, but these were the strongest associations in the data. Because the data are MCAR, the auxiliary variables should have minimal impact on the parameter estimates, but they can provide a slight increase in power.

Researchers typically begin a regression analysis by examining the omnibus *F* test. In Chapter 4, I illustrated how to use the likelihood ratio statistic to perform a test of the regression coefficients. Because the addition of auxiliary variables does not affect this procedure, there is no need to illustrate the test here. Table 5.6 gives the regression model parameter estimates along with those from a maximum likelihood analysis with no auxiliary variables. I omit the estimates from the auxiliary variable portion of the model (e.g., the correlations between the auxiliary variables and the predictors) because they are not of substantive interest. As seen in the table, the parameter estimates from the saturated correlates model are virtually identical to those from Chapter 4. The analysis results suggest that psychological well-being was a significant predictor of job performance, $\hat{\beta}_1 = 0.475$, $z = 8.798$, $p < .001$, but job satisfaction was not, $\hat{\beta}_2 = 0.035$, $z = 0.605$, $p = .55$. Although the effect is subtle, the estimates from the saturated correlates model have slightly smaller standard errors and slightly larger *z* statistics, indicating that the auxiliary variables produced a slight increase in power. Auxiliary variables with stronger correlations would have produced more noticeable gains.

5.14 DATA ANALYSIS EXAMPLE 3

The final example is a confirmatory factor analysis that illustrates the use of auxiliary variables and the corrective procedures for nonnormal data.* 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 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. I generated the EAT questionnaire items to have discrete 7-point scales with positive skewness and kurtosis (e.g., skewness values typically ranged between 0.50 and 1.00, and kurtosis values of 1.00 were the norm). Inasmuch as the methodological literature has established that nonnormal data can bias standard errors and distort the likelihood ratio test, one of the goals of this analysis example is to illustrate how to correct for these problems.

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

This analysis used the same two-factor model as the example in Chapter 4, but it included three auxiliary variables: body mass index, anxiety, and beliefs about Western standards of beauty. Figure 5.6 shows a path diagram of the model. Notice that the saturated correlates model includes all possible correlations among the auxiliary variables as well as all possible correlations between the auxiliary variables and the manifest variable residuals (the auxiliary variables do not correlate with the latent factors). The two-factor model fit the data reasonably well according to conventional standards (Hu & Bentler, 1998, 1999), $\chi^2(34) = 49.044$, $p = .046$, RMSEA = 0.033, SRMR = 0.026. It is worth reiterating that the auxiliary variables do not affect the degrees of freedom for the model fit statistic because the auxiliary variable portion of the model includes all possible associations between the auxiliary variables and the manifest indicators (i.e., the auxiliary variable portion of the model is saturated).

One consequence of non-normal data is that the likelihood ratio test no longer follows the appropriate central chi-square distribution. I used the rescaling procedure and the

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

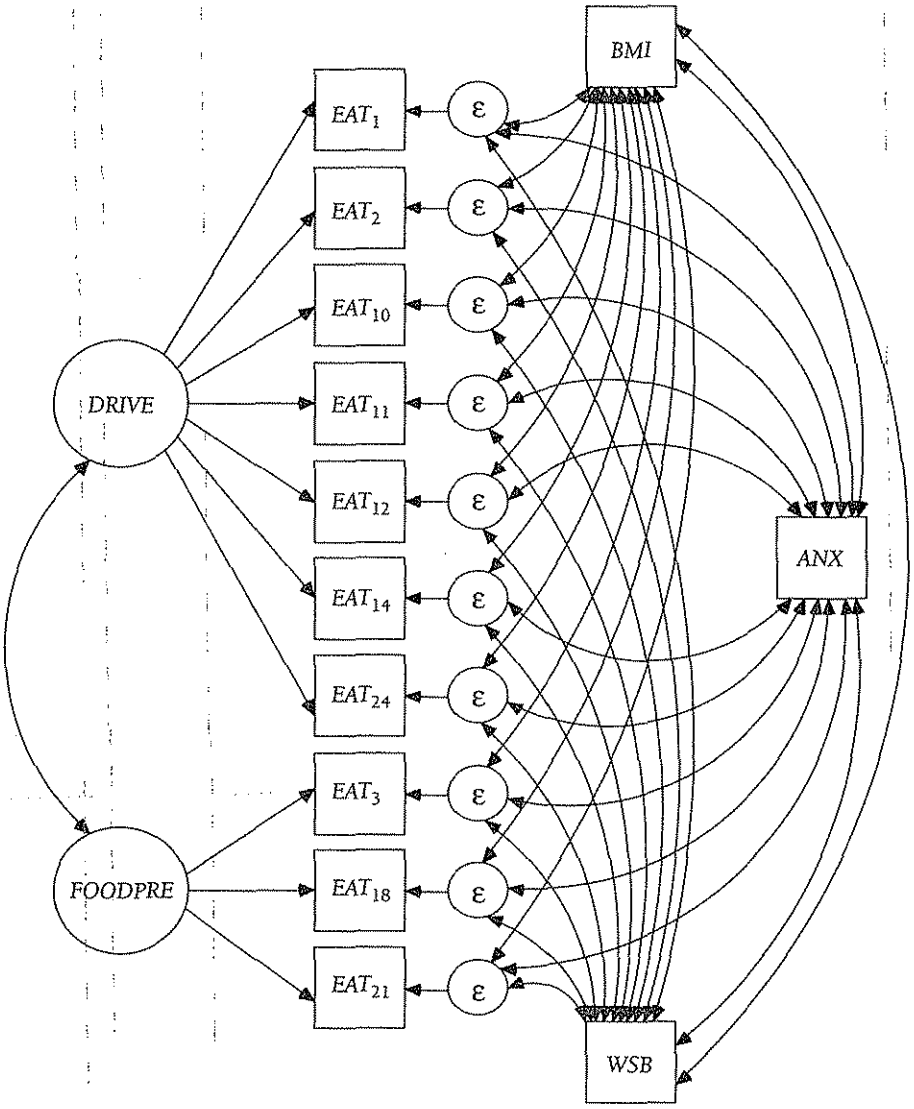

FIGURE 5.6. The Eating Attitudes Test CFA model with three auxiliary variables. Note that *DRIVE* = drive for thinness; *FOODPRE* = food preoccupation; *BMI* = body mass index; *ANX* = anxiety; *WSB* = Western standards of beauty. The saturated correlates model specifies correlations among the auxiliary variables and correlations between the auxiliary variables and the residual terms of the manifest indicators.

Bollen–Stine bootstrap to correct for bias in the likelihood ratio test (i.e., the chi-square test of model fit). In practice, there is no reason to implement both procedures, but I do so here for illustration purposes. To begin, the rescaled likelihood ratio test requires a scaling factor that pools information from the factor analysis model (i.e., the restricted model) and a saturated model that specifies all possible associations among the manifest variables (i.e., the full model). The two-factor model with 70 parameters produced a log-likelihood value and a scaling factor of $\log L_{\text{Restricted}} = -6957.674$ and $c_{\text{Restricted}} = 1.084$, respectively. The corresponding values from the saturated model with 104 parameters were $\log L_{\text{Full}} = -6933.152$ and $c_{\text{Full}} = 1.088$, respectively. Substituting these values into Equation 5.15 yields a scaling factor of $c_{\text{LR}} = 1.095$. Finally, dividing the likelihood ratio test by its scaling factor gives the rescaled test statistic, $\text{LR}_{\text{RS}} = 44.796$, and referencing this test to a central chi-square distribution with

34 degrees of freedom produces a probability value of $p = .102$. Notice that the rescaled test statistic is somewhat smaller than the original likelihood ratio test (LR = 49.044, $p = .046$) and is no longer statistically significant. Because the data are slightly leptokurtic, the downward adjustment in the test statistic makes intuitive sense.

The rescaling procedure transforms the likelihood test to more closely approximate the appropriate theoretical chi-square distribution. In contrast, the Bollen–Stine bootstrap leaves the likelihood ratio statistic intact and constructs a new empirical reference distribution. The null hypothesis for the likelihood ratio test states that the restricted model (i.e., the factor model) is true in the population, so the Bollen–Stine procedure transforms the sample data to have a perfect fit to the two-factor model. This transformation requires the mean vector and the covariance matrix from the sample data as well as the model-implied mean vector and covariance matrix from the factor analysis. These quantities are standard output in structural equation modeling packages. After applying the Bollen–Stine transformation, I drew 2,000 samples of $N = 400$ (the original sample size) with replacement from the transformed data set. Next, I used maximum likelihood missing data handling to estimate both models and saved the likelihood ratio statistic from each bootstrap sample.

Figure 5.7 shows the empirical sampling distribution of the likelihood ratio test along with the theoretical sampling distribution of a chi-square statistic with 34 degrees of freedom. For comparison purposes, the figure also shows the empirical sampling distribution for 2,000 naïve bootstrap samples. I previously explained that the naïve bootstrap is inappropriate because it yields a distribution that reflects natural sampling fluctuation as well as model misfit. The fact that the naïve sampling distribution is centered on an inappropriately large chi-square value illustrates this point. In contrast, the Bollen–Stine sampling distribution is unaffected by model misfit and reflects the sampling fluctuation of the likelihood ratio test

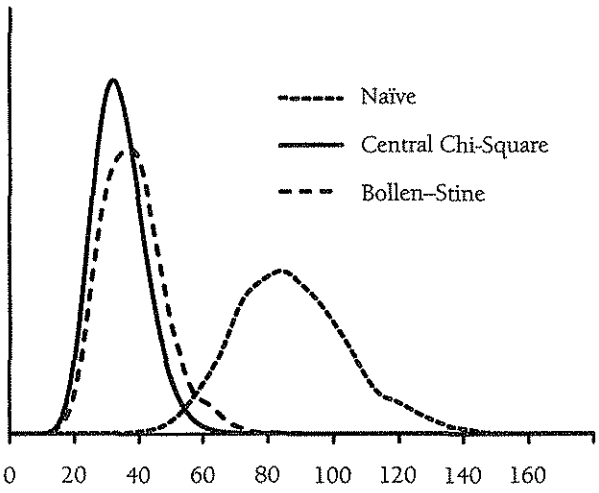

FIGURE 5.7. The theoretical central chi-square distribution and empirical sampling distributions of the likelihood ratio test generated from the Bollen–Stine bootstrap and naïve bootstrap. The naïve sampling distribution is centered on an inappropriately large chi-square value because it reflects natural sampling fluctuation as well as model misfit. In contrast, the Bollen–Stine sampling distribution is unaffected by model misfit and reflects the sampling fluctuation of the likelihood ratio test with nonnormal data. The Bollen–Stine distribution has a slightly thicker tail than the theoretical chi-square distribution, which is a result of the leptokurtic data.

with nonnormal data. The Bollen–Stine distribution has a slightly thicker tail than the theoretical chi-square distribution, which makes sense given that the data are slightly leptokurtic.

Next, I obtained an adjusted probability value for the likelihood ratio test by computing the proportion of test statistics from the Bollen–Stine sampling distribution that were equal to or greater than the value of the original likelihood ratio statistic. The Bollen–Stine probability value is $p = .208$, which means that 416 of the 2,000 bootstrap samples produced a test statistic that was equal to or greater than LR = 49.044. Visually, this probability value is the area under the Bollen–Stine sampling distribution that falls to the right of 49.044. You can see that this area is larger than that of the central chi-square distribution, which explains why the probability value increased from $p = .046$ to $p = .208$.

Incorporating auxiliary variables into a structural equation model yields invalid incremental (i.e., comparative) fit indices because the auxiliary variables inappropriately penalize the fit of the independence model. Earlier in the chapter, I illustrated how to remedy this problem by estimating a special independence model. The correct independence model for this analysis estimates the variances of the manifest variables, constrains the covariances among the manifest variables to zero, estimates the covariances between the auxiliary variables and the manifest variables, and estimates the covariances among the auxiliary variables. Estimating the modified independence model yields a likelihood ratio test of $LR_I = 1956.345$ with $df_I = 45$. Substituting these values into Equation 5.1 gives corrected CFI value of 0.993, as follows:

$$CFI = \frac{(1956.345 - 45) - (49.044 - 34)}{(1956.345 - 45)} = 0.993$$

The new CFI value is virtually identical to the incorrect value from the original analysis, but this will not always be the case. Although I only illustrate the correction to the CFI, the same corrective procedure applies to other incremental fit indices (e.g., the TLI, NFI).

Turning to the parameter estimates, note that Table 5.7 gives the factor loadings from the saturated correlates model along with those from a corresponding analysis with no auxiliary variables. I constrained the factor variances to unity in order to identify the model, so that the loadings reflect the expected change in the EAT item for a one-standard-deviation increase in the latent factor. Unlike the factor analysis from Chapter 4, this analysis satisfies the MAR assumption because the "causes" of missing data (i.e., body mass index and anxiety) appear as auxiliary variables in the model. Interestingly, the two sets of factor loadings in Table 5.7 are quite similar, so the addition of the auxiliary variables did little to change these estimates. Although not shown in the table, the auxiliary variable model did produce measurement intercepts (i.e., item means) that more closely resembled those of the complete data. The body mass index and anxiety variables have relatively modest correlations with the EAT questionnaire items. The impact of these auxiliary variables would therefore have been more pronounced had these correlations been stronger in magnitude.

The standard errors are of particular interest because non-normal data can distort these values. Table 5.7 shows the normal-theory standard errors along with those of the sandwich estimator and the naïve bootstrap. As seen in the table, the normal-theory standard errors were generally lower than the robust standard errors and the bootstrap standard errors. Lep-

TABLE 5.7. Confirmatory Factor Analysis Loading Estimates from Data Analysis Example 3

	No auxiliary		Saturated correlates model			
EAT Item	Estimate	*SE*	Estimate	*SE*	SE_R	SE_{BS}
EAT_1	0.741	0.050	0.741	0.050	0.049	0.048
EAT_2	0.650	0.045	0.649	0.045	0.050	0.050
EAT_{10}	0.807	0.043	0.808	0.043	0.052	0.053
EAT_{11}	0.764	0.040	0.764	0.040	0.049	0.049
EAT_{12}	0.662	0.047	0.662	0.047	0.055	0.056
EAT_{14}	0.901	0.041	0.901	0.041	0.047	0.048
EAT_{24}	0.623	0.048	0.622	0.048	0.053	0.053
EAT_3	0.772	0.046	0.772	0.046	0.052	0.052
EAT_{18}	0.749	0.048	0.751	0.048	0.056	0.056
EAT_{21}	0.862	0.045	0.862	0.046	0.053	0.053

Note. SE_R = robust standard errors; SE_{BS} = bootstrap standard errors.

tokurtic data tend to attenuate standard errors, so these differences make sense. The limited research to date suggests that significance tests based on robust standard errors may be slightly conservative (i.e., standard errors are a bit too large), whereas bootstrap significance tests may be slightly liberal (i.e., standard errors are a bit too small). However, the difference between the two procedures is often trivial, and both approaches provide a rather dramatic improvement over standard errors that assume normality (Enders, 2001).

5.15 SUMMARY

This chapter describes techniques that can improve the accuracy of maximum likelihood missing data handling. The first half of the chapter is devoted to the use of auxiliary variables. The definition of MAR states that the probability of missing data on a variable *Y* can relate to some other measured variable (or variables) but not to the values of *Y* itself. Although this definition seems to be satisfied when a correlate of missingness is a variable in the data set, the variables in the analysis dictate the missing data mechanism. For this reason, methodologists recommend an inclusive analysis strategy that incorporates a number of auxiliary variables. An auxiliary variable is one that is ancillary to the substantive research questions but is a potential correlate of missingness or a correlate of the missing variable. Including auxiliary variables in a maximum likelihood analysis can reduce or eliminate bias (e.g., by making the MAR assumption more plausible) and can increase in power (e.g., by recapturing some of the lost information in the missing variable).

Two strategies can be used for implementing an inclusive analysis strategy: the saturated correlates model and the two-stage analysis procedure. The saturated correlates model uses structural equation modeling software to incorporate auxiliary variables as correlates of the analysis variables. A set of rules guide the specification of the model, and the basic idea behind these rules is to transmit the information from the auxiliary variables to the analysis model without affecting the interpretation of the parameters. The two-stage approach is an

alternative to the saturated correlates model that deals with missing data in two steps: the first stage uses maximum likelihood missing data handling to estimate the mean vector and covariance matrix, and the second stage uses the resulting estimates as input data for subsequent analyses. The advantage of the two-stage approach is that it can readily incorporate any number of auxiliary variables into the first step of the procedure, so there is no need to include the auxiliary variables in the subsequent analysis step. The problem with the two-stage approach is that it requires complex standard error computations that have not yet been implemented in computer software programs. However, given the ease with which the two-stage approach incorporates auxiliary variables, it will likely become a viable alternative to the saturated correlates model in the near future.

The second half of the chapter is devoted to corrective procedures for nonnormal data. Maximum likelihood estimation relies heavily on the multivariate normality assumption, both for identifying the most likely parameter values and for computing standard errors. The methodological literature shows that non-normal data tend to have a minimal impact on the parameter estimates themselves but can bias standard errors and distort the likelihood ratio test. This chapter outlined two strategies for correcting the bias in standard errors, the so-called sandwich estimator (i.e., "robust" standard errors) and bootstrap resampling. With non-normal data, the second derivative formulas from the normal distribution are missing terms that depend on the skewness and kurtosis of the population distribution. The absence of these terms can overestimate or underestimate the standard errors, depending on whether the data are leptokurtic or platykurtic. Robust standard errors correct this problem by using a "sandwich" of terms that involve first and second derivatives. In contrast, bootstrap resampling uses Monte Carlo simulations to generate standard errors. The basic idea behind the bootstrap is to treat the sample data as a miniature population and draw repeated samples of size N from the sample data set. Estimating the analysis model from the bootstrap samples yields an empirical sampling distribution for each parameter, the standard deviation of which estimates the standard error.

Non-normal data can also distort the likelihood ratio test. The basic problem with the test is that its sampling distribution no longer follows the appropriate central chi-square distribution. One solution is to rescale the test statistic so that it more closely approximates its theoretical chi-square distribution. This rescaling procedure divides the normal-theory likelihood ratio test by a correction factor that depends on the kurtosis of the data. The scaling factor can increase or decrease the value of the likelihood ratio test, depending on the distribution shape of the data. The Bollen–Stine bootstrap is a second method that can correct the bias in the likelihood ratio test. Whereas the rescaling procedure attempts to adjust the value of the likelihood ratio test so that it more closely approximates its theoretical sampling distribution, the bootstrap leaves the test statistic intact and uses Monte Carlo simulation to generate a new sampling distribution. The bootstrap corrects the probability value for the test by referencing the likelihood ratio statistic to the empirical sampling distribution rather than to the theoretical chi-square distribution. Although the Bollen–Stine bootstrap follows the same procedure as the naïve bootstrap, it applies an algebraic transformation to the sample data prior to drawing the samples. This transformation aligns the mean and covariance structure of the data to the null hypothesis and produces an empirical distribution that reflects only the sampling fluctuation of the likelihood ratio test.

The next chapter takes a break from missing data issues and provides an introduction to Bayesian estimation. Chapters 7 through 9 focus on a second "modern" missing data technique, multiple imputation. The mathematical machinery behind multiple imputation is heavily entrenched in Bayesian methodology. At one level, you can effectively implement multiple imputation in your own research without fully understanding its Bayesian underpinnings. However, understanding multiple imputation at a deeper level requires a background in Bayesian statistics; accessing the seminal missing data work can be difficult without this knowledge. The purpose of Chapter 6 is to provide a user-friendly introduction to Bayesian statistics, while still providing a level of detail that will serve as a springboard for accessing the technically oriented missing data literature.

5.16 RECOMMENDED READINGS

Bollen, K. A. & Stine, R. A. (1992). Bootstrapping goodness-of-fit measures in structural equation models. *Sociological Methods and Research*, *21*, 205–229.

Collins, L. M., Schafer, J. L, & Kam, C-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*, 330–351.

Finney, S. J., & DiStefano, C. (2006). Nonnormal and categorical data in structural equation models. In G. R. Hancock & R. O. Mueller (Eds.), *A second course in structural equation modeling* (pp. 269–314). Greenwich, CT: Information Age.

Freedman, D. A. (2006). On the so-called "Huber sandwich estimator" and "robust standard errors." *The American Statistician*, *60*, 299–302.

Graham, J. W. (2003). Adding missing-data relevant variables to FIML-based structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*, 80–100.

Yuan, K-H., Bentler, P. M., & Zhang, W. (2005). The effect of skewness and kurtosis on mean and covariance structure analysis: The univariate case and its multivariate implication. *Sociological Methods and Research*, *24*, 240–258.