Biases in SPSS 12.0 Missing Value Analysis

Paul T. von Hippel

In addition to SPSS Base software, SPSS Inc. sells a number of add-on packages, including a package called Missing Value Analysis (MVA). In version 12.0, MVA offers four general methods for analyzing data with missing values. Unfortunately, none of these methods is wholly satisfactory when values are missing at random. The first two methods, listwise and pairwise deletion, are well known to be biased. The third method, regression imputation, uses a regression model to impute missing values, but the regression parameters are biased because they are derived using pairwise deletion. The final method, expectation maximization (EM), produces asymptotically unbiased estimates, but EM's implementation in MVA is limited to point estimates (without standard errors) of means, variances, and covariances. MVA can also impute values using the EM algorithm, but values are imputed without residual variation, so analyses that use the imputed values can be biased.

KEY WORDS: EM algorithm; Ignorable missingness; Imputation; Maximum likelihood; Missing at random; Missing data; Missing values; Multiple imputation; Nonignorable missingness.

1. INTRODUCTION

In addition to SPSS Base software, SPSS Inc. sells a variety of add-on packages, including SPSS Complex Samples, SPSS Exact Tests, and SPSS Advanced Models. In this review, we focus on the add-on package called SPSS Missing Value Analysis (MVA). MVA has received only passing mention in reviews of missing-value packages (e.g., Horton and Lipsitz 2001), probably because it does not support the increasingly popular techniques of multiple imputation. In an environment where SPSS is used heavily, however, MVA might be tempting as a second-best solution.

Unfortunately, the methods implemented in MVA are not second best. In this review, we discuss MVA's methods and demonstrate their potential biases and limitations.

2. MISSING VALUE MECHANISMS

When a dataset has missing values, the difficulty of obtaining valid parameter estimates depends on the mechanism that causes values to be missing. A useful classification for missing-value mechanisms was introduced by Rubin (1976); summaries appear in Little (1992) and Schafer (1997). Here we review the essentials.

Informally, let *missingness* be the probability that a value is missing rather than observed. If missingness depends on both observed and missing values, then values are *not missing at random* (NMAR). If missingness depends only on observed values (and not on missing values), then values are *missing at random* (MAR). And if missingness depends neither on observed nor on missing values, then values are *missing completely at random* (MCAR). The analyst can ignore the missing data mechanism provided that values are MAR or MCAR. If values are NMAR, however, then the mechanism cannot be ignored.

More formally, let **Z** be a data matrix representing n cases of k random variables whose joint distribution depends on the parameter vector $\boldsymbol{\theta}$. **Z** is only partially observed—that is, **Z** is made up of observed values $\mathbf{Z}_{\text{obs}}$ and missing values $\mathbf{Z}_{\text{mis}}$. The locations of the missing values are summarized by a matrix **R** of the same dimension as **Z**. **R** contains dummy variables that indicate whether each value in **Z** is observed or missing—that is, $R_{ij} = 1$ if Z_{ij} is missing, and $R_{ij} = 0$ if Z_{ij} is observed.

In general, the distribution of **R** is written

$$p(\mathbf{R}|\mathbf{Z}_{\text{mis}}, \mathbf{Z}_{\text{obs}}, \boldsymbol{\phi}) \tag{1a}$$

to show that missingness can depend on both missing and observed **Z** values, and on a parameter vector $\boldsymbol{\phi}$, which we assume to be distinct from $\boldsymbol{\theta}$.

Values are NMAR if missingness depends on both observed and missing values—that is, if the distribution in (1a) cannot be written in simpler form.

Values are MAR if missingness depends on observed values but not on missing values—that is, if the distribution of **R** can be simplified to depend on $\mathbf{Z}_{\text{obs}}$ and $\boldsymbol{\phi}$ but not on $\mathbf{Z}_{\text{mis}}$:

$$p(\mathbf{R}|\mathbf{Z}_{\text{mis}}, \mathbf{Z}_{\text{obs}}, \boldsymbol{\phi}) = p(\mathbf{R}|\mathbf{Z}_{\text{obs}}, \boldsymbol{\phi}). \tag{1b}$$

Finally, values are MCAR if missingness depends neither on observed nor on missing values—that is, if the distribution of **R** can be simplified to depend on $\boldsymbol{\phi}$ but not on $\mathbf{Z}_{\text{obs}}$ or $\mathbf{Z}_{\text{mis}}$:

$$p(\mathbf{R}|\mathbf{Z}_{\text{mis}}, \mathbf{Z}_{\text{obs}}, \boldsymbol{\phi}) = p(\mathbf{R}|\boldsymbol{\phi}). \tag{1c}$$

Typically the analyst's goal is to estimate the parameters $\boldsymbol{\theta}$ that govern the distribution of **Z**. If values are MCAR or MAR, then the missing data mechanism is said to be *ignorable*, and valid estimates can be obtained without an explicit model of the missing data mechanism. If values are NMAR, then the missing value mechanism is *nonignorable* and a model of it must be incorporated into the estimation process. This can be quite difficult in practice, because the missing data mechanism is rarely known with much certainty.

2.1 Example

An example may help to illustrate patterns of missing data and the methods used to analyze them. Let X and Y be standard

Paul T. von Hippel is Statistician, Department of Sociology and Initiative in Population Research, Ohio State University, 300 Bricker Hall, 190 N. Oval Mall, Columbus, OH 43210 (E-mail: von-hippel.1@osu.edu). Data were simulated using the %mvn macro under SAS 8, and calculations were carried out using Mathematica 5. SPSS MVA was purchased using a Small Grant from the Ohio State University College of Social and Behavioral Sciences to Douglas B. Downey. I thank Colin Odden for technical support.

DOI: 10.1198/0003130043204

normal variables with a correlation of $\rho = 0.5$. Then the first two moments of X and Y are as follows:

$$\begin{aligned} \mu &= E\left(\begin{bmatrix} X \\ Y \end{bmatrix}\right) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\ \Sigma &= \text{var cov}\left(\begin{bmatrix} X \\ Y \end{bmatrix}\right) = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}. \end{aligned} \quad (2)$$

These moments are not hard to estimate if the data are complete, but suppose that X is missing half of its values. Echoing a more complicated example (Allison 2000; Horton and Lipsitz 2001), we consider three different mechanisms that may govern the missingness of X:

MCAR: X is missing with probability 0.5;
MAR: X is missing if $Y < 0$;
NMAR: X is missing if $X < 0$.

If values are MCAR, estimates obtained from SPSS MVA software will be unbiased. But this is not impressive, because the observed values are a random sample from the population. If values are NMAR, estimates from MVA may be biased. But this bias is not surprising, because MVA has no provisions for modeling the missing data mechanism.

In the ensuing discussion, we therefore focus on the MAR pattern. When values are MAR, it is in principle possible to obtain unbiased estimates without using a model of the missing data mechanism. Nonetheless, most estimates obtained from MVA are biased. We demonstrate MVA's biases by closed-form calculations, and by applying MVA to a large simulated dataset that follows the MAR pattern above. The simulated data consist of 50,000 observations from the joint distribution of X and Y, where the value of X is missing whenever $Y < 0$.

3. ESTIMATION WITH MISSING VALUES

MVA can produce a tabular summary of the missing value patterns in a dataset. In our simulated sample of 50,000 cases, MVA's summary shows us that 25,079 cases have values for both X and Y, while the remaining 24,921 cases have values for Y but not for X.

MVA has also implemented Little's (1988) test of the hypothesis that values are MCAR. For our sample this hypothesis is convincingly rejected ($\chi^2(1) = 31{,}809, p \ll 0.0005$), which is reassuring because values are MAR by design. (MVA rounds the p value to three digits, reporting it as 0.000.)

Although the features above may be appealing, most analysts will want to estimate parameters when values may be missing at random. MVA offers four methods for this purpose. The first two methods are based on deleting cases. The last two are based on imputation and likelihood.

3.1 Deletion Methods

The first two methods are *listwise deletion* and *pairwise deletion*, also known as complete-case and available-case analysis. These methods are not really selling points for MVA, because they are also implemented in SPSS Base software. In fact, the implementation in SPSS Base software is more comprehensive, since it provides listwise and pairwise estimates for a variety of models including regression and factor analysis. MVA's implementation, by contrast, can only give point estimates for means, variances, and covariances.

We review MVA's deletion methods below, partly for the sake of completeness, and partly because they lead naturally into MVA's imputation and likelihood methods.

3.1.1 Listwise Deletion

In listwise deletion (LD), all cases with missing values are deleted. Following deletion, conventional methods are used to derive estimates from the remaining, complete cases. LD estimates for the first two moments are represented by $\hat{\mu}_{\text{LD}}$ and $\hat{\Sigma}_{\text{LD}}$.

Unfortunately, LD can produce biased estimates when values are MAR. In our MAR example, LD means deleting cases where X is missing since $Y < 0$. But when we derive estimates from the remaining cases, we are not estimating the unconditional moments μ and Σ. Instead, we are estimating the moments conditionally on $Y \geq 0$. It can be shown (Rose and Smith 2002, p. 226) that these conditional moments are

$$\begin{aligned} E(\hat{\mu}_{\text{LD}}) &= E\left(\begin{bmatrix} X \\ Y \end{bmatrix} \mid Y \geq 0\right) = \sqrt{\frac{2}{\pi}}\begin{bmatrix} \rho \\ 1 \end{bmatrix} \approx \begin{bmatrix} 0.399 \\ 0.798 \end{bmatrix} \\ E(\hat{\Sigma}_{\text{LD}}) &= \text{var cov}\left(\begin{bmatrix} X \\ Y \end{bmatrix} \mid Y \geq 0\right) \\ &= \frac{1}{\pi}\begin{bmatrix} \pi - 2\rho^2 & (\pi-2)\rho \\ (\pi-2)\rho & \pi - 2 \end{bmatrix} \approx \begin{bmatrix} 0.841 & 0.182 \\ 0.182 & 0.363 \end{bmatrix}, \end{aligned} \quad (3)$$

which differ substantially from the unconditional moments μ and Σ in (2). Table 1 shows that the LD estimates from our simulated sample are quite close to their expectations $E(\hat{\mu}_{\text{LD}})$ and $E(\hat{\Sigma}_{\text{LD}})$.

3.1.2 Pairwise Deletion

In pairwise deletion (PD), each moment is estimated separately using cases with values for the pertinent variables. In our MAR example, $E(Y)$ and var(Y) would be estimated using all the cases, but $E(X)$, var(X), and cov(X, Y) would be estimated using only the cases with values for X. PD estimates for the first two moments are represented by $\hat{\mu}_{\text{PD}}$ and $\hat{\Sigma}_{\text{PD}}$.

Like LD estimates, PD estimates can be biased when values are MAR. In our MAR example, the PD estimates of $E(Y)$ and var(Y) are unconditional, but the estimates of $E(X)$, var(X), and cov(X, Y) are conditional on $Y \geq 0$:

$$\begin{aligned} E(\hat{\mu}_{\text{PD}}) &= E\left(\begin{bmatrix} X \mid Y \geq 0 \\ Y \end{bmatrix}\right) = \begin{bmatrix} \rho\sqrt{2/\pi} \\ 0 \end{bmatrix} \approx \begin{bmatrix} 0.399 \\ 0 \end{bmatrix} \\ E(\hat{\Sigma}_{\text{PD}}) &= \begin{bmatrix} \text{var}(X \mid Y \geq 0) & \text{cov}(X, Y \mid Y \geq 0) \\ \text{cov}(X, Y \mid Y \geq 0) & \text{var}(Y) \end{bmatrix} \\ &= \frac{1}{\pi}\begin{bmatrix} \pi - 2\rho^2 & (\pi-2)\rho \\ (\pi-2)\rho & \pi \end{bmatrix} \approx \begin{bmatrix} 0.841 & 0.182 \\ 0.182 & 1 \end{bmatrix}. \end{aligned} \quad (4)$$

The PD estimates of $E(X)$, var(X), and cov(X, Y) are biased, with expectations quite different from the population values in (2). Table 1 shows that the PD estimates from our simulated sample are quite close to their expectations $E(\hat{\mu}_{\text{PD}})$ and $E(\hat{\Sigma}_{\text{PD}})$.

In addition to being biased, PD can yield "impossible" covariance matrices that fail to be nonnegative definite. For example,

Table 1. Means and Covariances for Bivariate Normal Variables X and Y. Simulated sample of 50,000 cases, with X missing whenever Y<0.

	Population values			Listwise deletion			Pairwise deletion			Regression imputation			EM estimates			EM imputation		
		Covariances			Covariances			Covariances			Covariances			Covariances			Covariances	
	Means	X	Y	Means	X	Y	Means	X	Y	Means	X	Y	Means	X	Y	Means	X	Y
X	0	1		0.397	0.841		0.397	0.841		0.325	0.837		−0.006	1.005		−0.006	0.633	
Y	0	0.5	1	0.798	0.185	0.363	0.005	0.185	0.996	0.005	0.185	0.996	0.005	0.507	0.996	0.005	0.507	0.996

Arbuckle (1996) constructed a small dataset where the PD estimate for a correlation is an impossible -1.45. The absolute correlation exceeds 1 because the absolute covariance exceeds the product of the standard deviations. Such a configuration is impossible when all moments are estimated from the same cases, but quite possible in PD, where different cases may be used to estimate the covariance and each standard deviation.

MVA actually implements two methods for PD estimation. In the method described above, the mean and variance of Y are estimated using all cases with values for Y. In the alternative method, the mean and variance of Y would be estimated using only cases where Y can be paired with an observed value of X. In our simulated data, these alternative PD estimates are identical to those obtained by LD. The alternative PD estimates play no role in subsequent calculations, and will not be discussed further.

3.2 Imputation and Likelihood-Based Methods

As mentioned earlier, deletion methods are not a reason to buy MVA, because deletion methods are available in SPSS Base software. The selling points for MVA are its methods based on imputation and likelihood. Although imputation and likelihood-based methods have the potential to produce valid point estimates and standard errors, their implementation in MVA is disappointing. When values are MAR, only one of MVA's method (EM) produces valid results, and MVA's implementation of that method is limited to point estimates of means, variances, and covariances.

3.2.1 Regression Imputation

In *regression imputation* (RI), a regression model is used to impute (or fill in) all missing values. Following imputation, conventional methods are used to derive estimates from the complete and imputed cases together. RI estimates for the first two moments are $\hat{\mu}_{\mathrm{RI}}$ and $\hat{\Sigma}_{\mathrm{RI}}$.

RI poses some practical difficulties. One problem is that there may be many possible regressors. To address this problem, MVA can reduce the number of regressors using forward stepwise selection. A second difficulty is that the regressors may themselves have missing values. To address this problem, MVA estimates several different regression equations—one for each pattern of missing values—so that each missing value can be imputed using only the regressors that are observed for that case.

The difficulties are more modest in our MAR example, where X is the only variable with missing values, and Y is the only possible regressor. Applied to our simulated data, MVA imputes the missing values of X using the following model:

$$X_{\mathrm{imp}} = \hat{\alpha}_{\mathrm{PD}} + \hat{\beta}_{\mathrm{PD}} Y + e, \quad \text{where} \quad e \sim N(0, \hat{\sigma}^2_{\mathrm{PD}}). \quad (5)$$

Here the model is shown with normal residuals. MVA can also impute values using t residuals, resampled residuals, or no residuals at all. MVA can use the imputed data to estimate means, variances, and covariances, or it can save the imputed data for further analysis.

Regardless of the options that the user chooses, MVA's RI method has a fundamental flaw. The regression estimates $\hat{\alpha}_{\mathrm{PD}}$, $\hat{\beta}_{\mathrm{PD}}$, and $\hat{\sigma}^2_{\mathrm{PD}}$ are biased, because they are derived from the biased moment estimates $\hat{\mu}_{\mathrm{PD}}$ and $\hat{\Sigma}_{\mathrm{PD}}$ (SPSS Inc. 2002). Substituting $\hat{\mu}_{\mathrm{PD}}$ and $\hat{\Sigma}_{\mathrm{PD}}$ into the usual formulas for regression estimates, we can show that $\hat{\alpha}_{\mathrm{PD}}$, $\hat{\beta}_{\mathrm{PD}}$, and $\hat{\sigma}^2_{\mathrm{PD}}$ converge in probability to the following values

$$\begin{aligned}
\operatorname{plim}\left(\hat{\alpha}_{\mathrm{PD}}\right) &= \rho\sqrt{2/\pi} - \hat{\beta}_{\mathrm{PD}} 0 \approx 0.399 \\
\operatorname{plim}\left(\hat{\beta}_{\mathrm{PD}}\right) &= \frac{(\pi-2)\rho}{\pi} \approx 0.182 \\
\operatorname{plim}\left(\hat{\sigma}^2_{\mathrm{PD}}\right) &= (\pi - 2\rho^2 - \hat{\beta}^2_{\mathrm{PD}}\pi)/\pi \approx 0.808, \qquad (6)
\end{aligned}$$

which are rather different from the true regression parameters $\alpha = 0$, $\beta = 0.5$, and $\sigma^2 = 0.5$.

Because the PD regression estimates are biased, the RI estimates $\hat{\mu}_{\mathrm{RI}}$ and $\hat{\Sigma}_{\mathrm{RI}}$ are biased as well. To see this, split the regression-imputed data $\mathbf{Z}_{\mathrm{RI}}$ into two parts, $\mathbf{Z}_{\mathrm{comp}}$ and $\mathbf{Z}_{\mathrm{imp}}$. $\mathbf{Z}_{\mathrm{comp}}$ consists of complete cases—cases where both X and Y are observed because $Y \geq 0$. $\mathbf{Z}_{\mathrm{imp}}$ consists of imputed cases—cases where X must be imputed because $Y < 0$. The moments of $\mathbf{Z}_{\mathrm{comp}}$ are simply the moments of (X, Y) conditional on $Y \geq 0$:

$$\begin{aligned}
\mu_{\mathrm{comp}} &= E\left(\begin{bmatrix} X \\ Y \end{bmatrix} \mid Y \geq 0\right) = \begin{bmatrix} \rho\sqrt{2/\pi} \\ \sqrt{2/\pi} \end{bmatrix} \\
\Sigma_{\mathrm{comp}} &= \operatorname{var}\operatorname{cov}\left(\begin{bmatrix} X \\ Y \end{bmatrix} \mid Y \geq 0\right) \\
&= \frac{1}{\pi}\begin{bmatrix} \pi - 2\rho^2 & (\pi-2)\rho \\ (\pi-2)\rho & \pi - 2 \end{bmatrix}. \qquad (7)
\end{aligned}$$

We have already presented these, in (3), as the moments estimated by listwise deletion. The moments of $\mathbf{Z}_{\mathrm{imp}}$ are the moments of (X_{imp}, Y) conditional on $Y < 0$. If values are imputed using model (5), the moments of $\mathbf{Z}_{\mathrm{imp}}$ are

$$\begin{aligned}
\mu_{\mathrm{imp}} &= E\left(\begin{bmatrix} X_{\mathrm{imp}} \\ Y \end{bmatrix} \mid Y < 0\right) \\
&= E\left(\begin{bmatrix} \hat{\alpha}_{\mathrm{PD}} + \hat{\beta}_{\mathrm{PD}} Y + e \\ Y \end{bmatrix} \mid Y < 0\right) \\
&= \begin{bmatrix} \hat{\alpha}_{\mathrm{PD}} - \hat{\beta}_{\mathrm{PD}}\sqrt{2/\pi} \\ -\sqrt{2/\pi} \end{bmatrix} \\
\Sigma_{\mathrm{imp}} &= \operatorname{var}\operatorname{cov}\left(\begin{bmatrix} X_{\mathrm{imp}} \\ Y \end{bmatrix} \mid Y < 0\right) \\
&= \operatorname{var}\operatorname{cov}\left(\begin{bmatrix} \hat{\alpha}_{\mathrm{PD}} + \hat{\beta}_{\mathrm{PD}} Y + e \\ Y \end{bmatrix} \mid Y < 0\right)
\end{aligned}$$

$$= \frac{1}{\pi}\left[\begin{array}{cc} \hat{\beta}^2_{\mathrm{PD}}(\pi-2)+\hat{\sigma}^2\pi & \hat{\beta}_{\mathrm{PD}}(\pi-2) \\ \hat{\beta}_{\mathrm{PD}}(\pi-2) & \pi-2 \end{array}\right]. \quad (8)$$

Now the moments of $\mathbf{Z}_{\mathrm{RI}}$ can be obtained by combining the moments of $\mathbf{Z}_{\mathrm{comp}}$ and $\mathbf{Z}_{\mathrm{imp}}$:

$$\begin{aligned} \mu_{\mathrm{RI}} &= \frac{1}{2}\left(\mu_{\mathrm{comp}}+\mu_{\mathrm{imp}}\right) \\ \Sigma_{\mathrm{RI}} &= \frac{1}{2}\left(\Sigma_{\mathrm{comp}}+\Sigma_{\mathrm{imp}}\right) \\ &\quad +\frac{1}{2}\left(\mu_{\mathrm{comp}}-\mu_{\mathrm{imp}}\right)\frac{1}{2}\left(\mu_{\mathrm{comp}}-\mu_{\mathrm{imp}}\right)^T. \quad (9) \end{aligned}$$

These moments μ_{RI} and Σ_{RI} are the expectations of the RI estimators $\hat{\mu}_{\mathrm{RI}}$ and $\hat{\Sigma}_{\mathrm{RI}}$. If we plug in the probability limits for $\hat{\alpha}_{\mathrm{PD}}$, $\hat{\beta}_{\mathrm{PD}}$, and $\hat{\sigma}^2_{\mathrm{PD}}$, which we obtained in (6), we find that $\hat{\mu}_{\mathrm{RI}}$ and $\hat{\Sigma}_{\mathrm{RI}}$ converge in probability to the following limits:

$$\begin{aligned} \mathrm{plim}\left(\hat{\mu}_{\mathrm{RI}}\right) &\approx \left[\begin{array}{c} 0.326 \\ 0 \end{array}\right] \\ \mathrm{plim}\left(\hat{\Sigma}_{\mathrm{RI}}\right) &\approx \left[\begin{array}{cc} 0.836 & 0.182 \\ 0.182 & 1 \end{array}\right]. \quad (10) \end{aligned}$$

These limits are close to the PD expectations $E(\hat{\mu}_{\mathrm{PD}})$ and $E(\hat{\Sigma}_{\mathrm{PD}})$ in (4), and far from the population parameters μ and Σ in (2). Table 1 shows that the RI estimates from our simulated sample are quite close to their probability limits plim$(\hat{\mu}_{\mathrm{RI}})$ and plim$(\hat{\Sigma}_{\mathrm{RI}})$.

3.2.2 *EM Algorithm*

MVA's final method is the *expectation-maximization* (EM) algorithm (Dempster, Laird, and Rubin 1977). EM is based on iterating the process of regression imputation. In our MAR example, EM's first step uses initial regression estimates $\hat{\alpha}_0, \hat{\beta}_0$ to impute X from Y:

$$X_{\mathrm{imp}} = \hat{\alpha}_0 + \hat{\beta}_0 Y. \quad (11)$$

Any convenient values will do for $\hat{\alpha}_0, \hat{\beta}_0$; MVA's implementation uses the PD estimates $\hat{\alpha}_{\mathrm{PD}}, \hat{\beta}_{\mathrm{PD}}$ (SPSS Inc. 2002). Note that the imputations include no residual variation (SPSS Inc. 2002). Imputed residuals would add noise to the algorithm, introducing random variation that is not inherent in the data.

Using the complete and imputed cases together, EM reestimates the means, variances, and covariances, using a formula that compensates for the lack of residual variation in the imputed values of X (SPSS Inc. 2002). The newly estimated moments $\hat{\mu}_1, \hat{\Sigma}_1$ imply new estimates $\hat{\alpha}_1, \hat{\beta}_1$ of the regression parameters. These new regression estimates are used to generate new imputations of X, and the process iterates until convergence. (For MVA's convergence criterion, see SPSS Inc. 2002.) By default MVA sets a limit of 25 iterations, but generates an error message if that limit is not sufficient. In our MAR example, convergence was achieved after 56 iterations.

The moments estimated in the final iteration are the EM estimates $\hat{\mu}_{\mathrm{EM}}, \hat{\Sigma}_{\mathrm{EM}}$. Dempster et al. (1977) showed that EM estimates are maximum likelihood estimates. Because maximum likelihood estimates are consistent, the EM estimates converge in probability to the population parameters:

$$\begin{aligned} \mathrm{plim}\left(\hat{\mu}_{\mathrm{EM}}\right) &= \mu = \left[\begin{array}{c} 0 \\ 0 \end{array}\right] \\ \mathrm{plim}\left(\hat{\Sigma}_{\mathrm{EM}}\right) &= \Sigma = \left[\begin{array}{cc} 1 & \rho \\ \rho & 1 \end{array}\right] = \left[\begin{array}{cc} 1 & 0.5 \\ 0.5 & 1 \end{array}\right]. \quad (12) \end{aligned}$$

Table 1 shows that the EM estimates from our simulated sample are quite close to their probability limits μ and Σ.

3.2.3 *EM Imputation*

Although asymptotically unbiased, the EM estimates in Table 1 are of limited use. In most analyses, researchers are not interested in means, variances, and covariances, but in the parameters of some causal or descriptive model. Sometimes these parameters can be estimated by transforming the EM means, variances, and covariances. More often, though, it would be convenient to analyze the EM-imputed data in SPSS Base software or elsewhere.

MVA can indeed save the EM-imputed data for further analysis. In our MAR example, missing X values would be imputed using the following model,

$$X_{\mathrm{imp}} = \hat{\alpha}_{\mathrm{EM}} + \hat{\beta}_{\mathrm{EM}} Y, \quad (13)$$

where $\hat{\alpha}_{\mathrm{EM}}$ and $\hat{\beta}_{\mathrm{EM}}$ are the regression estimates obtained in the final iteration of the EM algorithm. Because the EM algorithm converges to maximum likelihood estimates, $\hat{\alpha}_{\mathrm{EM}}$ and $\hat{\beta}_{\mathrm{EM}}$ are consistent estimators of the true regression parameters α and β.

Nevertheless, analyses that use the EM-imputed data can be biased. The reason is that values imputed using model (13) lack residual variation. The lack of imputed residuals means that analyses using the imputed values will be biased by insufficient variation in X_{imp}. For example, let $\hat{\mu}_{\mathrm{EMI}}$ and $\hat{\Sigma}_{\mathrm{EMI}}$ be the estimated moments of the EM-imputed data. Calculations similar to those in Section 3.2.1 show that the EM-imputed estimates converge in probability to the following limits:

$$\begin{aligned} \mathrm{plim}\left(\hat{\mu}_{\mathrm{EMI}}\right) &= \left[\begin{array}{c} 0 \\ 0 \end{array}\right] \\ \mathrm{plim}\left(\hat{\Sigma}_{\mathrm{EMI}}\right) &= \left[\begin{array}{cc} (1+\rho^2)/2 & \rho \\ \rho & 1 \end{array}\right] = \left[\begin{array}{cc} 0.625 & 0.5 \\ 0.5 & 1 \end{array}\right]. \quad (14) \end{aligned}$$

It is clear that the variance of X is underestimated unless X is perfectly correlated with Y—that is, unless $|\rho| = 1$. If $|\rho|$ is 1, then the lack of imputed residuals is unimportant since the residual variance is 0. As ρ gets closer to 0, however, the underestimation of var(X) gets more and more serious. In our MAR example, $\rho = 0.5$, and var(X) is underestimated by 37.5%. Table 1 shows that the EMI estimates from our simulated sample are quite close to their probability limits plim$(\hat{\mu}_{\mathrm{EMI}})$ and plim$(\hat{\Sigma}_{\mathrm{EMI}})$.

Note that residuals are not just omitted from the final iteration of the EM algorithm. Earlier iterations omit residuals as well. As noted earlier, the EM algorithm compensates for this omission by using specialized formulas. Since such formulas are not built into SPSS Base software (or other packages), it is inadvisable to use the EM-imputed data outside the EM module.

3.3 Standard Errors

In our discussion, we have emphasized MVA's point estimates of means, variances, and covariances. We have shown that, except for the EM algorithm, MVA's point estimates can be biased when values are missing at random.

Point estimates, of course, are only part of the story. Usually the analyst would also like to have good estimates of standard errors, along with concomitant p values and confidence intervals. When values are MAR or MCAR, valid standard errors can be obtained using either maximum likelihood or multiple imputation techniques (Little and Rubin 1987; Rubin 1987). These techniques, however, are not implemented in MVA. Even MVA's implementation of the EM algorithm does not estimate standard errors, though EM methods for estimating standard errors are well known (McLachlan and Krishnan 1997).

4. CONCLUSION

When normally distributed values are missing at random, a variety of software packages can produce asymptotically unbiased estimates of distributional parameters. AMOS and other software produce such estimates using maximum likelihood (Allison 2002; Arbuckle 1996), and a growing number of packages can do so using multiple imputation (Allison 2002; Horton and Lipsitz 2001).

Unfortunately, most of the methods in SPSS Missing Value Analysis fall short of this standard. When bivariate normal values are missing at random, MVA's listwise and pairwise deletion methods can produce biased estimates. MVA's regression imputation method can be biased as well, since it uses a regression model whose parameters are obtained by pairwise deletion.

The one bright spot in MVA is its implementation of the EM method, which can produce maximum likelihood point estimates of means, variances, and covariances. The EM method can also be used to impute missing values. Unfortunately, MVA imputes these values without residual variation. Analyses based on the EM imputations can therefore be biased.

MVA does not estimate standard errors, and does not support the likelihood or multiple-imputation methods that can produce valid standard error estimates.

Communication from SPSS technical support suggests that the company has been aware of MVA's problems for more than three years (e.g., Nichols 2000). We assume that SPSS is not aware of how serious the biases can be, and we hope that the present review helps draw attention to the issue.

[Received December 2003. Revised February 2004.]

REFERENCES

Allison, P. D. (2000), "Multiple Imputation for Missing Data: A Cautionary Tale," *Sociological Methods and Research*, 28, 301–309.

——— (2002), *Missing Data*, Thousand Oaks, CA: Sage.

Arbuckle, J. L. (1996), "Full Information Estimation in the Presence of Incomplete Data," in *Advanced Structural Equation Modeling: Issues and Techniques*, eds. George A. Marcoulides and Randall E. Schumacker, Mahwah, NJ: Lawrence Erlbaum Associates, pp. 243–277.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data Via the EM Algorithm," *Journal of the Royal Statistical Society*, Series B, 39, 1–38.

Horton, N. J., and Lipsitz, S. R. (2001), "Multiple Imputation in Practice," *The American Statistician*, 55, 244–254.

Little, R. J. A. (1988), "A Test of Missing Completely at Random for Multivariate Data With Missing Values," *Journal of the American Statistical Association*, 83, 1198–1202.

——— (1992), "Regression With Missing X's: A Review," *Journal of the American Statistical Association*, 87, 1227–1237.

Little, R. J. A., and Rubin, D. B. (1987), *Statistical Analysis With Missing Data*, New York: Wiley.

McLachlan, G. J., and Krishnan, T. (1997), *The EM Algorithm and Extensions*, New York: Wiley.

Nichols, D. (2000), E-mail message posted to the Structural Equation Modeling Discussion Network (SEMNET), accessed March 15, 2004; available at http://bama.ua.edu/cgi-bin/wa?A2=ind0008&L=semnet&P=R5358&I=1.

Rose, C., and Smith, M. D. (2002), *Mathematical Statistics With Mathematica*, New York: Springer Verlag.

Rubin, D. B. (1976), "Inference and Missing Data," *Biometrika*, 63, 581–592.

——— (1987), *Multiple Imputation for Survey Nonresponse*, New York: Wiley.

Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*, Boca Raton, FL: Chapman and Hall.

SPSS Inc. (2002), "MVA [specification of algorithms]," Accessed April 1, 2004; available at http://support.spss.com/tech/stat/Algorithms/12.0/mva.pdf.