Bayesian Multiple Imputation and Maximum Likelihood Methods for Missing Data

Min Sun and Ferry Butar Butar
Sam Houston State University
Huntsville, TX 77341

Abstract

Bayesian multiple imputation and maximum likelihood provide useful strategy for dealing with dataset including missing values. Imputation methods affect the significance of test results and the quality of estimates. In this paper, the general procedures of multiple imputation and maximum likelihood described which include the normal-based analysis of a multiple imputed dataset. A Monte Carlo simulation is conducted to compare the performances of the methods.

Key words: multiple imputation, Bayesian method, maximum likelihood, Monte Carlo, bootstrap

1. Introduction

In sample surveys, the problem of nonresponse is very important and is difficult to handle. Little and Rubin (1987) gave an explication of various missing data patterns, which has been helpfully abstracted by Roth (1994) and Schafer and Graham (2002). According to Rubin (1976), there are even more possible missing data patterns when considering two or more variables at once.

1.1. Types of missing data

Schafer and Graham (2002) described that missing data can informally be thought of as being caused in some combination of three ways: random processes, processes which are measured, and processes which are not measured. Modern missing data methods generally work well for the first two causes, but not for the last. More formally, missing data mechanisms are commonly described as falling into one of three categories, described by Little and Rubin (1987).

The first type of missing data can be “Missing completely at Random”, or MCAR. Suppose there are missing data on a particular variable *Y*. The data on Y are said to be missing completely at random (MCAR) if the probability of missing data on *Y* is unrelated to the value of *Y* itself or to the values of any other variables in the data set. When this assumption is satisfied for all variables, the set of individuals with complete data can be regarded as a simple random subsample from the original set of observations. Note that MCAR does allow for the possibility that "missingness" on *Y* is related to "missingness" on some other variable *X*.

Second, data can be “Missing at Random”, or MAR. In this case, missing data depends on known values and thus is described fully by variables observed in the data set. Accounting for the values which “cause” the missing data will produce unbiased results in an analysis. Supposing there are only two variables *X* and *Y*, where *X* always is observed and *Y* sometimes is missing. MAR can be expressed as

$$P\ (Y_{\text{missing}}\ |Y;\ X\ ) = P\ (\text{Y}_{\text{missing}}\ |\ X).$$

It means that given both *Y* and *X*, the conditional probability of missing data on *Y* is equal to the probability of missing data on *Y* given *X* alone. In general, data are not missing at random if those individuals with missing data on a particular variable tend to have lower (or higher) values on that variable than those with data present, controlling for other observed variables.

The third type of missing data can be missing in an unmeasured fashion, termed “nonignorable” (also called “Missing Not at Random” (MNAR) and “Not Missing at Random” (NMAR)). The missing data mechanism is said to be ignorable if (a) the data are MAR and (b) the parameters that govern the missing data process are unrelated to the parameters to be estimated. Ignorability basically means that there is no need to model the missing data mechanism as part of the estimation process.

If the data are not MAR, we say that the missing data mechanism is nonignorable. In that case, usually the missing data mechanism must be modeled to get good estimates of the parameters of interest. One widely used method for nonignorable missing data is Heckman's (1976) two-stage estimator for regression models with selection bias on the dependent variable. Unfortunately, for effective estimation with nonignorable missing data, very good prior knowledge about the nature of the missing data process usually is needed, because the data contain no information about what models would be appropriate and the results typically will be very sensitive to the choice of model. Since the missing data depends on events or items which the researcher has not measured, this is a damaging situation described by Sinharay et al. (2001).

In a summary of the three types of missing data patterns, Schafer and Graham (2002) mentioned that MCAR, MAR and NMAR can be distinguished by delineating the antecedents of the missing data on variable *Y*. That is, the probability that data are missing on *Y* can depend on (a) neither *X* nor *Y* (MCAR), (b) *X* but not *Y* when *X* is controlled (MAR), or (c) *Y* itself (NMAR).

1.2 Methods to treat missing data

The intent of any analysis is to make valid inferences regarding a population of interest. Missing data threatens this goal if it is missing in a manner which makes the sample different than the population from which it was drawn, that is, if the missing data creates a biased sample. Therefore, it is important to respond to a missing data problem in a manner which reflects the population of inference.

It is important to understand that once data are missing, it is impossible not to treat them – once data are missing, any subsequent procedure with that data set represents a response in some form to the missing data problem. As a result, there are many different methods of managing missing data. The two most commonly used techniques are *complete-case analysis* and *available case analysis*.

The first is "complete-case analysis", also known as listwise deletion, and is accomplished by deleting from the sample any observations that have missing data on any variables in the model of interest and then applying conventional methods of analysis for complete data sets. The other technique is "available case analysis", also commonly known as pairwise deletion, which is a simple alternative that can be used for many linear models. The technique of single imputation is to substitute some reasonable imputation for each missing value and then proceed to do the analysis as if there were no missing data. There are lots of ways to impute missing values such as hot-deck imputation, mean substitution and regression imputation.

None of these methods adjusts for the fact that the imputation process involves uncertainty about the missing values. The better methods are maximum likelihood (ML) and multiple imputations (MI) (Allison, 2000). This paper tries to display and compare these two methods. In the next section, we describe maximum likelihood (ML) method. In section 3, Bayesian multiple imputation is considered. The comparisons of the two methods are showed in section 4 and the conclusion is given in section 5.

2. Maximum likelihood

Maximum likelihood (ML) is a very general approach to statistically estimate a set of parameters that maximize the probability of getting the data that was observed. It is particularly adept at handling missing data problems. Under MAR, the maximum likelihood can be obtained by summing the usual likelihood over all possible values of the missing data. Consider a two variable sample of *n* independent observations. Let $f(x, y \mid \theta)$ represent the likelihood, where θ is a set of unknown parameters that govern the distribution of *X* and *Y*. Assuming that *X* is discrete (When *X* is continuous, use an integral to replace the summation), the marginal distribution of *Y* provides the correct likelihood for θ,

$$g(y \mid \theta) = \sum_{x} f(x, y \mid \theta)$$

For the n observations, we obtain all values for *Y*, but only the first m observations for *X*,

The likelihood for the entire sample can be expressed as,

$$L(\theta) = \prod_{i=1}^{m} f(x_i, y_i \mid \theta) \prod_{i=m+1}^{n} g(y_i \mid \theta)$$

ML is to find values of θ to maximize this likelihood. There is a variety of methods to solve this optimization problem. A general method was described by Dempster et al. (1977) in their article on the expectation-maximization (EM) algorithm. EM consists of two steps, an expectation step and a maximization step. These two steps are repeated multiple times in an iterative process that eventually converges to the ML estimates.

3. Bayesian multiple imputation

3.1 Introduction

Single imputation treats the missing values as if they were known, thereby resulting in unreliable inferences, because the variability from not knowing the missing values is ignored. However, multiple imputations provide a useful strategy for dealing with data sets with missing values (Little & Rubin, 1987). Instead of filling in a single value for each missing value, Rubin's (1987) multiple imputations procedure replaces each missing value with a set of plausible values that represent the uncertainty about the right value to impute. Multiple imputations (MI) construct a distribution for the missing observations. A complete data set is then formed by using this distribution to impute values for all the missing observations. Then these multiply imputed data sets are analyzed using standard procedures just as if the imputed data were the real data obtained from the nonrespondents (Davey et al., 2001). This process is repeated several times, and in each repetition a new set of imputed values is chosen for the missing observations. This collection of complete-data inferences can be combined to form one inference. The inference reflects the uncertainty due to nonresponse more properly than that if just one set of imputed values is considered (Rubin, 1987; Rubin & Wang, 2000). This results in valid statistical inferences that properly reflect the uncertainty due to missing values.

The performance of several multiple imputation methods have been studied by Rubin and Schenker (1986), which include Single random (*SR*) imputation, Bayesian bootstrap (*BB*) imputation, Approximate Bayesian bootstrap (*ABB*) imputation, designed for discrete data, and Fully normal (*FN*) imputation, Imputation adjusted for uncertainty in the mean and variance (*MV*), designed for continuous data . The certain general result is that the *FN* and the *MV* methods are superior to all of the other methods.

3.2 Multiple imputation procedure

In multiple imputations, the missing values are drawn from an appropriate distribution that characterizes the conditional relation of the imputed variables to other variables. Of course, the imputed values for any subject are not real values and have no interpretation. They are utilized simply as statistical tools to effectively use other nonmissing variables from that subject to make an inference about the quantity of interest. The procedure of drawing missing values from the distribution is repeated *M* times. Because the missing values are drawn from a distribution, there will be a range of values imputed for each missing value, with this variation appropriately reflecting the uncertainty about that value. After imputation, each of the *m* completed data sets is analyzed separately, and the results are combined (Sinharay et al., 2001). The Bayesian theoretical underpinnings of the method require a statistical model for the joint probability distribution of all of the variables. On the basis of this model, each missing value is drawn from an appropriate distribution (Little & Rubin, 1987).

Let Q denote the quantity of interest, Y_o denote observed data and Y_m the missing data. If there were no missing values, let $\hat{Q}$ be an estimate of Q and the estimated variance is $\hat{V}(\hat{Q})$ and assume that $\hat{Q}$ -Q is normally distributed with mean 0 and variance $\hat{V}(Q)$. Now suppose that due to nonresponse, only n_1 of the total data values n are observed, that is,

Observed values, $Y_{oi,}$ $i= 1,\dots, n_1$

Missing values, Y_{ml}, $l = 1,\dots, n_0$

$$n = n_1+n_0$$

For each missing value give m imputations created under a single nonresponse model, there are m completed data sets yielding m values of $\hat{Q}$ and $\hat{V}(\hat{Q})$, say

$$(\hat{Q}_{*l}, \hat{V}(\hat{Q}_{*l})),\ l = 1, ..., m.$$

Then, $Q - \hat{Q}_{*l}$ is $N(0, T_*)$

We have the estimated value after inserting the imputed values,

$$\hat{\bar{Q}}_* = \frac{1}{m}\sum \hat{Q}_{*l}$$

and its variance formula is given by

$$T_* = \frac{1}{m}\sum_{l=1}^{m} SE_l{}^2 + \left(\frac{m+1}{m}\right)\frac{1}{m-1}\sum_{l=1}^{m}(\hat{Q}_{*l} - \hat{\bar{Q}}_*)^2 .$$

Note that the first term is the within-imputation variance and the second is between-imputation variance. The factor of $(m + 1) / m$ are an improvement for modest m because it reflects the extra variability of $\hat{\bar{Q}}_*$ based on a finite rather than an infinite number of imputations (Rubin & Schenker, 1986).

Interval inferences for Q is given by

$$I(Y_1) = \hat{\bar{Q}}_* \pm k T_*^{1/2}$$

When giving the confidence interval for population mean after multiple imputations, the quantities introduced above take the following special forms:

$$Q = \mu ,\ \hat{Q} = \bar{y} ,\ \hat{V}(\hat{Q}) = s^2/n$$

The following quantities denoting the values of $\bar{y}$ and s^2/n in the l th data set completed by multiple imputation, $l = 1, ..., m$.

$$\hat{Q}_{*l} = \bar{y}_{*l} \text{, and } \hat{V}(\hat{Q}_{*l}) = \frac{s_{*l}^2}{n}$$

Then, we have the estimated mean after inserting the imputation values

$$\hat{\bar{y}}_* = m^{-1}\sum_{l=1}^{m}\hat{\bar{y}}_{*l}$$

and estimated variance of $(\hat{\hat{y}}_* - \mu)$,

$$T_* = \frac{1}{m}\sum_{l=1}^{m} SE_l^{\ 2} + \left(\frac{m+1}{m}\right)\frac{1}{m-1}\sum_{l=1}^{m}(\hat{Q}_{*l} - \hat{\hat{Q}}_*)^2$$

$$= \frac{1}{m}\sum_{l=1}^{m}\frac{s_{*l}^2}{n} + \left(\frac{m+1}{m}\right)\frac{1}{m-1}\sum_{l=1}^{m}(\hat{\hat{y}}_{*l} - \hat{\hat{y}}_*)^2.$$

The confidence interval is given by

$$I(Y_1) = \hat{\hat{y}}_* \pm kT_*^{1/2} \ , \quad \text{where } k = z_{\alpha/2},$$

when m is small use $k = t_{v,\alpha/2}$, where v is given by

$$v = (m-1)\left[1 + \frac{m-1}{n(m+1)}\frac{\sum_{l=1}^{m} s_{*l}^2}{\sum_{l=1}^{m}(\hat{\hat{y}}_{*l} - \hat{\hat{y}}_*)^2}\right]^2$$

4. Comparisons

We simulated dataset consist of 1,000 students for six courses: Math, Science, Biology, Art, English and History.

Of 1000 observations, we generated 15%, 25%, 33.5% and 50% missing data at random. The methods of Maximum likelihood (ML) and Multiple imputations (MI) are used to estimate and to analyze the data.

Maximum likelihood estimations are provided by Amos, a software package designed for estimating structural equation models with latent variables. Amos assumes that data values that are missing are missing at random (MAR). It is not always easy to know whether this assumption is valid or what it means in practice (Rubin 1976). But if the MAR condition is satisfied, Amos provides maximum likelihood estimates efficiently and consistently even in the presence of missing data.

The SAS procedures, PROC MI and PROC MIANALYZE, are used to give Multiple imputations. The MI procedure is a multiple imputation procedure that creates multiply imputed data sets for incomplete p-dimensional multivariate data. Then it uses methods that incorporate appropriate variability across multiple imputations. Once the multiple complete data sets are analyzed using standard SAS/STAT procedures, PROC MIANALYZE is used to generate valid statistical inferences about these parameters by combining the results.

Table 1 and Figure 1 show the comparison of the estimated means for ML and MI and Table 2 and Figure 2 presents the comparison of their standard errors. From the results, several trends can be observed.

Variable	Mean								
		ML				MI			
	No Missing	15%	25%	33%	50%	15%	25%	33%	50%
Math	74.709	74.687	74.677	74.872	74.729	74.704	74.708	74.798	74.896
Science	74.816	74.835	74.813	74.907	75.109	74.826	74.802	74.848	75.185
Biology	75.096	75.184	74.951	74.87	75.107	75.080	75.077	75.216	75.027
Art	75.046	75.055	75.071	75.045	75.17	75.044	75.066	75.017	75.134
English	74.600	74.591	74.622	74.484	74.82	74.588	74.580	74.446	74.824
History	71.764	71.626	72.91	70.182	71.555	71.509	72.760	70.321	71.567

Table 1. Means of ML and MI estimations

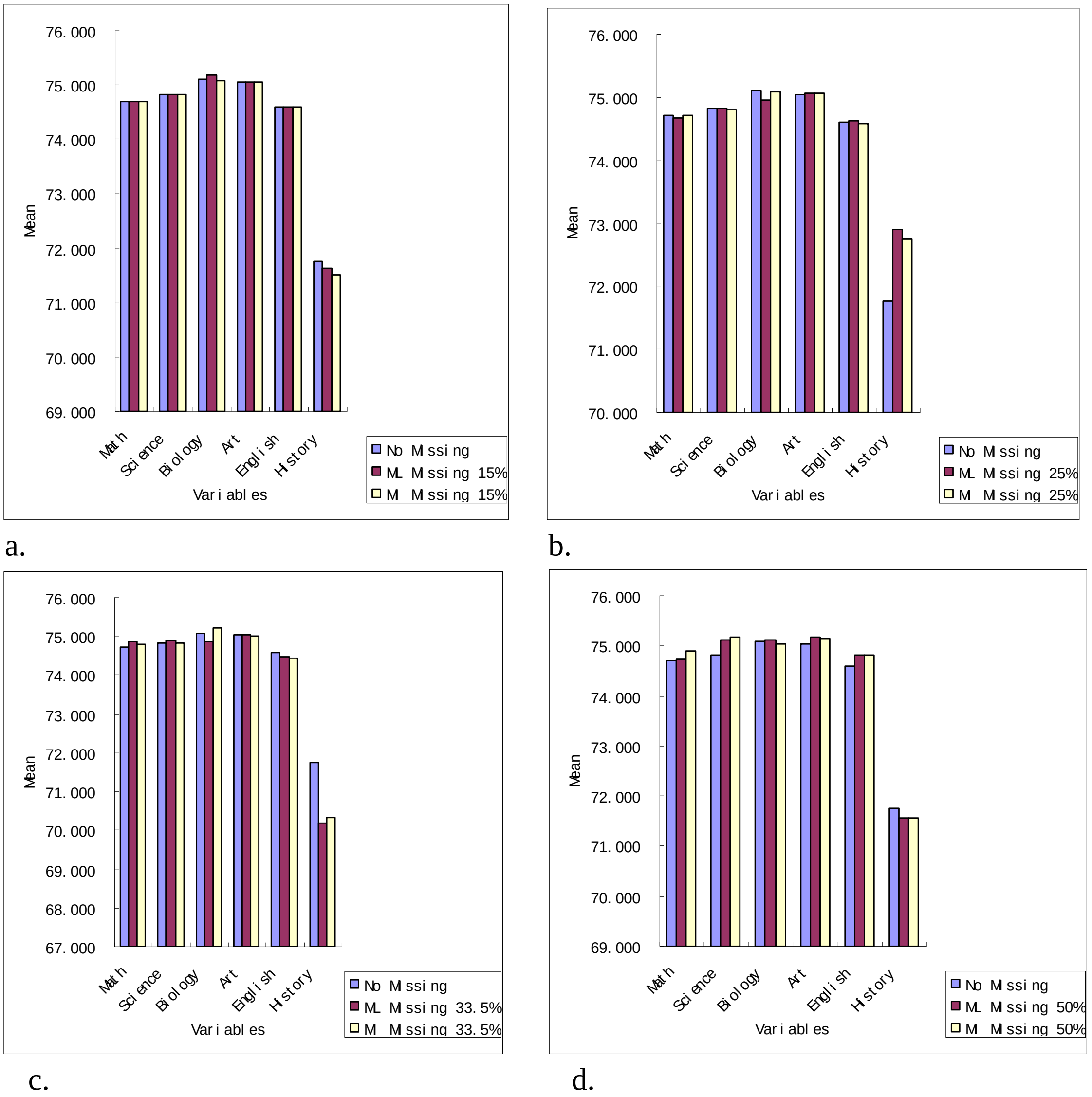
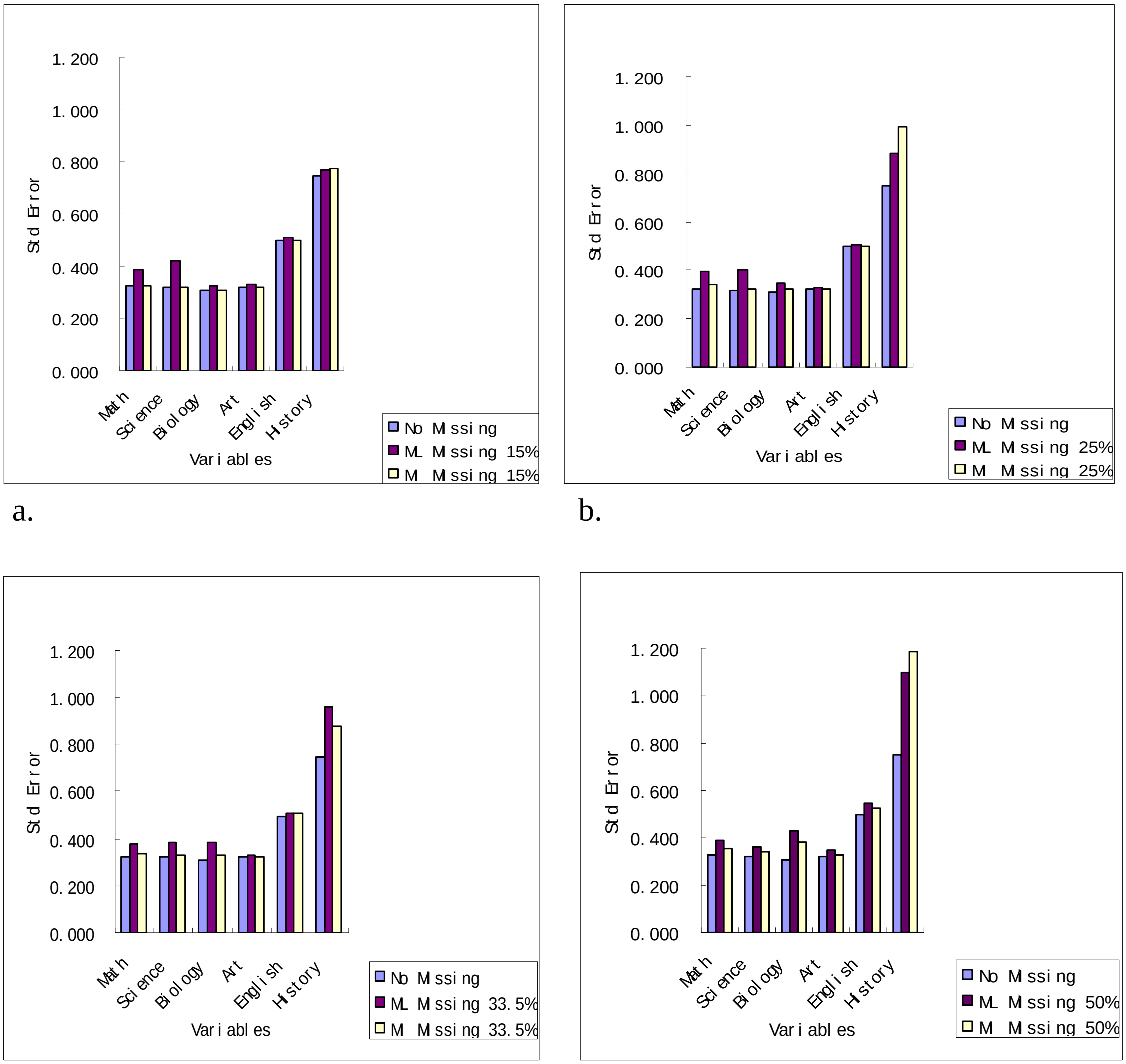


Figure 1. Means Comparison of ML and MI estimations.

Variable	Standard Error								
		ML				MI			
	No Missing	15%	25%	33%	50%	15%	25%	33%	50%
Math	0.324	0.389	0.398	0.379	0.389	0.324	0.341	0.334	0.355
Science	0.319	0.42	0.405	0.384	0.358	0.320	0.326	0.332	0.339
Biology	0.308	0.326	0.35	0.387	0.43	0.309	0.323	0.330	0.384
Art	0.321	0.329	0.327	0.326	0.346	0.322	0.323	0.325	0.329
English	0.497	0.509	0.505	0.509	0.544	0.501	0.502	0.508	0.524
History	0.748	0.769	0.881	0.961	1.095	0.772	0.991	0.880	1.188

Table 2. Standard errors of ML and MI estimations.



a.

b.

c.

d.

Figure 2. Standard errors comparison of ML and MI estimations.

1. For each of the percentages of missing data the means for both ML and MI produce similar results compare with the means of the completed data. For 15% and 50% missing data, both the ML and the MI methods produce similar estimations. However, for 25% and 33.5% data are missing, the MI's results are closer to the means of the completed data than that of the ML's estimations.

2. The MI's standard errors are almost the same with the standard errors of the completed data for 15% and 25% data are missing. They become quite different when more data are missed as shown in the figure 2.c. and 2.d except for the last variable.

3. The ML's standard errors are bigger than both standard errors of the completed data and the MI's estimations at all data sets except the last variables in the 25% and 50% missing data.

5. Conclusions

In summary, maximum likelihood (ML) and Bayesian multiple imputation (MI) are highly useful paradigms for handing missing values in many settings. Bayesian multiple imputation (MI) separates the imputation phase from the analysis phase which has some advantageous when the models are different under different conditions.

It has been noted that the different data missingness can influence the performance of ML and MI estimations. The difference of the ML and the MI methods will need future comparisons for the MCAR and the MNAR with different missing percentages.

References

Allison, P. D. (2000). Multiple imputations for missing data: A cautionary tale. *Sociological Methods and Research*, 28, 301-309.

Amos Software, *Version 5.0.1*. Amos Development Corporation. http://amosdevelopment.com

Davey, A., Shanahan, M. J., and Schafer, J. L. (2001). Correcting for selective nonresponse in the national longitudinal survey of youth using multiple imputation. *The Journal of Human Resources,* 36(3), 500-519.

Dempster, A. P., Laird, N. H., and Rubin, D. B. (1997). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*. B39, 1-38.

Heckman, J. J. (1976). The common structure of statistical models of the truncated, sample selection and limited dependent variables, and a simple estimator of such models. *Annals of Economic and Social Measurement*, 5, 475-492.

Little, R. J. A., and Rubin D. B. (1987). *Statistical analysis with missing data*. New York, Wiley.

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personal Psychology*. 17(3), 537-560.

Rubin, D. B. (1976). Inference and misiing data. *Biometrika*, 63, 581-592.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in Survey*. New York, Wiley.

Rubin, D. B., and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366-374.

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

Rubin, D. B., and Wang, N. (2000). Inference for imputation estimators. *Biometrika*, 87, 113-124.

SAS Software, Version 9.1. (2003). SAS Institute Inc., Cary, NC, USA.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London, Chapman & Hall.

Schafer, J. L., and Graham, J. W. (2002). Missing data: Our view of the State of the Art. *Psychological Methods*, Vol. 7, No. 2, 147-177.

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