Imputation Methods for Incomplete Dependent Variables in Finance

by

Paul Kofman
School of Finance and Economics
University of Technology, Sydney
Sydney, NSW 2007
Tel.+61.2.95147728
Fax.+61.2.95147711
paul.kofman@uts.edu.au

and

IAN SHARPE
School of Banking and Finance
University of New South Wales
Sydney, NSW 2052
Tel. +61.2.93855856
Fax.+61.2.93856347
i.sharpe@unsw.edu.au

Preliminary, January 2000[♦]

♦ The authors gratefully acknowledge support from the ARC small grant scheme. Excellent research assistance was provided by Karen Lew. Comments and suggestions by Guyonne Kalb and seminar participants at Monash University and the University of Technology, Sydney are much appreciated.

Imputation Methods for Incomplete Dependent Variables in Finance

ABSTRACT

Missing observations in dependent variables is a common feature of many financial applications. Standard ad hoc missing value imputation methods invariably fail to deliver efficient and unbiased parameter estimates. A number of recently developed classical and Bayesian iterative methods are evaluated for the treatment of missing dependent variables when the independent variables are completely observed. These methods are compared by simulation to commonly applied alternative missing data methodologies in the finance literature. The methods are then applied to a system of simultaneous equations modelling the maturity, secured status, and pricing of U.S. bank revolving loan contracts. Two of the four dependent variables in this application are characterised by severe missingness. The system of equations approach allows us to also exploit the additional information contained in the interdependencies among these features. The results indicate that proper treatment of missingness can be important for many financial applications.

Keywords: missing data, EM-algorithm, IP-algorithm, multiple imputations, revolving loan characteristics
JEL-code: G00, G21, C80

MISSING DATA PROBLEMS occur frequently in applied research in finance. While it is relatively easy to acknowledge the existence of such problems, it seems much harder to admit to any potential consequences for the investigator's research. These consequences can be manifold, but ultimately boil down to a questioning of the reliability of the research outcomes.

While some researchers have ignored (and not reported) the problem, others have used one of the following three approaches. The predominant approach – the *listwise deletion* method – is to exclude the observations with missing data from the study and only use complete records. Unfortunately, the damage done is an efficiency cost due to lost observations. A further difficulty with this approach is that authors sometimes find that the observations where data are missing appear to have characteristics more attuned to a particular outcome of the dependent variable. In these circumstances exclusion of the missing data introduces a systematic bias to the estimates. A second common approach – the *omitted variable* method – is to exclude the variables with missing data from the analysis. King et al. (1998) show that this method also risks bias though not inefficiency. Also, variable omission will not be an option when the variables with missing data are dependent variables in the analysis. The third alternative is to use some kind of imputation method.

The typical imputation approach – the *ad hoc value imputation* method – is to assume the blank fields take some ad hoc (subjective) value. This could take the form of imputing zeroes (or ones) for all missing values of a discrete (0/1) variable or using a binomial process to randomly allocate zeroes and ones to missing values. For continuous variables, researchers often use mean imputation, i.e., replacing all missing values with the mean of the observed values. Yet another ad hoc value imputation alternative is to look for 'matching' observations (those that are in every respect identical, but are completely observed). This occurs, for example, when proxy observations (or even replacing the variable with missing values) are used. While the sample size is maximised, potential measurement error is introduced in each of the ad hoc approaches.

Assuming that the missingness pattern is not completely at random, values for the observed variables provide indirect evidence about the likely values of the unobserved variables. Under certain conditions, this then implies a predictive probability distribution for missing values over which one could average in statistical analysis of the data. To exploit this predictability some researchers adopt a 'one-shot' fitted value approach – the *regression imputation* method – to handle missing data. Using the complete observations only, an auxiliary regression is run. The parameter estimates are then used to fit the missing observations. This paper will show that such an approach is one step towards the preferred imputation method. The application illustrates that under certain circumstances such a method might provide satisfactory results.

The missing data problems are accentuated in simultaneous equation studies of interrelated dependent variables. The missing data exclusion method – now known as the *pairwise deletion* method – applied commonly across all equations may dramatically reduce sample size and introduce bias into estimates. The *regression imputation* methodology can then also be adopted and is intuitively even more appealing given that there are more 'information channels' than in the single equation context. However, in both instances the implications of *regression imputation* are not straightforward. Whereas *pairwise deletion* in simultaneous equations systems may lead to inconsistency in the covariance matrix and biased estimates, *regression imputation* may also lead to biased estimates. Also, the imputed values are really estimates. Ignoring the uncertainty of missing value prediction will lead to standard errors that are too small.

The current paper proposes a novel statistical framework that minimizes the problems associated with missing values. Based on either maximum likelihood or Bayesian estimation methods it is often possible to obtain unbiased imputed values for the missing observations with correct standard errors. The statistical literature on missing data problems dates back to the mid-1970s with Rubin's (1976, 1978) and Dempster, Laird and Rubin's (1977) papers. These papers proposed an iterative maximum likelihood based procedure to impute 'most likely' values for the

missing data. Tanner and Wong (1987) introduced a Bayesian equivalent of this methodology. There are many advantages to both iterative maximization methods when direct maximization of the incomplete data likelihood is not an option. This likelihood is computationally complex and convergence of maximization is not guaranteed. The iterative methods on the other hand are computationally straightforward, easy to program in specific cases, and also generate fitted values for the missing data.

Dempster, Laird and Rubin first suggested the by now standard *EM*-algorithm for imputing values of incomplete observations. This estimation procedure consists of iteratively computing the conditional expected values of incomplete observations, substituting these for the incomplete observations, and then estimating the unknown parameters to maximize the complete data likelihood. Whereas its convergence properties are very attractive, the *EM*-algorithm does not explicitly account for the uncertainty surrounding the missing value imputations (a shortcoming it shares with the previously described naive imputation methods).

An alternative approach suggested by Tanner and Wong (1987), the *Imputation Posterior (IP)*-method, explicitly accounts for the imputation uncertainty. This Bayesian alternative to the *EM*-methodology is also based on an iterative (posterior density) maximization procedure where the *I*-step imputes missing values based on an initial random draw of parameter values and the *P*-step then computes new parameter values from a Bayesian posterior distribution. *IP* convergence occurs in distribution to the exact data likelihood. Therefore, according to Schafer (1997), it should be considered the "gold method" to deal with missing data problems. Unfortunately, exactly when this convergence occurs is difficult to assess and requires some Bayesian expertise. Moreover, the method is also computationally intensive and might not be suitable for large datasets that are common in financial applications.

Hence, it would be optimal to use a method that has the 'exactness' of the *IP*-method while retaining the computational simplicity of the *EM*-method. Rubin (1978) and more recently Schafer (1997) develop a number of extensions to the *EM*-algorithm which solve its basic shortcoming. The *EM-sampling* and *EM-importance sampling* methods add a Bayesian flavour to the classical *EM* approach. That is, they allow for the uncertainty regarding the imputations of missing values. Thus, they approach the exactness of the *IP*-method.

This paper first surveys the pervasiveness of missing data in the applied finance literature. As it turns out, financial researchers typically treat missing data by discarding observations or, at best, by using ad hoc methods. The novel imputation methodologies referred to above have better

properties than these methods and should therefore be considered for application. Whereas the statistical literature has focused on single equation estimation with continuous dependent variables, this paper also considers simultaneous equations with discrete dependent variables. This leads to some interesting simulation evidence that has not been recorded before. The imputation methodologies are then applied to a financial dataset with an extremely high degree of missing data. To test the imputation methods on this real dataset, a cross-validation experiment is designed which predicts missing values. The iterative *EM-importance sampling* method is found to outperform the alternatives.

The paper is organized as follows. The next section discusses the missing data problem in the financial literature. Then, as these methods are not well known in applied finance, a brief overview of the multiple imputation methods is given. To illustrate these methods, a financial application is given for a simultaneous equation model where the interdependent variables are characterized by a high degree of missingness. The results of the consistent and efficient *EM*-extended and *IP* estimates are then compared with the 'naive' approaches typically used in the finance literature. Finally, the predictive ability of these imputation methods is evaluated.

I. Survey

An examination of papers published in four recent volumes of five international journals in banking and finance,[1] as summarised in Table I, suggests that missing observations is a common feature of many financial applications. In total, 175 articles (out of 1057)[2] were identified where authors explicitly recognised their treatment of missing data. These authors frequently describe how their sample has been reduced, or observations amended, because of missing values of <u>independent</u> variables. Somewhat less common, however, is a recognition that samples have been reduced, or modified, because of missing <u>dependent</u> variable observations. This may reflect the fact that missing dependent variable observations are often easily concealed in single equation studies by a statement such as *"N observations of the (dependent) variable were available and collected for study."* Thus, the results in Table I potentially underestimate the extent of the missing data problem in finance. Complicating the investigation is the fact that data descriptions are frequently incomplete and/or are hardly informative with regard to sources, availability, completeness, and

[1] The journals included the *Journal of Banking and Finance*, Vols.19-22 , the *Journal of Finance*, Vols.50-53, the *Journal of Financial Economics*, Vols.37-50, the *Journal of Financial and Quantitative Analysis*, Vols.30-33, and the *Review of Financial Studies*, Vols.9-11.
[2] These include theoretical as well as empirical papers.

transformations applied. For replication purposes, and to shed light on the reliability and significance of the results, it would be desirable if such information became more commonly available.

A second reason why missing data problems do not seem to occur as often in dependent variables might be that editors and/or referees may be more inclined to reject studies with missing dependent variables. There may also be significant self-selection. Whereas it may be considered straightforward to estimate a model with some missing observations in the explanatory variables, a similar missingness among the dependent variables may often seem more complicated.

Table I

Literature Survey of Missing Data in Finance: 1995-1998

This table presents information regarding the number of papers that acknowledge the presence of a missing data problem; in what type of variable the problem occurs; and, in what type of empirical application the problem occurs. The total number of papers appearing in this journal is given in parentheses. The journals include the *Journal of Banking and Finance* (JBF), the *Journal of Finance* (JF), the *Journal of Financial Economics* (JFE), the *Journal of Financial and Quantitative Analysis* (JFQA), and the *Review of Financial Studies* (RFS).

Journals	Papers (1)	Independent Variable (2)	Dependent Variable (3)	Both (4)	Cross Section (5)	Time Series (6)
JBF	42 (327)	32	4	6	31 (179)	20 (175)
JF	70 (289)	61	7	2	59 (191)	15 (198)
JFE	38 (194)	29	-	9	36 (166)	12 (159)
JFQA	15 (112)	13	1	1	11 (66)	5 (69)
RFS	10 (135)	7	2	1	4 (54)	9 (58)
	175 (1057)	142	14	19	143 (656)	63 (659)

Table I also investigates the relationship between the missingness problem and the analysis type. The numbers in columns (5)-(6) measure the number of papers with a missing data problem in a time-series application, in a cross-section application, or (if it uses a combination of these) in both applications.[3] The numbers in parentheses in these columns are the total numbers of papers of the analysis type. Event studies (or asset pricing models) use a mixture of time-series analysis to estimate excess returns (to estimate asset betas) and cross-section analysis to estimate event parameters (to estimate risk premia). Hence, these studies are counted under both columns (5) and

[3] Note that even if the empirical analysis was based on a combination of time-series and cross-section, it often occurred that the missing data problem was only relevant for either the time-series or the cross-section, but not for both.

(6). The high frequency with which this occurs highlights the predominance of these studies in the finance literature. Missing data problems are often assumed to prevail in cross-section studies. There is some evidence for this phenomenon. However, editors and referees may deem it more acceptable for cross-section studies to have missing data than for time-series studies.

Table II

Treatment of Missing Data in Finance: 1995-1998

This table presents information regarding the treatment of a missing data problem in those papers that acknowledge the presence of a missing data problem. The journals include the *Journal of Banking and Finance* (JBF), the *Journal of Finance* (JF), the *Journal of Financial Economics* (JFE), the *Journal of Financial and Quantitative Analysis* (JFQA), and the *Review of Financial Studies* (RFS).

Journals	Papers	Listwise Deletion	Regression Imputation	Ad Hoc Imputation	Proxy Imputation
JBF	42	35	4	2	1
JF	70	54	6	6	4
JFE	38	30	2	3	3
JFQA	15	14	-	-	1
RFS	10	4	3	3	-
	175	137	15	14	9

Table II summarises the missing data methodology applied when the researcher was confronted with a missing data problem. In most of the missing value cases, the solution adopted in the paper was listwise deletion (in 137 cases) while regression imputation, ad hoc imputation and proxy imputation were infrequently used. Note, however, that in a substantial number of cases this information was derived indirectly from the data description.

II. *EM* and *IP* Imputation Algorithms

Excellent statistical treatments of missing data imputation methodologies can be found in Rubin (1987) and Schafer (1997). Since this literature is not commonly adopted by finance practitioners, it seems worthwhile to provide a synopsis of the theory. To understand the consequences of missing values and potential solutions for statistical analysis with missing values, some idea of why and how missing values occur is needed. The occurrence of missing values can be captured by three distinct missingness schemes. They are distinguished by whether the source of missingness is

internal to the dataset, external to the dataset, or completely independent from the dataset. Suppose x is a completely observed independent variable for an incomplete dependent variable y. The missingness pattern in y is now said to be:

- *Missing Completely At Random (MCAR)*
  when the missingness in y is independent of both x and y. The missing data are then missing-at-random, while the observed data are observed-at-random
- *Missing At Random (MAR)*
  when the missingness in y depends on x but not on y. Missing data are still missing-at-random, but the observed data are no longer observed-at-random.
- *Non-Ignorable (NI)*
  when the missingness in y depends on y and possibly also on x.

To illustrate the distinction in missingness types, consider the following application based on Pulvino (1998). The dependent variable in Pulvino's hedonic regression model is the transaction price of used aircraft. Independent variables in this regression are aircraft characteristics (e.g., the age of the aircraft). The data series is based on aircraft transactions from 1978 to 1993. Post-1991 transactions are, however, excluded due to some missing observations in the transaction prices. Let us assume that these post-1991 transaction prices are missing simply because of data handling. If it can reasonably be assumed that the data manager makes these mistakes at random, then the missingness type is MCAR. Instead, consider that the data manager is not to blame, but missingness depends on the age of the aircraft. The older the aircraft, the less likely it is that its transaction price gets reported. The missingness is now of type MAR. However, for this particular data series Pulvino (p.947) notes that: "*...1991 transactions are included in the Avmark database only when prices were voluntarily disclosed or reported in other public sources. To preclude sample selection bias, transactions that occurred after 1991 are excluded from the analyses that follow*." This could imply that parties involved in a fire sale are less likely to report extreme transaction prices (i.e., very high and very low). Hence, missingness in the dependent variable is now a function of the dependent variable itself. This is known as type NI.

Clearly, when the missingness is determined outside the dataset, as it is with NI, it is impossible to infer likely values for the missing data from the observed data. As long as the missing data scheme is not NI, however, likelihood-based imputation methods can be used to generate unbiased estimates for the complete data statistical model. Hence, if the MAR assumption is reasonable, then among the x the distribution of y is the same for y_{obs} (the observed dependent variables) as it is for y_{mis} (the missing observations). It implies that the relationship between x and y for the observed data can be extrapolated to the missing data, for which we do observe the x-values.

This is known as the ignorability assumption. Of course, it is impossible to test the ignorability assumption against the NI assumption.

For now, assume that the data satisfies the ignorability condition. Consider the complete data matrix Y (of dimension nxp), that is not completely observed. It can be partitioned according to missingness status, such that $Y=(Y_{obs},Y_{mis})$. Assuming the observations are independently and identically distributed (*i.i.d.)*, the probability density function (*pdf)* of the complete data can be written as

$$P(Y|\theta)=\prod_{i=1}^{n} f(y_i \mid \theta) \tag{1}$$

the product of the n densities $f(.)$ for the individual observations y_i. These densities are conditional on a set of parameters θ for which unbiased estimates and their correct standard errors are of interest. Given that there are missing values in Y, this is not trivial. That is, the parameters θ pertain to the complete data, but this dataset is only partially observed. A matrix Π of the same dimension nxp as the data matrix can be introduced to indicate which part of Y is observed (zeroes in Π) and which part is missing (ones in Π). The probability of encountering a missing value (a one in Π) is conditional on the observed data, the missing data and a set of nuisance parameters ϑ in the most general missingness model. Such a model coincides with the assumption of NI missingness. Assuming the missingness type is MCAR, the probability distribution of this missingness matrix Π simplifies to

$$P(\Pi \mid Y_{obs},Y_{mis},\vartheta)=P(\Pi) \tag{2}$$

It does not depend on any available information in the data (missing or observed). Missingness is truly random. Assuming the missingness type is MAR, the probability distribution of this missingness matrix Π simplifies to

$$P(\Pi \mid Y_{obs},Y_{mis},\vartheta)=P(\Pi \mid Y_{obs},\vartheta) \tag{3}$$

It does not depend on the missing data, but it does depend on the observed data. Combining the simplification in (3) with equation (1) and assuming that θ,ϑ are distinct nuisance parameters, see Schafer (1997), leads to

$$P(\Pi,Y_{obs} \mid \theta,\vartheta)=P(\Pi \mid Y_{obs},\vartheta)\int P(Y \mid \theta)dY_{mis} = P(\Pi \mid Y_{obs},\vartheta)P(Y_{obs},\theta) \tag{4}$$

These two assumptions (MAR and distinctness) allow ‘ignorability of the missingness model,’ see Rubin (1987). Equation (4) then implies that likelihood estimation of θ (the parameters of interest)

is unaffected by the model for missingness. That, however, does not imply that the missing observations are of no consequence to inference on θ. It does say that all the necessary information to 'complete' the data (i.e., to fill in the missing values) is contained in the observed data.

As an example, consider the case where a dataset consists of two possibly related variables. Y_1 is complete, but Y_2 has some missing observations. If the data are rearranged such that observations 1 to j are completely observed (both Y_1 and Y_2 are available), and observations j+1 to n have missing values for Y_2, the observed data likelihood can be written as:

$$L(\theta \mid Y_{obs}) = \int \prod_{i=1}^{j} P(y_{i1}, y_{i2} \mid \theta) \prod_{i=j+1}^{n} P(y_{i1} \mid \theta) \prod_{i=j+1}^{n} P(y_{i2} \mid y_{i1}, \theta) dY_{mis} \tag{5}$$

where the first two terms do not involve missing values and the last term integrates to one, then

$$L(\theta \mid Y_{obs}) = \prod_{i=1}^{j} P(y_{i1}, y_{i2} \mid \theta) \prod_{i=j+1}^{n} P(y_{i1} \mid \theta) = \prod_{i=1}^{n} P(y_{i1} \mid \theta) \prod_{i=1}^{j} P(y_{i2} \mid y_{i1}, \theta) \tag{6}$$

If, e.g., the data density is assumed bivariate normally distributed with means μ, and variance-covariance matrix Σ, then the data-likelihood for the parameters can be written as

$$L(\theta \mid Y_{obs}) \propto |\Sigma|^{-\frac{j}{2}} \exp\left\{ -\tfrac{1}{2} \sum_{i=1}^{j} (y_i - \mu)' \Sigma^{-1} (y_i - \mu) \right\} \times \sigma_{11}^{-(n-j)/2} \exp\left\{ -\frac{1}{2\sigma_{11}} \sum_{i=j+1}^{n} (y_{i1} - \mu_1)^2 \right\} \tag{7}$$

Generalizing this to multivariate normally distributed data with arbitrary missingness patterns throughout the p variables, there will be 2^p possible missingness patterns. Of course, not all occur and the unique missingness patterns can be summarized as $\pi=1,2,..,\Pi$, and $I(\pi)$ is then a subset of rows with this particular pattern. The observed data likelihood then looks like

$$L(\theta \mid Y_{obs}) = \prod_{\pi=1}^{\Pi} \prod_{i \in I(\Pi)} \left| \Sigma_{\pi}^{*} \right|^{-\frac{1}{2}} \exp\left\{ -\tfrac{1}{2} \left(y_i^{*} - \mu_{\pi}^{*} \right)' \Sigma_{\pi}^{*-1} \left(y_i^{*} - \mu_{\pi}^{*} \right) \right\} \tag{8}$$

where a * indicates the observed part in missingness pattern π. Clearly, this can become a complicated function of individual means and (co)variances. Its derivatives with respect to these means and (co)variances are very complicated. There are no closed form solutions, nor is it straightforward to compute this expression numerically. Instead, some kind of iterative procedure is needed. The iteration has intuitive appeal. It exploits the interdependence between the missing values and the complete data parameters of interest. Under the MAR assumption, the complete data parameters have information relevant to estimating likely missing values. At the same time the missing values have relevant information with regard to the parameter estimates. This interaction

has been captured by the following two iterative techniques that are commonly considered and used for this purpose.

A. IP Method

The *Imputation-Posterior* methodology, discussed in Tanner and Wong (1987), is an iterative Bayesian procedure imputing values for the missing observations and making inference about unknown parameters in a stochastic manner. *IP* initially imputes missing observations randomly based on a suitable prior for the parameters $\tilde{\theta}$ of the imputation model, and then samples new parameter values from a Bayesian posterior distribution based on both observed and imputed data. Thus, we begin by selecting starting values ($\tilde{\theta}^{(0)}$) and sample from

$$\tilde{Y}_{mis}^{(i)} \sim P\left(Y_{mis} \mid Y_{obs}, \tilde{\theta}^{(i-1)}\right) \tag{9}$$

which is the Imputation (*I*) step, and then sample from

$$\tilde{\theta}^{(i)} \sim P\left(\theta \mid Y_{obs}, \tilde{Y}_{mis}^{(i)}\right) \tag{10}$$

which is the Posterior (*P*) step. Iteratively sampling for i=1,2,..,N whilst updating the conditioning variables produces a sample of N parameter sets (of parameter values) which converge in distribution to the posterior distribution $P(\theta \mid Y)$. Features of the posterior distribution, its marginal densities and probability intervals, can be extracted from this sample. An advantage of the *IP* approach is that the parameter distribution converges to a posterior distribution averaging over the missing observations. The parameter distribution and missing observations distribution converge to an (exact) predictive distribution for θ and Y_{mis} respectively. However, a disadvantage of the *IP*-method according to Schafer (1997) is that its convergence is difficult to evaluate and the method is computationally time-intensive (N will typically be large). The subjectivity involved in judging convergence is of particular concern. Even though the method is theoretically exact, it will be difficult to determine when this exactness has been achieved. Incorrect judgments may then still lead to biased outcomes.

B. EM Algorithm

A less computationally intensive alternative to the Bayesian *IP*-method is the *Expectation-Maximization* algorithm, first introduced by Dempster, Laird and Rubin (1977). The distribution of the complete (but incompletely observed) data Y can be split up as follows:

$$P(Y \mid \theta) = P(Y_{obs} \mid \theta) P(Y_{mis} \mid Y_{obs}, \theta) \tag{11}$$

and writing down the data likelihood for the parameters of interest

$$L(\theta \mid Y) = L(\theta \mid Y_{obs}) + \ln P(Y_{mis} \mid Y_{obs}, \theta) + c \tag{12}$$

with *c*, a constant. The second term on the right hand side of (12) is a predictive distribution of the missing data given the observed data and the parameters. Given that we do not observe the missing values, this term cannot be computed. Instead, by replacing θ by $\tilde{\theta}^{(i)}$, which for i=0 is an initial estimate of the unknown parameters, the *EM*-method averages iteratively over the predictive distribution. Each iteration consists of two steps, the Expectation (*E*) step and the Maximization (*M*) step. The *E*-step estimates the sufficient statistics of the complete data *Y*, given the observed data Y_{obs} and the initial parameter estimate $\tilde{\theta}^{(0)}$. It then computes expected values for the missing data $\tilde{Y}_{mis}^{(1)}$. The *M*-step then takes the estimated complete data $\tilde{Y}^{(1)}$ and estimates the unknown parameters $\tilde{\theta}^{(1)}$ by maximum likelihood as though the estimated complete data were the observed data. Then these two steps are iterated (where $\tilde{\theta}^{(1)}$ is the new parameter estimate to find updated expected values for the missing data $\tilde{Y}_{mis}^{(2)}$) until convergence is achieved based on $\tilde{\theta}^{(i)} \approx \tilde{\theta}^{(i-1)}$. Convergence occurs without any assumptions on the derivatives or the starting values and will also occur for small sample sizes. This is clearly an advantage over the subjectivity involved in judging *IP* convergence. However, convergence is based on the parameter estimates' contribution to the likelihood. While the *EM* algorithm finds the maximum of the likelihood function for the parameters and missing values, unlike *IP* it does not identify the full parameter distribution for θ and Y_{mis}. Both the *EM*-imputations, and the *EM*-parameters are single values (maximum posterior), instead of a complete distribution. According to Schafer (1997) the method therefore ignores estimation uncertainty and consequently underestimates the standard errors.

Despite its computational simplicity and good convergence properties, the *EM*-algorithm has so far attracted little attention outside theoretical analyses and applications with a purely statistics focus. Very few finance and/or economic applications have so far utilized the *EM* methodology. That is somewhat surprising given that incomplete data problems can and do occur in almost every type of financial research as witnessed in Section I of this paper. There are a few exceptions. Kalb, Kofman and Vorst (1995) for example, illustrate how the algorithm can be operationalized in single equation missing observation problems for actuarial applications where there are mixtures of continuously and discretely valued variables. Malhotra (1987) shows how this can be done for probit, and tobit estimation problems in marketing research.

C. EM extensions

Tanner (1996) provides a related method, known as *EM-sampling (EM-s)*, that reintroduces imputation uncertainty into the *EM*-algorithm. *EM-s* first applies the standard *EM*-algorithm to generate the maximum posterior parameter estimates $\tilde{\theta}^*$ (the converged mean parameter values) and computes its variance matrix $V(\tilde{\theta}^*)$. To determine an imputation uncertainty adjusted parameter estimate $\hat{\theta}$, a simulated parameter value is drawn from a normal distribution with mean $\tilde{\theta}^*$ and variance $V(\tilde{\theta}^*)$. By sampling from this parameter distribution, an imputation for the missing values $\hat{Y}_{mis}$ is obtained. This sampling procedure is repeated *m* times (after the initial *EM*-step and variance computation) which generates a sampling distribution for Y_{mis}. According to King et al. (1998) the *EM-s* approach tends to find the true mode of this distribution well, but with highly skewed (non-normal) categorical data it can produce incorrect standard errors. This is a serious shortcoming for financial applications where non-normality is a common characteristic.

Rubin (1987), Tanner (1996), and King et al. (1998) use the same procedure as *EM-sampling*, except that draws from the initial *EM* are treated as first approximations to the true posterior. Their *EM-importance sampling* procedure then uses an acceptance-rejection algorithm that keeps parameter draws $\hat{\theta}$ with probability proportional to the *importance ratio*:

$$IR = \frac{L(\hat{\theta} \mid Y_{obs})}{N(\hat{\theta} \mid \tilde{\theta}^*, V(\tilde{\theta}^*))} \tag{13}$$

and discards the rest. The *IR* is a ratio of the actual posterior distribution to the asymptotic normal approximation. This implies that the likelihood is evaluated at the important segments of the range more frequently than otherwise. The parameter simulations are then used to produce the imputations $\hat{Y}_{mis}$ similar to *EM-sampling*. This method is fast, and its imputations are based on the exact finite sample posterior distribution. Unlike the *IP*-method it does not require Markov chains and convergence is therefore still easily determined. As suggested by Schafer (1997), it might not do so well for seriously non-normal likelihoods.

D. Multiple Imputation

Rather than estimate the incomplete data likelihood function directly, the iterative *IP* and *EM*(extended) methods start with an initial parameter vector, and within each iteration the incomplete data problem is converted to a complete data problem by replacing each of the incomplete observations by their simulated expected values, conditioning upon all extraneous information available. That is, implementation requires the researcher to include as many variables

as possible in the imputation stage. Including as much information as possible might prevent the *NI* problem, when a variable outside the analysis model could explain missing values. This suggests that the imputation model will most likely diverge from the analysis model, which is of ultimate interest to the researcher. The analysis model will typically be chosen based on model selection rules to prevent an overspecified model. King et al. (1998) argues that the risk of overspecification is not an issue of concern in the imputation stage.

To accommodate this 'separation' between imputation and analysis model, the two iterative imputation approaches can be used in either of two ways: (i) to simulate *m* parameter values $\tilde{\theta}^{*(1)},\tilde{\theta}^{*(2)},...,\tilde{\theta}^{*(m)}$ from the observed data posterior distribution for the parameters $P(\theta|Y_{obs})$; or (ii) to simulate *m* missing values $\tilde{Y}_{mis}^{*(1)},\tilde{Y}_{mis}^{*(2)},...,\tilde{Y}_{mis}^{*(m)}$ from the observed data posterior distribution for the missing values $P(Y_{mis}|Y_{obs})$. The first way, known as *parameter simulation*, seems attractive since it directly provides the required parameter estimates. However, the encompassing imputation model does not necessarily match the (smaller) analysis model. The larger imputation model may violate proper model selection rules, and the $\tilde{\theta}^{*(i)}$ estimates may not be the parameter estimates of interest. Also, *parameter simulation* imposes a certain structure on the parameter estimates based on the assumptions underlying the imputation methodology. The ultimate parameter estimates may be seriously misleading if these assumptions are violated. The second way to utilize the iterative imputation methods was introduced by Rubin (1978), known as *multiple imputation*, where the Y_{mis} are replaced successively by simulated values $\tilde{Y}_{mis}^{*(i)}$ and each of the *m* complete datasets are analyzed by standard analysis. The variability among the *m* analysis results provides a measure of the uncertainty due to missingness and, when combined with sample variation, gives a single measure for the parameter of interest. Unlike *parameter simulation*, now only the missing values are 'affected' by the imputation methodology. Violating the assumptions underlying the imputation methodology will then be less dramatic since its impact on the parameter estimates will be diminished by the actually observed observations.

Assuming multivariate normality implies that the missing observations can be imputed linearly, like simulating from a regression

$$\tilde{y}_{ij}^{*} = y_{i,-j}\tilde{\lambda}^{*} + \tilde{\varepsilon}_{i} \tag{14}$$

where $\tilde{y}_{ij}^{*}$ is a simulated value for missing observation *i* and variable *j*, $y_{i,-j}$ is the vector of all observed variables for this observation, and $\tilde{\lambda}^{*}$ is computed from a random draw of the observed data posterior distribution for the parameters $\tilde{\theta}^{*}$. The disturbance term $\tilde{\varepsilon}_{i}$ is a random draw from a

standard normal distribution. Hence, multiple imputation requires m independent draws for the missing observations from a posterior predictive distribution for θ.

Having computed the m different imputed datasets,[4] inference can then be drawn with respect to the parameters b, and their variances s_b for the analysis model. These parameters will then take explicit account of the uncertainty regarding the imputed missing values. They may not match the imputation model parameters λ, since the analysis model will typically be of smaller dimension than the imputation model. The successive estimation of the analysis model generates m equally likely estimates $(b_1,..,b_m;s_1,..,s_m)$ estimates for the parameters and their variances, respectively. The combined multiple imputation estimate is then:

$$\bar{b} = \frac{1}{m}\sum_{i=1}^{m} b_i \tag{15}$$

and *within-* and *between-imputation* variance, respectively of:

$$\begin{aligned} \bar{s} &= \frac{1}{m}\sum_{i=1}^{m} s_i \\ s_m &= \frac{1}{m-1}\sum_{i=1}^{m}\left(b_i - \bar{b}\right)^2 \end{aligned} \tag{16}$$

and total variance, standard error and confidence interval,

$$\begin{aligned} s_b &= \bar{s} + \left(1+\tfrac{1}{m}\right)s_m \\ se_{\bar{b}} &= \sqrt{s_b} \\ \bar{b} &\pm t_{df}\, se_{\bar{b}} \quad with \quad df = (m-1)\left(1+\frac{\bar{s}}{\left(\frac{1}{m}+1\right)s_m}\right)^2 \end{aligned} \tag{17}$$

respectively. Hence, the uncertainty with regard to the parameter estimate decreases with the number of imputations through the degrees of freedom expression in (17). The parameter uncertainty increases with the ratio

$$r_b = \frac{\left(\frac{1}{m}+1\right)s_m}{\bar{s}} \tag{18}$$

[4] The imputation models for this paper have been estimated using the Gauss programs readily available on the WWW-page of King: http://Gking.Harvard.edu explained in Honaker, Joseph, King, and Scheve (1999). S-PLUS programs for *IP* and *EM* imputation are available on the WWW-page of Schafer: http://www.stat.psu.edu/~jls explained in Schafer (1997). Standard packages like Stata and SPSS nowadays have some missing data imputation options, but they are restricted to *regression imputation* and/or *mean imputation.* The analysis models for this paper have been estimated using Gauss programs written by the authors of this paper, and are available on request.

which measures the relative increase in variance due to missingness, Rubin (1987). The efficiency of an estimate for a single parameter *b* based on *m* imputations is approximately $(1+(\pi_b/m))^{-1}$ with π_b the fraction of missing information with regard to *b*, see Rubin (1987):

$$\pi_b = \frac{r_b + \frac{2}{df+3}}{r_b + 1} \tag{19}$$

Typically, only a few imputations (<10) are necessary to achieve a high degree of efficiency. Further imputations will only contribute minor efficiency gains. It is nevertheless worthwhile to compute the fraction of missingness about *b*.

III. Monte Carlo Experiments

To assess the relative size of the bias and/or (in)efficiency for these missing value imputation methodologies vis-à-vis simply discarding observations with missing values, a series of Monte Carlo experiments were designed. Consider a complete dataset based on the single equation model $Y = X\beta + e$ with a single explanatory variable *X*, a known correlation parameter ρ and i.i.d. innovations *e* that are drawn from a standard normal distribution. The Monte Carlo experiments are conducted for two different types of dependent variable *Y*, continuous and discrete (0/1)[5] respectively. The latter is particularly relevant for the application that follows.

Three missing value datasets were then created from this complete dataset. The first is based on the MCAR scheme, where a fraction κ of the dependent variable values are eliminated completely at random by drawing from a binomial distribution independent of both *X* and *Y*. The second dataset is based on the MAR scheme, where a fraction κ of the dependent variable values are eliminated based on a missingness function *g(X)*. Hence, the missingness depends on the values of the explanatory variable *X*. The third dataset is based on the NI scheme, where a fraction κ of the dependent variable values are eliminated based on a missingness function *g(Y)*. Hence, the missingness depends on the values of the dependent variable *Y* itself.

The models are then estimated based on listwise deletion and the *EM-is* algorithm,[6] respectively. In addition the Monte Carlo is run over a range of values for ρ, (0, 0.1,...,0.9) and a range of values for κ (0.25, 0.3,..., 0.75) in order to examine the sensitivity of the mean squared error (*mse*) to strength of correlation, and degree of missingness, respectively. For each

[5] The benchmark experiments, i.e., the continuous variables examples, can also be found in Schafer (1997) and King et al. (1998).

combination, the estimation procedure is repeated a large number of times (1000 random draws of innovations). The results are summarized in Figures 1, 2 and 3 for a discrete dependent variable, for MCAR, MAR, and NI respectively. The continuous dependent variable equivalents are given in Figures 1a, 2a and 3a.

INSERT FIGURES 1-2-3

To assess the relative contribution of bias and (in)efficiency to *mse*, the bias function is given separately. The bias and *mse* are calculated for the mean (expected) value of the 'completed' dependent variable distribution $P(Y|\theta)$. The results are very clear. For the experiment with a discrete dependent variable, MCAR missingness does not generate significant bias nor inefficiency as illustrated in Figure 1. Listwise deletion performs just as well as the more complicated *EM-is* method. With MAR missingness in Figure 2 listwise deletion results in significant negative bias. This bias increases with increasing ρ, and (less so) with increasing missingness. The first effect dominates the latter. Also noteworthy is the fact that even for small ρ *(<0.5)*, the bias is already significant. The *EM-is* method, on the other hand, now clearly outperforms listwise deletion. Nevertheless, it still has some (significant) bias at high degrees of correlation. This seems to be driven by the fact that the dependent variable is discrete. For a continuous dependent variable *EM-is* has insignificant bias. Both methods do not fare well when the missingness is NI as illustrated in Figure 3. For both methods, the bias is increasing with missingness, but decreasing in correlation.

Next, the Monte Carlo experiment is extended to a simultaneous equation setting where the related dependent variables are drawn from a bivariate normal distribution with known correlation matrix and repeated for the case where one dependent variable is a discrete (0/1) variable. Once again, a fraction κ of the dependent variables values are eliminated by MCAR, MAR, and NI respectively. Now, missingness occurs in both dependent variables. The models that are reported below are based on *pairwise deletion* and the *EM-is* algorithm. As before, simulations are based on a range of missingness fractions as well as a range of correlation. Note that the relevant correlation measure is the one between the dependent variables. The simultaneous estimation procedure is replicated a large number of times (1000 random draws of innovations), and the results illustrating the bias and *mse* in the parameter estimates are summarized in Figures 4, 5 and 6. Results for the case where one dependent variable is discrete (and the other is continuous) are displayed in Figures 4, 5 and 6. For comparison, the equivalents where both dependent variables are continuous are

[6] For the purpose of comparison, only the preferred imputation methodology (EM-is) and the most common naive (listwise/pairwise deletion) imputation methodology are presented here. Results for the other EM and IP estimation methods are available from the authors upon request.

given in Figures 4a, 5a, and 6a. The correlations of the dependent variables with their respective explanatory variables are fixed at 0.75. The bias and *mse* are calculated for the mean (expected) value of the discrete 'completed' dependent variable distribution $P(Y_1|\theta)$.

INSERT FIGURES 4-5-6

The results are very similar to those obtained from single equation experiments. First, for the MCAR case in Figure 4, the *pairwise deletion* and *EM-is* method perform similarly. As before with *listwise deletion* in the single equation results, *pairwise deletion* performs badly when the missingness scheme is MAR in Figure 5. Bias is increasing with missingness and in correlation. In contrast to the single equation results, *EM-is* now performs really well. Even for high degrees of correlation and missingness, there is very limited (and fairly constant) evidence of bias. The information embedded in the feedback structure of the simultaneous equations seems to have eliminated the bias at higher correlations found in Figure 2. The results for the NI-case in Figure 6 are less dramatic than for the single equation examples. Bias is still significant for both *pairwise deletion* and *EM-is* method, though less pronounced than in the single equation. The negative bias still decreases with correlation and increases with missingness.

IV. Illustrative Example

In a recent study, Dennis, Nandy and Sharpe (1999), hereafter referred to as DNS, model debt contract terms (duration, secured status, all-in-spread and undrawn commitment fee) on bank revolving lines of credit where two of the four dependent variables in the simultaneous equation system were subject to missing observations. DNS adopt a different approach for each of the problem variables. For the secured status equation they use listwise deletion, effectively losing 1331 of the total 2634 observations. On the other hand, in the commitment fee equation they use ad hoc imputation, assuming that the 877 missing values of the 2634 observations have zero commitment fees. However, where secured status appears as a determinant of each of the other three contract terms, the authors adopt a regression imputation method. In this case the secured status reduced form estimates using 1303 observations were used to obtain out-of-sample fitted values for the N=1331 missing observations allowing the duration, the all-in-spread, and the commitment fee equations to be estimated on the full N=2634 sample.

The DNS model and data provides a particularly interesting application for which to compare the imputation methods proposed in the previous sections. Two of the four dependent variables within a simultaneous equation system are subject to a high rate of missing observations

(51% for secured status, and 33% for commitment fee, respectively). This is even further inflated when these variables are considered jointly, with the sample size reducing to just 896 observations (i.e., 66% missingness). Moreover, the dependent variables involve a mix of continuous dependent variables (duration and all-in-spread), a discrete choice dependent variable (secured status) and a censored-from-below-at-zero dependent variable (commitment fee).[7] Finally, the data is interesting as the missing observations have somewhat different characteristics vis-à-vis the non-missing data.

The first step is to use an informal tool to assess whether the missingness scheme is likely to be MCAR or MAR (note that NI cannot be established from the observed data). The dataset is divided according to missingness status for the dependent variable. The frequency distribution of an explanatory variable can then be plotted for those observations that have a dependent variable value missing and for those that are complete. Under MCAR, a similar distribution shape is expected regardless of missingness status. Under MAR, a distinctly different shape should be expected for the missing dataset in comparison with the complete dataset. The graphs in Figure 7 clearly illustrate this difference.

INSERT FIGURE 7

The panels on the left in Figure 7 condition the data on the missingness status of the commitment fee (COMFEE) dependent variable. There is no apparent difference between the conditional histograms for the Z-score (ZSCR) explanatory variable. The top two panels on the right in Figure 7 condition the data on the missingness status of the secured status (SECURED) dependent variable. There is a distinct difference in shape between the two histograms for the all-in-spread (SPRD) explanatory variable. To further investigate this difference, the data for which secured status (SECURED) is observed is further conditioned on its outcome (whether it is a 0 or a 1). The histograms for the all-in-spread (SPRD) explanatory variable are displayed in the bottom two panels on the right in Figure 7. The histogram for missing secured status now looks similar to the histogram for SECURED=0. Hence, the missing values for SECURED do not seem to be missing <u>completely</u> at random, but seem more likely to be zeroes than ones. Clearly, such statements are not very satisfactory for a complicated model like this. A more formal procedure is needed.

DNS model the four debt contract features as an interrelated system in which borrowers trade off loan characteristics. However, only three of the contract features may be independently chosen. Reflecting this independence feature, they jointly model the choice of DURATION and

[7] The commitment fee is treated as a continuous variable in the imputation stage for reasons of comparison with the results of Dennis, Nandy and Sharpe (1999). In the next section, the commitment fee is treated as a discrete variable to

SECURED status and then model ALL-IN-SPREAD and COMFEE as being determined by the choice of DURATION and SECURED status. While DNS assume unidirectional relationships from both DURATION and SECURED to ALL-IN-SPREAD and COMFEE, they allow bi-directional relationships between ALL-IN-SPREAD and COMFEE and between DURATION and SECURED. The model takes the following form:

$$\begin{array}{lllll} DURATION & = & \gamma_{DS} SECURED & & +\beta_D' X_D + e_D \\ SECURED & = \gamma_{SD} DURATION & & & +\beta_S' X_S + e_S \\ ALL-IN-SPREAD & = \gamma_{AD} DURATION & +\gamma_{AS} SECURED & +\gamma_{AC} COMFEE & +\beta_A' X_A + e_A \\ COMFEE & = \gamma_{CD} DURATION & +\gamma_{CS} SECURED & +\gamma_{CA} ALL-IN-SPREAD & +\beta_C' X_C + e_C \end{array} \quad (20)$$

where SECURED is a dichotomous [0,1] variable, COMFEE is censored from below at zero, γ_{ij} are the interdependence parameters between contract terms *i* and *j*, X_k are vectors of explanatory variables relevant to the specific contract term *k*, and e_k are the disturbances for contract term *k*.

The X_k vectors for the determinants of DURATION, SECURED status, ALL-IN-SPREAD and COMFEE together with the expected sign of the relationship are summarised in the following:

$$\begin{aligned} X_D &= \left[\begin{array}{l} \text{constant, market/book}^{\text{-ve}}, \text{unexpected earnings}^{\text{-ve}}, \text{tax/assets}^{\text{-ve}}, \text{earnings variance}^{\text{-ve}}, \\ \text{term premium}^{\text{+ve}}, \text{interest volatility}^{\text{+ve}}, \text{z-score}^{\text{+ve}}, \text{z-squared}^{\text{-ve}}, \text{leverage}^{\text{-ve}}, \\ \text{asset maturity}^{\text{+ve}}, \text{firm size}^{?}, \text{capital adequacy}^{\text{-ve}}, \text{loan purpose}^{?}, \text{deal structure}^{?} \end{array}\right] \\ X_S &= \left[\begin{array}{l} \text{constant, market/book}^{\text{+ve}}, \text{unexpected earnings}^{\text{+ve}}, \text{z-score}^{?}, \text{leverage}^{\text{+ve}}, \text{firm size}^{?}, \\ \text{loan concentration}^{\text{-ve}}, \text{loan purpose}^{?}, \text{deal structure}^{?} \end{array}\right] \\ X_A &= \left[\begin{array}{l} \text{constant, market/book}^{\text{+ve}}, \text{unexpected earnings}^{\text{+ve}}, \text{term premium}^{?}, \text{interest} \\ \text{volatility}^{\text{+ve}}, \text{z-score}^{-\text{ve}}, \text{leverage}^{\text{+ve}}, \text{libor}^{\text{+ve}}, \text{firm size}^{?}, \text{capital adequacy}^{\text{-ve}}, \text{loan} \\ \text{concentration}^{?}, \text{loan purpose}^{?}, \text{deal structure}^{?} \end{array}\right] \\ X_C &= \left[\begin{array}{l} \text{constant, market/book}^{\text{+ve}}, \text{unexpected earnings}^{\text{+ve}}, \text{z-score}^{-\text{ve}}, \text{leverage}^{\text{+ve}}, \text{capital} \\ \text{adequacy}^{\text{-ve}}, \text{loan concentration}^{?}, \text{loan purpose}^{?}, \text{deal structure}^{?} \end{array}\right] \end{aligned} \quad (21)$$

DNS estimate the model using a two-stage estimation procedure suggested by Nelson and Olson (1978). In the first stage the reduced form estimates of the model from OLS, logit and tobit estimators are used to obtain fitted values for each of the dependent variables. Then the four structural equations are estimated in the second stage by OLS, logit and tobit as appropriate and using the fitted values as instruments for the interdependent endogenous variables. The correct

more closely approximate the assumptions of the imputation model.

asymptotic covariance matrix of the structural estimates is then obtained following Amemiya (1979).

This approach was modified to take account of an endogeneity problem with the leverage variable. While leverage is not modelled in the paper, agency theory suggests that leverage, maturity and secured status are alternative mechanisms for limiting underinvestment and other agency problems in firms. To overcome this problem DNS use an instrumental variable approach and estimate a reduced form equation for leverage. Fitted values for this reduced form are then substituted for leverage in the second stage estimates of the four structural equations in the model.

The two-stage estimation procedure used by DNS is a particularly good example of a case where multiple imputation is much preferred to parameter simulation. The analysis model is rather complicated and generating thousands of parameter estimates (let alone the appropriate two-stage standard errors) will be highly impractical. Specifying an imputation model that encompasses all possible interactions of variables is, however, much easier. The sample is also of sufficient size for the asymptotic approximations to be valid.

As outlined in Section II, the missing value analysis then consists of three stages. In the imputation stage, an imputation methodology is used to generate multiple *(=m)* imputations for each missing value in the original dataset. This creates *m* completed datasets. At this stage all instruments and exogenous variables are used to generate the imputations. In the analysis stage, the model is then estimated according to the specification and two-stage estimation methodology used by DNS for each of the *m* datasets. In the final stage, the *m* sets of parameter estimates are combined according to equations (15) to (17). The results are given in Table III.

INSERT TABLE III

In a multivariate context it is difficult to evaluate the alternative sets of regression coefficients and standard errors in Table III. While there is considerable variability in the point estimates, in many cases the differences are unlikely to be statistically significant. Assuming the IP estimates are correct or exact, then visual comparisons can be made between the IP and alternative treatments of the missing data in Table III. This comparison is facilitated by an examination of the number of statistically significant parameters in each of the treatments, as summarised in Table IV. For each equation the table shows the number of statistically significant coefficients at the 90% confidence level or higher. Also shown in parentheses is the number of significant coefficients in method *i* that are also significant in the 'correct' *IP* method. This allows an evaluation of the reliability of the method in producing significant coefficients consistent with those of the IP method, given this particular model and underlying data.

For this illustrative example, the *EM-is* method produces coefficients and *t*-values in Table III very similar to those of the *IP* method. Relative to the 28 significant coefficients in the *IP* method, the *EM-is* method produces 27 significant coefficients of which 24 are common to the *IP* method. This suggests that for this particular model and data, the *EM-is* method performs well in terms of producing relatively unbiased and efficient estimates.

TABLE IV

Statistical Significance of Parameter Estimates in the Dennis, Nandy and Sharpe (1999) Model for Alternative Treatments of Missing Values

This table displays the number of significant coefficients in each equation at the 90% confidence level or higher. The number in parentheses is the number of significant coefficients that are consistent with significant coefficients obtained in the *IP* method

MISSING VALUE TREATMENT	MATURITY EQUATION	SECURED STATUS EQUATION	SPREAD EQUATION	COMMITMENT FEE EQUATION	ALL EQUATIONS
IP METHOD	5	5	8	10	28
EM-is METHOD	7 (5)	4 (4)	7 (7)	9 (8)	27 (24)
PWD METHOD	6 (4)	4 (4)	3 (3)	3 (3)	16 (14)
Ad Hoc METHOD	8 (4)	5 (3)	12 (7)	7 (6)	32 (20)
DNS METHOD	10 (5)	5 (5)	11 (8)	7 (7)	34 (25)
EM METHOD	7 (5)	5 (5)	7 (7)	9 (9)	28 (26)

On the other hand, the *PWD* and *Ad Hoc imputation* methods are less successful for this data. While the *PWD* coefficient estimates are generally similar to those of the *IP* method, suggesting unbiasedness for this data and model, only 16 coefficients are statistically significant with 14 of those common to the *IP* estimates. The problem with the *PWD* method appears to be the loss of efficiency associated with the reduction in sample size from 2634 to 896 observations, rather than any bias in the estimates. It is also suggestive of the missing data being MCAR, missing completely at random. The *Ad Hoc* method appears less reliable, producing 32 significant coefficients of which only 20 agreed with those of the *IP* method. While the *Ad Hoc* method retains

the efficiency of the full sample size, the imputed zero values for the missing observations introduces measurement error and biased coefficients.

As the *DNS* method involves a mix of *regression imputation* (a variant of *EM*) for missing secured status observations and *Ad Hoc imputation* for the commitment fee equation, the results contain features of both. It produces 34 significant coefficients, of which 25 agreed with those of the *IP* method.

Finally, it is interesting to compare the classical *EM* method, which ignores estimation uncertainty, with the *EM-is* method. With upwardly biased *t*-values (which are difficult to detect in Table 2 because of the multivariate model) the *EM* method has 28 significant coefficients (compared to 27 in *EM-is*) of which 26 are consistent with those of the *IP* method. Thus, for this data and model, there is little difference between the *EM* and *EM-is* results, although the former appears to slightly better mimic the *IP* method results in terms of significant coefficients.

It is not possible to draw general conclusions from this illustrative example because the results of the various treatments of missingness depend on the underlying nature of the missingness and the data and model used. Nevertheless, the example highlights the need for caution in handling missing data and the potential efficiency losses and biased results that can arise from use of the *listwise* (or *pairwise*) *deletion* and *ad hoc imputation* methods which are commonly used in the Finance discipline.

V. Predicting Missing Values

In order to assess the power of the imputation methodologies from a different perspective, a cross validation experiment has been designed for the empirical data set. Given that the empirical results in Table 2 indicate surprisingly little bias in the pairwise deletion estimates as compared to the 'exact' *IP*, despite the high degree of missingness, this implies that the missingness type is most likely MCAR. Given that the explanatory model is reasonably strong, it should predict with some accuracy the most likely value for a particular missing observation.

To examine this proposition, the sample was restricted to the 896 observations without missing data on either secured status and/or commitment fee. To obtain the best possible imputations, a few modifications to the DNS methodology have been applied for this prediction exercise. A close examination of the bank loan commitment data revealed that the commitment fee data was strongly spiked (and somewhat skewed) with observations concentrated at multiples of 6.25 basispoints ($1/16^{th}$). The imputation methods require the distribution of the continuous

incomplete random variables to approximate normality. Hence the commitment fee (COMFEE) variable was recoded as an ordinal variable according to the following scheme:

$0 < COMFEE \leq 6.25$	0
$6.25 < COMFEE \leq 12.5$	1
$12.5 < COMFEE \leq 25$	2
$25 < COMFEE \leq 50$	3
$50 < COMFEE$	4

This transformed the COMFEE equation into an Ordered Probit model, rather than the tobit model as in DNS. Also, an inspection of the residuals of the secured status (SECURED) equation suggested using a probit model instead of the logit model as in DNS.

After recoding, a randomly selected sample of observed dependent variables (COMFEE and SECURED) were given missing status, and the imputation methodology was applied to this set to obtain predictive posterior distributions for the missing values. The imputation model was based on the simultaneous equations model, i.e., exploiting the interactions between the dependent variables. The mode of these distributions was then compared with the true value and the proportion of correct predictions computed. This exercise was repeated a thousand times for different sets of missing observations in the dependent variables to create a Monte Carlo distribution of 1000 modal values of the proportion correctly predicted. The properties of this Monte Carlo distribution of *correct predictive proportion*s can then easily be compared with a similar Monte Carlo distribution based on naive imputation prediction schemes. Two naive imputation methods were considered for this exercise: a '*binomial*' method where each outcome has the same probability of occurrence; and a *mean imputation* method. These methods are compared with the *EM-is* methodology.

INSERT FIGURE 8

The results are illustrated in Figure 8. A further dimension to this standard experiment is added by increasing the number of 'explanatory' variables included in the missingness model from one to eight. It is clear that the *EM-is* imputation method is doing a much better job in forecasting the correct missing values than its naive alternatives. Whereas the probit variable (SECURED) scores a success rate of about 77% (against 50% naive), the ordered probit variable (COMFEE) scores about 50% (against 20% naive). Interestingly, mean imputation performs just as well as *EM-is* for the probit (SECURED) equation, but performs significantly worse for the ordered probit (COMFEE) equation. As expected the distribution is slightly narrowing (becoming more precise) for increasing numbers of explanatory variables in the missingness model. However, it does not seem to give

biased proportions when omitting explanatory variables (i.e., including too few variables in the missingness model). This may be an artifact of this particular application.

VI. Conclusion

Missing data problems should be taken seriously in many financial applications. Dropping observations with missing information is at best inefficient and at worst influences inference. It would therefore be advisable if journal editors required authors to supply explicit data information with regard to missing data problems and treatment. The approach advocated in this paper is based on missing data imputation. New, vastly improved, imputation methods have recently become available. Their application has so far been restricted to the statistics literature. This is lamentable given the scope for application in the finance (and economics) literature. Even in cases where the missing values are missing completely at random (MCAR), more efficiency can be gained by performing 'complete-data' analysis. Standard errors of parameter estimates will generally be smaller. When the missing values can be related to the (complete) explanatory variables (i.e., the MAR scheme), formal imputation methods become imperative. The discussed imputation methods are generally fast (though the Bayesian methods still have a distinct disadvantage) and are intuitively appealing. They are however not magic and will fail in the case of non-ignorable missingness (NI). Recent papers by Rotnitsky et al. (1998) and Horowitz and Manski (1998) deal with non-ignorable (non-randomly) missing data. Horowitz and Manski derive bounds on the parameters for the case where no assumptions are made with regard to the inherently untestable missingness model. Inevitably, the confidence intervals will be larger (sometimes unacceptably so) than when the MAR or MCAR assumption is taken on face value. Interestingly, whereas NI has a severe impact on single equation parameters, the results in this paper indicate that its impact may be much less severe for simultaneous equations. The intricate feedback obtained by the simultaneity expands the available information that can be used for imputing missing values. Ultimately, whether one can reasonably assume ignorability of the missingness model depends on the specific application. Careful analysis of the potential reasons why certain values are missing in a particular dataset might indicate whether NI is a likely cause for missingness. In the case of the example used in this study, the DNS model, the ignorability assumption seems reasonable.

Careful application of the missing data imputation methods discussed in this paper opens up many opportunities for otherwise complicated data analysis. In many cases, more reliable parameter estimates with smaller standard errors can be achieved. These methods might even allow researchers to revisit 'old' issues through analysis on previously discarded datasets.

REFERENCES

Amemiya, Takeshi, 1979, "The Estimation of a Simultaneous Equation Tobit Model," *International Economic Review*, 20, 169-181.

Barclay, Michael J., and Clifford W. Smith Jr., 1995a, "The Maturity Structure of Corporate Debt, *Journal of Finance*, 50, 609-631.

_____, 1995b, "The Priority Structure of Corporate Liabilities, *Journal of Finance*, 50, 899-917.

Berger, A., and G. Udell, 1990, "Collateral, Loan Quality and Bank Risk," *Journal of Monetary Economics*, 25, 21-42.

Dennis, Steven, Debarshi Nandy, and Ian Sharpe, 1999, "The Determinants of Contract Terms in Bank Revolving Credit Agreements," *Journal of Financial and Quantitative Analysis*, forthcoming.

Dennis, Steven, and Ian Sharpe, 1998, "The Structure of Intermediated Term Debt: An Investigation of Middle Market and Large Corporate Lending," Working Paper, Department of Finance, California State University, Fullerton.

Dempster, A.P., Nan M. Laird, and Donald B. Rubin, 1977, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society* 39B, 1-22.

Guedes, Jose, and Tim Opler, 1996, "The Determinants of the Maturity of New Corporate Debt Issues," *Journal of Finance* 51, 1809-1833.

Honaker, James, Anne Joseph, Gary King, and Kenneth Scheve, 1999, *Amelia: A Program for Missing Data* (Gauss version), Cambridge, MA: Harvard University, http://Gking.Harvard.edu/.

Horowitz, Joel L., and Charles F. Manski, "Non-Parametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data," working paper.

King, Gary, James Honaker, Anne Joseph, and Kenneth Scheve, 1998, "Listwise Deletion is Evil: What to Do About Missing Data in Political Science," Paper presented at the Annual Meetings of the American Political Science Association, Boston.

Little, Rodrick J.A., and Donald B. Rubin, 1987, *Statistical Analysis with Missing Data*, John Wiley & Sons, New York.

Malhotra, Naresh K., 1987, "Analyzing Market Research Data with Incomplete Information on the Dependent Variable," *Journal of Marketing Research* 24, 74-84.

Melnik, A., and S. Plaut, 1986, "Loan Commitment Contracts, Terms of Lending, and Credit Allocation," *Journal of Finance* 41, 425-435.

Nelson, Forrest, and Lawrence Olson, 1978, "Specification and Estimation of a Simultaneous-Equation Model with Limited Dependent Variables," *International Economic Review* 19, 695-709.

Peterson, M., and R. Rajan, 1994, "The Benefits of Lending Relationships: Evidence from Small Business Data," *Journal of Finance* 49, 3-37.

Pulvino, Todd C., 1998, "Do Asset Fire Sales Exist? An Empirical Investigation of Commercial Aircraft Transactions," *Journal of Finance* 53, 939-978.

Rubin, Donald B., 1976, "Inference and Missing Data," *Biometrika* 63, 581-592.

_____, 1978, "Multiple Imputations in Sample Surveys," *ASA 1978 proceedings of the Survey Research Methods Section*, 20-34.

_____, 1987, *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, New York.

_____, 1996, "Multiple Imputation after 18 Years," *Journal of the American Statistical Association* 89, 475-478.

Saunders, Anthony, 1996, "Credit Spreads in the Market for Highly Leveraged Transactions Loans," Working Paper, New York University.

Schafer, Joseph L., 1997, *Analysis of Incomplete Multivariate Data*, Chapman & Hall, London.

Schafer, Joseph L., and Maren K. Olsen, 1998, "Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective," *Multivariate Behavioral Research*, forthcoming.

Stohs, M.H., and David C. Mauer, 1996, "The Determinants of Corporate Debt Maturity Structure," *Journal of Business* 69, 279-312.

Tanner, Martin A., and W.H. Wong, 1987, "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association* 82, 528-550.

Tanner, Martin A., 1993, *Tools for Statistical Inference: Methods for the Exploration of Posterior Distribution and Likelihood Functions*, 2nd edition, Springer-Verlag, New York.

Table III. Simultaneous Equation Imputation Model

Panels A through D give the parameter estimates for the DNS model based on different treatment of missing values in the Secured Status and Commitment Fee variables. *DNS* is the benchmark method used by Dennis, Nandy and Sharpe (1999), which is a combination of ad hoc imputation and regression imputation. *PWD* is the pairwise deletion method *AdHoc* is the ad hoc imputation method where all missing values are assumed to be zero. *IP* is the Bayesian imputation-posterior method. *EM* is the Expectation/Maximization imputation method. *EM-is* is the EM algorithm extended by importance sampling. *PWD* estimates are based on 896 observations. All other methods' estimates are based on 2634 observations. π gives the fraction of missing information with regard to *EM-is* parameters.

Table III.A. Maturity Equation

The explanatory variables consist of instruments for: Secured Status (ISECD), Leverage (ILEVG), and exogenous variables: Constant (CONS), Market-to-Book ratio (MKBK), Unexpected Earnings (ABNL), Interest Volatility (SDGB), Tax to Assets ratio (TAXA), Altman's Z-score (ZSCR), Z-score squared (ZSQD), Asset Maturity (AMAT), Earnings Variance (VAR), Firm Size (FSZE), Term Premium (TERM), a post-1993 dummy (POST93), Other Purpose (PMSC).

	PARAMETER ESTIMATES						
	DNS	*PWD*	*AdHoc*	*IP*	*EM*	*EM-is*	
ISECD	0.893	1.034	0.623	1.019	1.024	1.011	
ILEVG	-3.905	-5.250	-1.388	-2.084	-2.300	-2.438	
CONS	-1.519	-1.685	0.766	-1.627	-1.267	-1.258	
MKBK	-0.192	-0.254	-0.188	-0.110	-0.124	-0.134	
ABNL	-0.178	-0.105	-0.108	-0.114	-0.121	-0.167	
SDGB	0.969	0.423	0.677	0.534	0.598	0.776	
TAXA	-3.545	0.630	-2.324	-2.473	-2.473	-1.899	
ZSCR	0.576	0.569	0.290	0.502	0.500	0.467	
ZSQD	-0.090	-0.099	-0.039	-0.066	-0.076	-0.067	
AMAT	0.080	0.066	0.073	0.095	0.039	0.082	
VAR	-6.227	-6.910	-3.083	-4.407	-5.674	-5.437	
FSZE	0.595	0.750	0.463	0.586	0.567	0.553	
TERM	0.178	0.120	-0.061	0.110	0.098	0.098	
POST93	0.408	0.282	0.066	0.342	0.359	0.314	
PMSC	-0.346	-0.068	-0.154	-0.375	-0.352	-0.327	
					t-VALUES		π
							EM-is
ISECD	4.99	4.49	6.06	4.19	5.70	5.47	0.20
ILEVG	-2.97	-3.17	-1.97	-2.17	-2.40	-1.66	1.48
CONS	-1.47	-1.12	1.61	-1.32	-1.51	-1.30	0.43
MKBK	-1.91	-1.67	-2.81	-1.24	-1.29	-1.49	0.11
ABNL	-0.90	-0.46	-1.01	-0.70	-0.65	-0.92	0.15
SDGB	1.99	0.60	2.46	1.37	1.55	1.49	1.12
TAXA	-1.13	0.14	-1.21	-0.86	-0.87	-0.66	0.35
ZSCR	2.26	1.57	2.32	1.49	2.10	1.99	0.68
ZSQD	-2.20	-1.75	2.11	-1.25	-2.07	-1.78	0.84
AMAT	0.81	0.45	1.24	1.00	0.43	0.81	0.65
VAR	-2.85	-2.22	-2.82	-1.90	-2.91	-2.17	1.18
FSZE	6.45	6.75	8.41	5.86	7.30	6.12	0.54
TERM	1.90	0.85	-1.23	1.38	1.19	1.39	0.10
POST93	2.34	1.14	0.62	1.87	2.47	1.91	0.41
PMSC	-0.95	-0.12	-0.72	-1.18	-1.34	-1.02	0.56

Table III.B. Secured Status Equation

The explanatory variables consist of instruments for: Maturity (IMATY), Leverage (ILEVG), and exogenous variables: Constant (CONS), Market-to-Book ratio (MKBK), Unexpected Earnings (ABNL), Altman's Z-score (ZSCR), Firm Size (FSZE), Loan Concentration (LRSZ).

	PARAMETER ESTIMATES					
	DNS	*PWD*	*AdHoc*	*IP*	*EM*	*EM-is*
IMATY	0.536	0.412	0.129	0.445	0.365	0.265
ILEVG	6.694	7.584	5.316	4.136	4.820	5.557
CONS	1.387	1.130	0.536	1.601	1.314	1.263
MKBK	0.386	0.334	0.381	0.246	0.277	0.286
ABNL	0.212	0.149	0.171	0.120	0.131	0.168
ZSCR	-0.051	-0.072	-0.141	-0.097	-0.052	-0.066
FSZE	-0.649	-0.524	-0.445	-0.536	-0.492	-0.439
LRSZ	0.232	0.459	0.504	0.191	0.244	0.358

					t-VALUES		π
							EM-Is
IMATY	2.28	1.34	0.89	2.05	1.78	1.63	0.20
ILEVG	3.19	3.83	4.21	2.58	2.49	3.92	0.19
CONS	2.29	1.77	1.49	3.81	2.56	2.62	0.59
MKBK	3.36	2.47	5.04	2.97	2.47	3.37	0.21
ABNL	1.04	0.73	1.51	0.80	0.76	1.07	0.14
ZSCR	-0.59	-0.75	-3.01	-1.46	-0.84	-1.10	0.43
FSZE	-5.60	-3.21	-6.23	-5.91	-4.68	-5.00	0.51
LRSZ	0.81	1.32	2.89	0.88	0.95	1.63	0.58

Table III.C. Spread Equation

The explanatory variables consist of instruments for: Commitment fee (ICOMD), Maturity (IMATY), Secured Status (ISECD), Leverage (ILEVG), and exogenous variables: Constant (CONS), Market-to-Book ratio (MKBK), Unexpected Earnings (ABNL), Interest Volatility (SDGB), Altman's Z-score (ZSCR), Loan Concentration (LRSZ), London Interbank Offered Rate (LIBOR), Term Premium (TERM), Repayment/Recap (PNRM), Acquisitions (PACQ), Term Loan (TMLN), Syndicated loan (SYND).

	PARAMETER ESTIMATES					
	DNS	*PWD*	*AdHoc*	*IP*	*EM*	*EM-is*
ICOMD	1.677	10.691	1.645	4.977	5.246	6.197
IMATY	-0.675	-0.503	-0.614	-0.466	-0.455	-0.419
ISECD	0.166	-0.233	0.258	0.085	0.062	0.043
ILEVG	2.203	-0.910	1.981	0.862	0.903	-0.126
CONS	2.165	0.307	2.087	0.892	0.921	0.727
MKBK	0.091	0.007	0.071	0.068	0.062	0.010
ABNL	0.015	-0.169	0.022	-0.051	-0.054	-0.066
SDGB	0.734	0.417	0.727	0.337	0.288	0.240
ZSCR	-0.046	-0.065	-0.025	-0.007	0.001	0.000
LRSZ	0.334	0.040	0.252	0.152	0.142	0.053
LIBOR	0.049	0.031	0.085	0.023	0.009	0.013
TERM	0.160	-0.074	0.165	0.062	0.029	0.023
PNRM	0.204	0.214	0.147	0.166	0.154	0.169
PACQ	0.499	0.370	0.349	0.348	0.323	0.289
TMLN	-0.170	-0.321	-0.216	-0.150	-0.145	-0.133
SYND	-0.312	-0.249	-0.249	-0.249	-0.246	-0.209

					t-VALUES		π
							EM-is
ICOMD	2.12	1.82	2.08	3.02	3.34	3.37	0.70
IMATY	-6.82	-2.38	-5.28	-4.95	-4.42	-4.78	0.16
ISECD	2.48	-0.78	2.52	0.76	0.64	0.35	0.99
ILEVG	2.08	-0.29	1.85	0.84	0.79	-0.09	0.70
CONS	5.40	0.22	5.11	2.21	2.11	2.02	0.16
MKBK	1.47	0.04	1.16	1.38	1.06	0.13	0.95
ABNL	0.24	-1.22	0.36	-0.97	-0.89	-1.26	0.15
SDGB	4.71	0.85	4.71	2.45	1.65	1.13	0.60
ZSCR	-1.52	-0.70	-0.79	-0.28	0.04	0.01	0.70
LRSZ	3.26	0.13	2.18	1.55	1.21	0.42	0.57
LIBOR	1.57	0.48	3.03	0.88	0.26	0.48	0.32
TERM	2.50	-0.37	2.59	1.01	0.38	0.31	0.67
PNRM	1.93	0.85	1.28	1.74	1.58	1.76	0.48
PACQ	3.74	1.13	2.08	2.91	2.53	2.15	0.56
TMLN	-2.34	-1.88	-2.99	-2.82	-2.32	-2.37	0.32
SYND	-2.90	-0.87	-2.36	-2.70	-2.60	-1.88	0.68

Table III.D. Commitment Fee Equation

The explanatory variables consist of instruments for: Spread (ISPRD), Maturity (IMATY), Secured Status (ISECD), Leverage (ILEVG), and exogenous variables: Constant (CONS), Market-to-Book ratio (MKBK), Unexpected Earnings (ABNL), Altman's Z-score (ZSCR), Loan Concentration (LRSZ), a post-1993 dummy (POST93), Other Purpose (PMSC), Multiple Revolvers (MULT), Term Loan (TMLN), Syndicated loan (SYND).

	PARAMETER ESTIMATES					
	DNS	*PWD*	*AdHoc*	*IP*	*EM*	*EM-is*
ISPRD	0.063	0.088	-0.003	0.111	0.120	0.112
IMATY	0.020	0.019	0.009	0.015	0.019	0.017
ISECD	-0.004	0.002	0.046	-0.008	-0.010	-0.006
ILEVG	0.615	0.256	0.514	0.247	0.186	0.251
CONS	-0.133	0.054	0.055	0.061	0.049	0.048
MKBK	0.023	0.004	0.018	0.002	0.000	0.005
ABNL	0.014	0.020	0.014	0.017	0.016	0.014
ZSCR	-0.006	-0.002	-0.005	-0.005	-0.006	-0.005
LRSZ	0.055	0.023	0.037	0.024	0.020	0.023
POST93	-0.041	-0.003	-0.046	-0.016	-0.020	-0.011
PMSC	-0.070	0.050	-0.085	0.032	0.031	0.027
MULT	-0.060	-0.013	-0.056	-0.021	-0.022	-0.018
TMLN	0.044	0.022	0.028	0.020	0.021	0.017
SYND	0.073	0.035	0.060	0.039	0.042	0.034

					t-VALUES		π
							EM-is
ISPRD	2.01	3.50	-0.09	7.19	6.13	6.95	0.46
IMATY	1.03	1.49	0.55	1.74	1.75	1.77	0.33
ISECD	-0.25	0.15	2.52	-0.66	-0.70	-0.57	0.62
ILEVG	2.22	1.37	2.45	1.98	1.26	2.11	0.23
CONS	-1.47	0.67	0.55	1.63	0.93	1.03	0.66
MKBK	1.54	0.40	1.54	0.25	0.03	0.77	0.33
ABNL	0.95	2.32	1.13	2.40	1.89	2.17	0.11
ZSCR	-0.87	-0.36	-0.92	-1.65	-1.69	-1.19	1.14
LRSZ	1.82	1.06	1.55	1.55	1.23	1.66	0.31
POST93	-2.45	-0.27	-3.64	-1.66	-2.21	-1.16	0.96
PMSC	-2.19	1.52	-3.17	2.26	1.79	1.75	0.51
MULT	-2.96	-0.86	-3.55	-2.23	-1.80	-2.08	0.13
TMLN	2.64	1.53	1.87	2.14	2.10	2.16	0.34
SYND	3.03	1.69	3.00	3.34	2.97	2.26	1.20

MONTE CARLO EXPERIMENTS

Figure 1. Single Equation Bias and *MSE* - MCAR

The top two panels display the mean squared error (*mse*) for the mean value of the distribution for a discrete (0/1) dependent variable in a single equation. The bottom two panels display the bias for the mean value of this distribution. The panels on the left are based on *listwise deletion*. The panels on the right are based on the *EM-is* imputation methodology. The missingness scheme is *Missing Completely at Random* (MCAR)

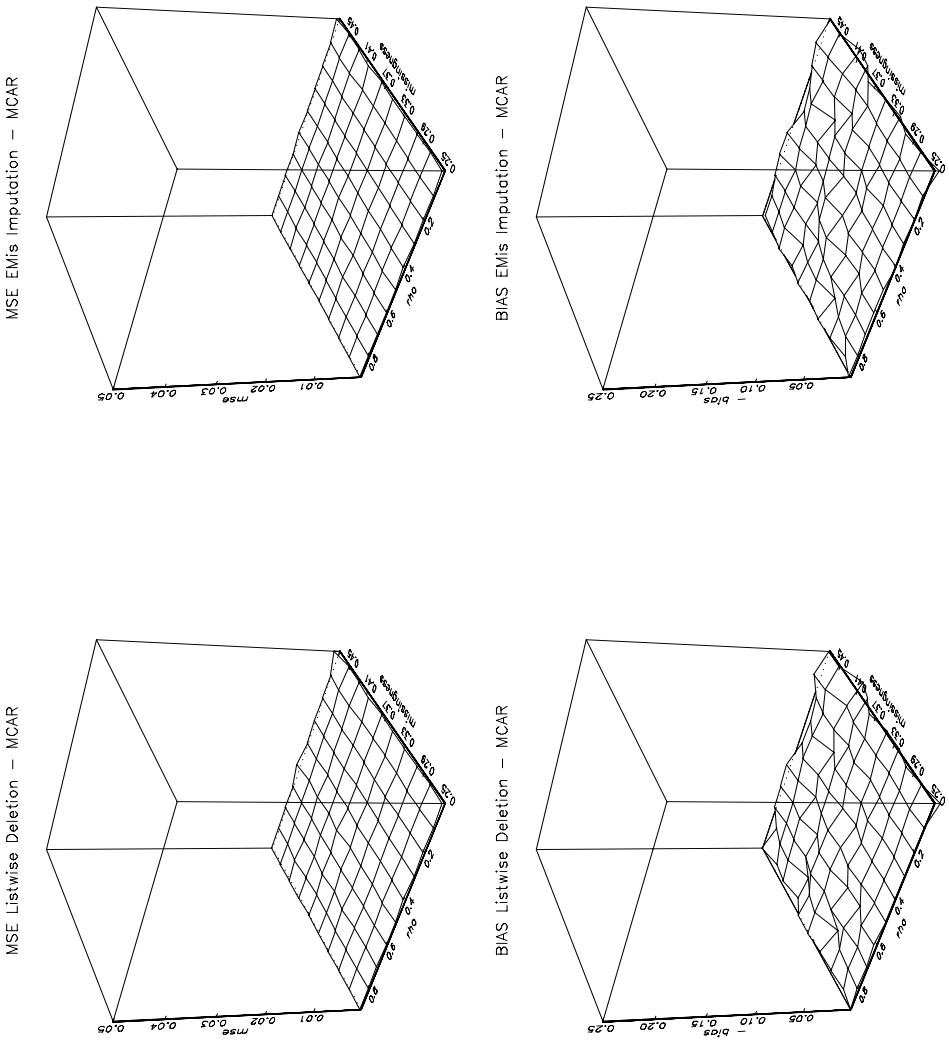

Figure 1a. Single Equation Bias and *MSE* - MCAR

The top two panels display the mean squared error (*mse*) for the mean value of the distribution for a continuous dependent variable in a single equation. The bottom two panels display the bias for the mean value of this distribution. The panels on the left are based on *listwise deletion*. The panels on the right are based on the *EM-is* imputation methodology. The missingness scheme is *Missing Completely at Random* (MCAR)

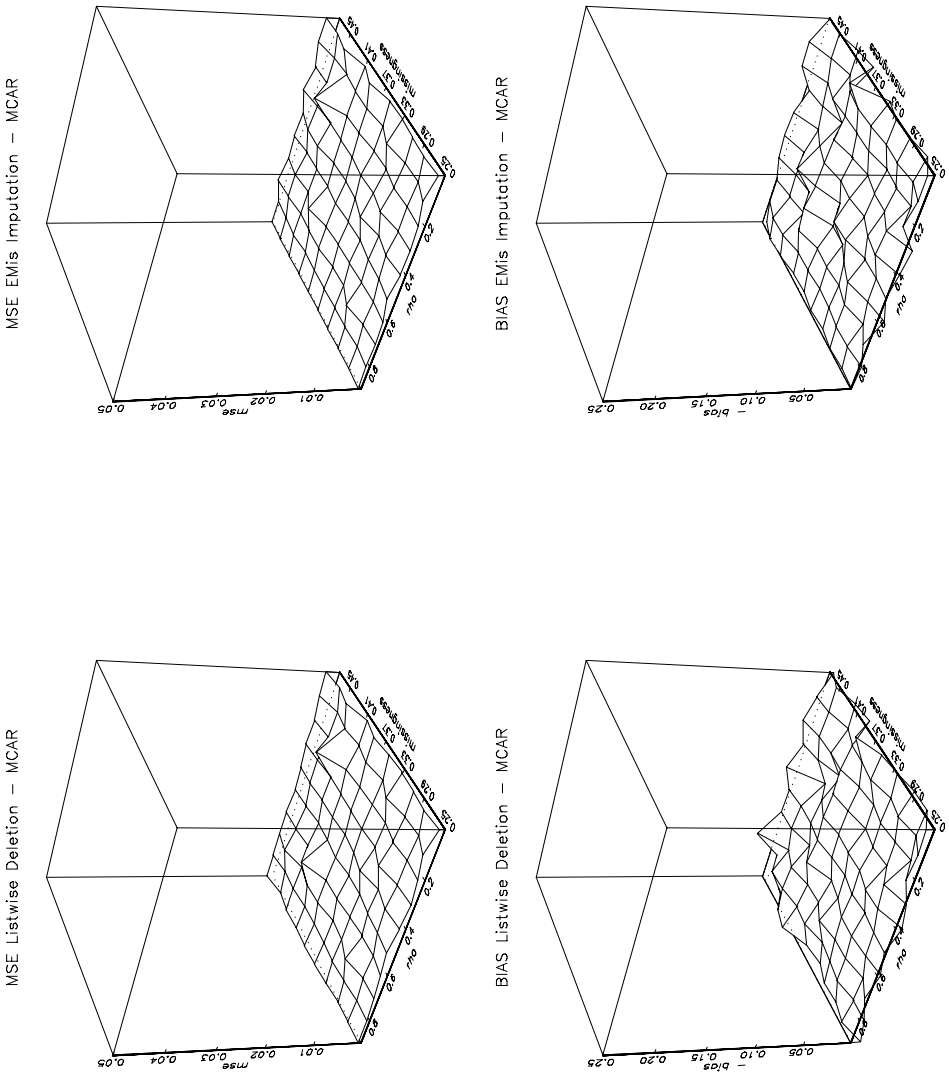

Figure 2. Single Equation Bias and *MSE* - MAR

The top two panels display the mean squared error (*mse*) for the mean value of the distribution for a discrete (0/1) dependent variable in a single equation. The bottom two panels display the bias for the mean value of this distribution. The panels on the left are based on *listwise deletion*. The panels on the right are based on the *EM-is* imputation methodology. The missingness scheme is *Missing at Random* (MAR).

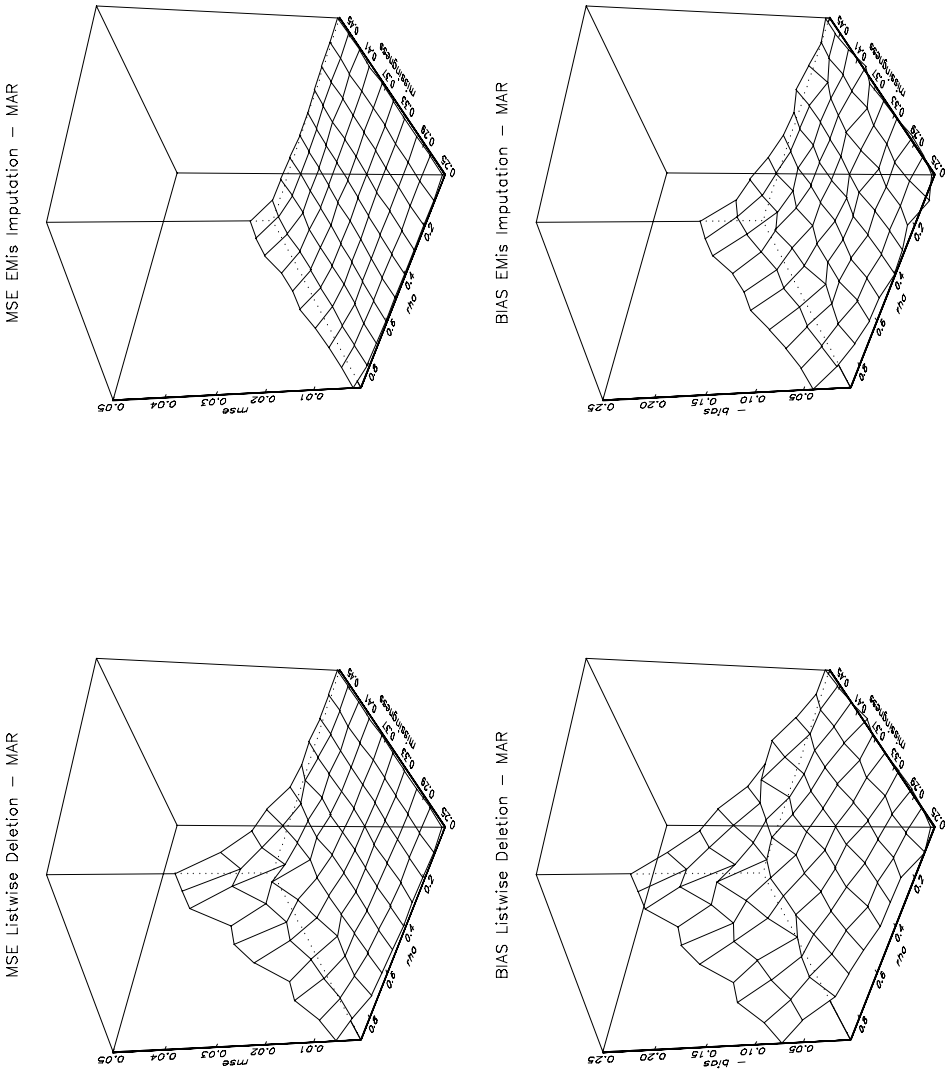

Figure 2a. Single Equation Bias and *MSE* – MAR

The top two panels display the mean squared error (*mse*) for the mean value of the distribution for a continuous dependent variable in a single equation. The bottom two panels display the bias for the mean value of this distribution. The panels on the left are based on *listwise deletion*. The panels on the right are based on the *EM-is* imputation methodology. The missingness scheme is *Missing at Random* (MAR).

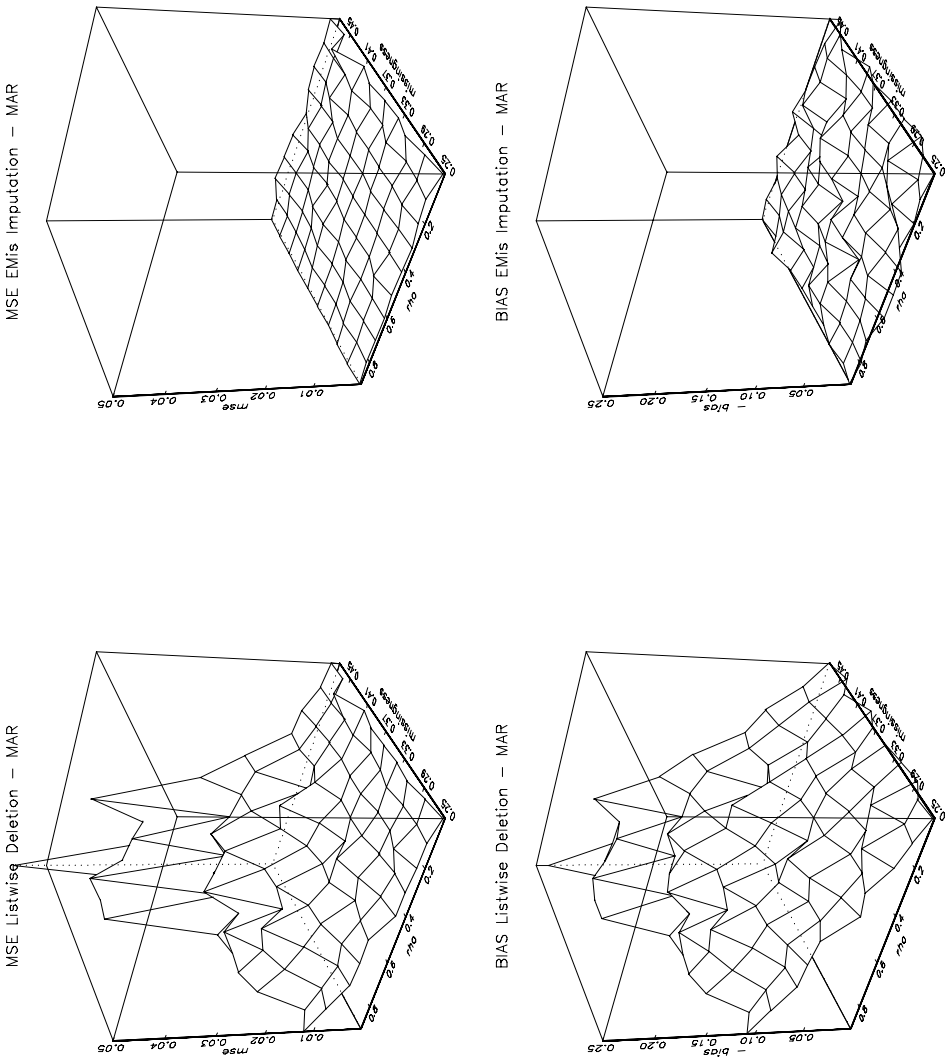

Figure 3. Single Equation Bias and *MSE* – NI

The top two panels display the mean squared error (*mse*) for the mean value of the distribution for a discrete (0/1) dependent variable in a single equation. The bottom two panels display the bias for the mean value of this distribution. The panels on the left are based on *listwise deletion*. The panels on the right are based on the *EM-is* imputation methodology. The missingness scheme is *Non-Ignorable* (NI).

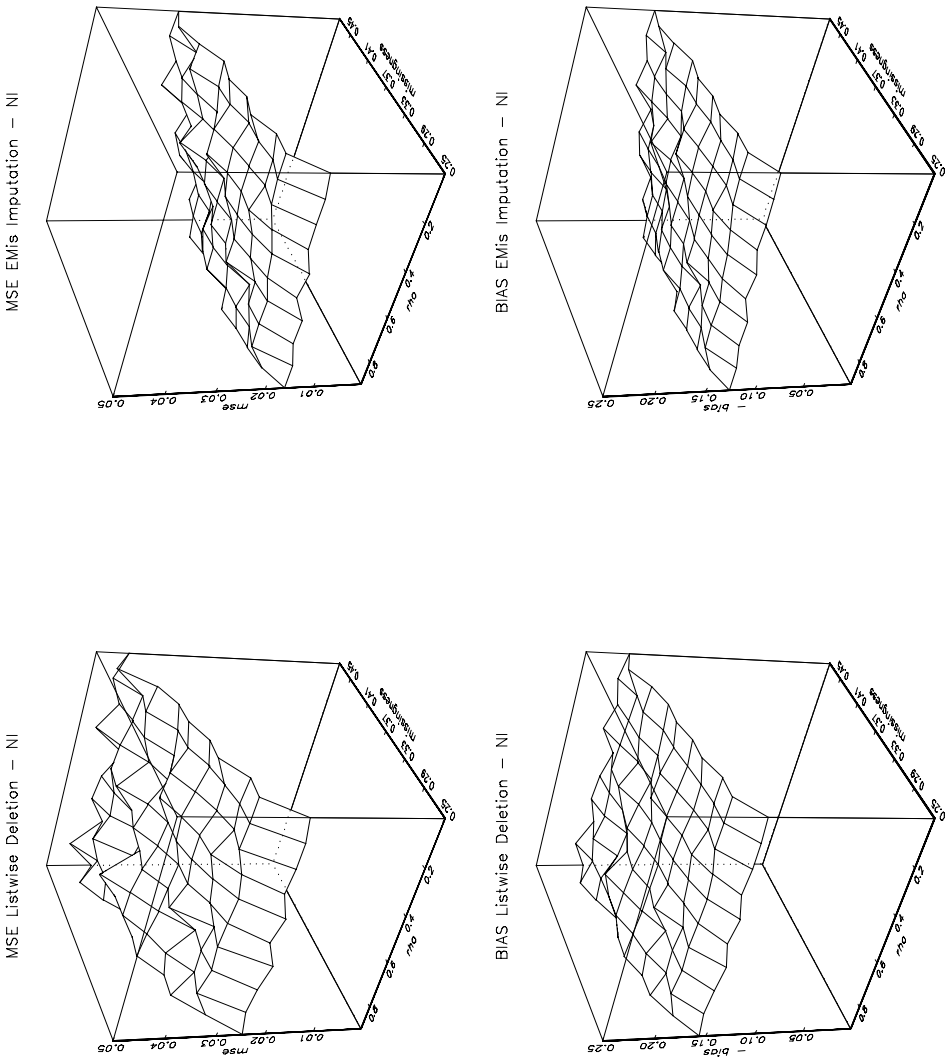

Figure 3a. Single Equation Bias and *MSE* - NI

The top two panels display the mean squared error (*mse*) for the mean value of the distribution for a continuous dependent variable in a single equation. The bottom two panels display the bias for the mean value of this distribution. The panels on the left are based on *listwise deletion*. The panels on the right are based on the *EM-is* imputation methodology. The missingness scheme is *Non-Ignorable* (NI).

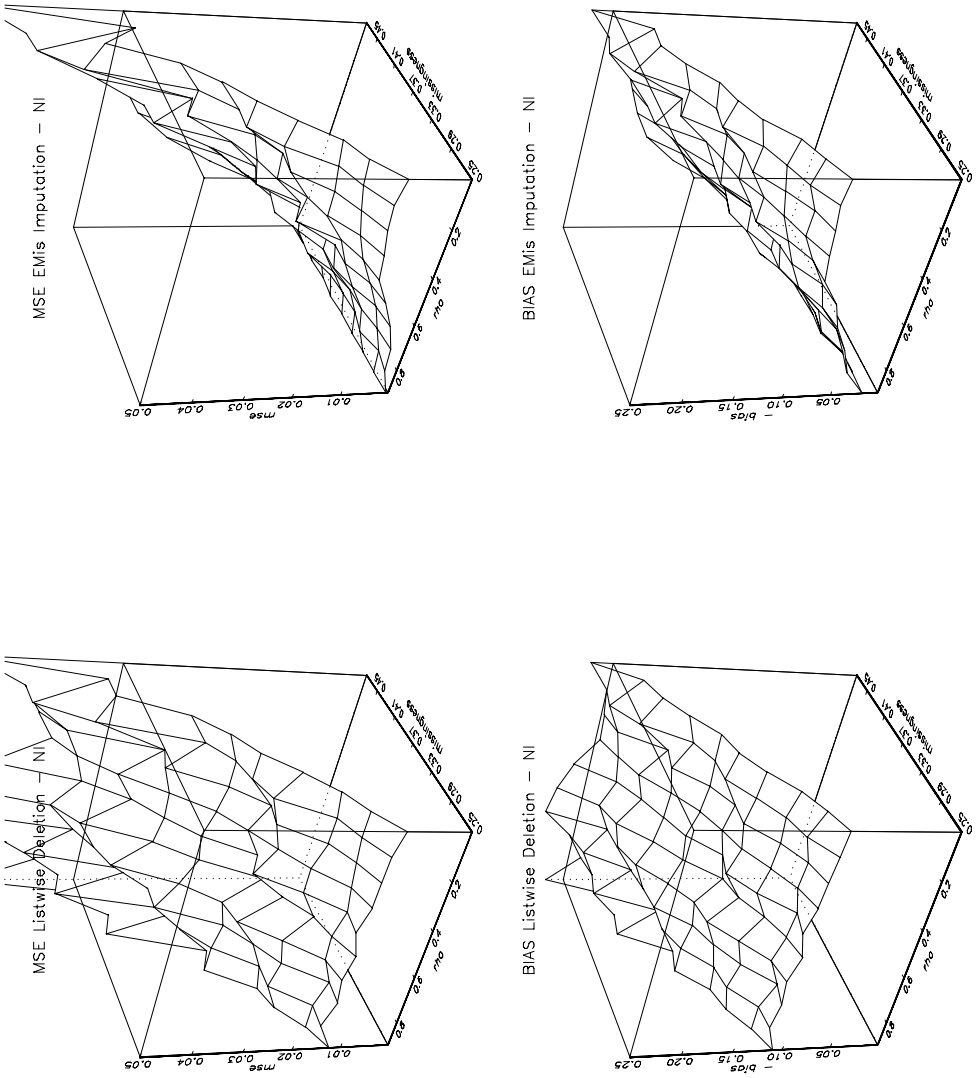

Figure 4. Simultaneous Equation Bias and *MSE* - MCAR

The top two panels display the mean squared error (*mse*) for the mean value of the distribution for a discrete (0/1) dependent variable in a simultaneous equations system where the other dependent variable is continuous. The bottom two panels display the bias for the mean value of this distribution. The panels on the left are based on *pairwise deletion*. The panels on the right are based on the *EM-is* imputation methodology. The missingness scheme is *Missing Completely at Random* (MCAR). Both dependent variables have missing values.

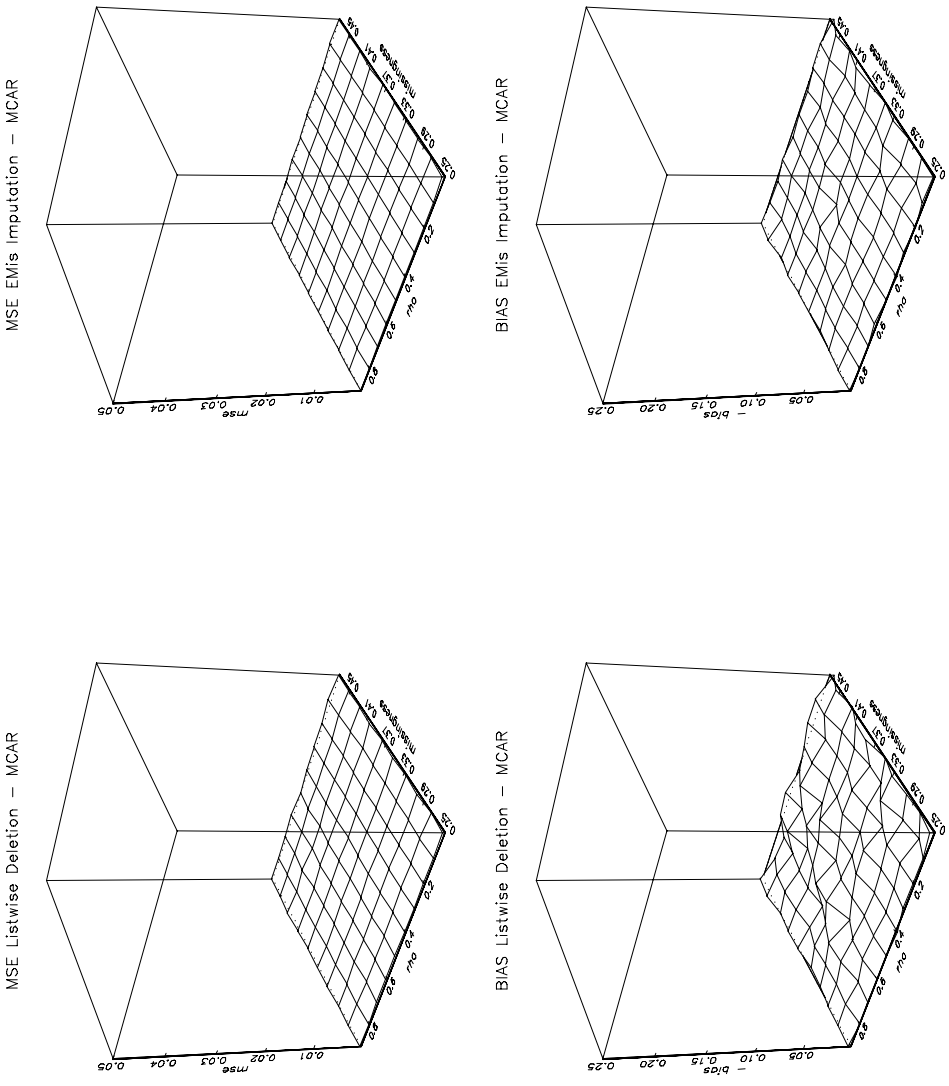

Figure 4a. Simultaneous Equation Bias and *MSE* - MCAR

The top two panels display the mean squared error (*mse*) for the mean value of the distribution for a continuous dependent variable in a simultaneous equations system where the other dependent variable is also continuous. The bottom two panels display the bias for the mean value of this distribution. The panels on the left are based on *pairwise deletion*. The panels on the right are based on the *EM-is* imputation methodology. The missingness scheme is *Missing Completely at Random* (MCAR). Both dependent variables have missing values.

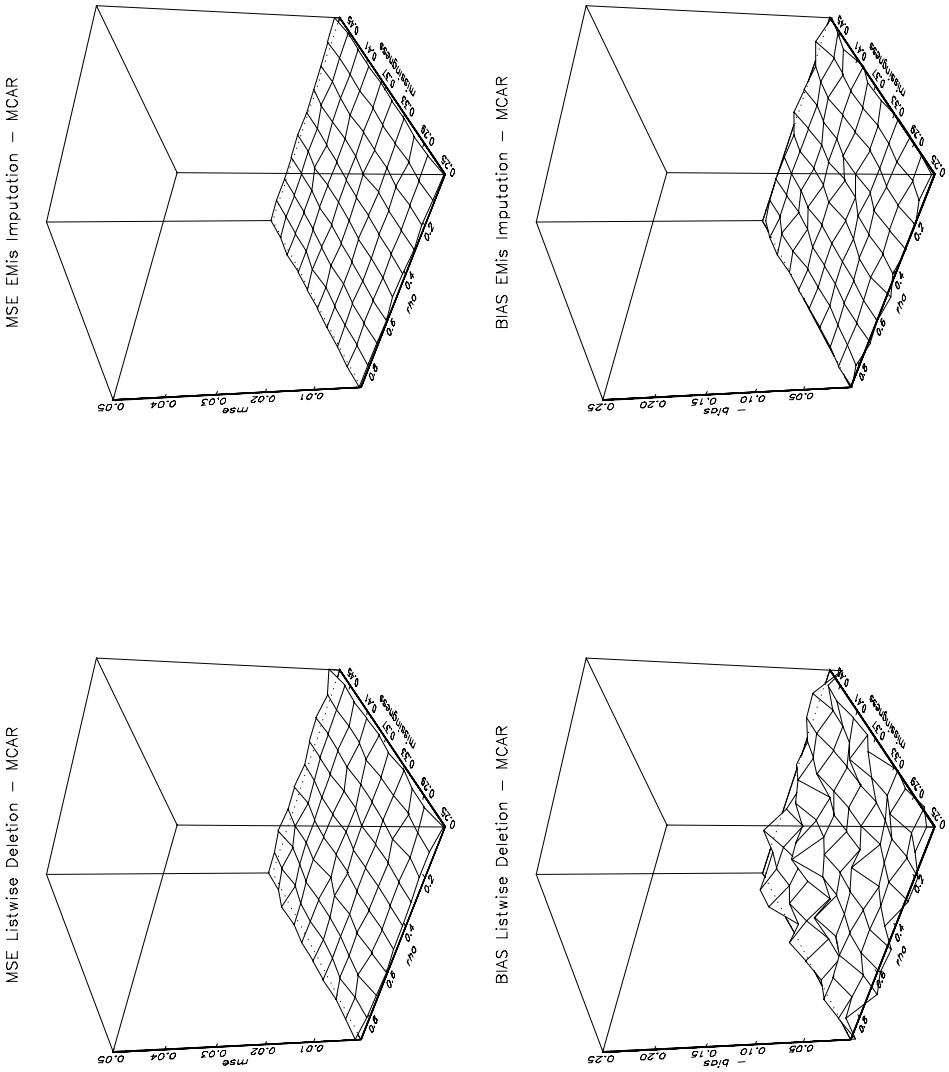

Figure 5. Simultaneous Equation Bias and *MSE* - MAR

The top two panels display the mean squared error (*mse*) for the mean value of the distribution for a discrete (0/1) dependent variable in a simultaneous equations system where the other dependent variable is continuous. The bottom two panels display the bias for the mean value of this distribution. The panels on the left are based on *pairwise deletion*. The panels on the right are based on the *EM-is* imputation methodology. The missingness scheme is *Missing at Random* (MAR). Both dependent variables have missing values.

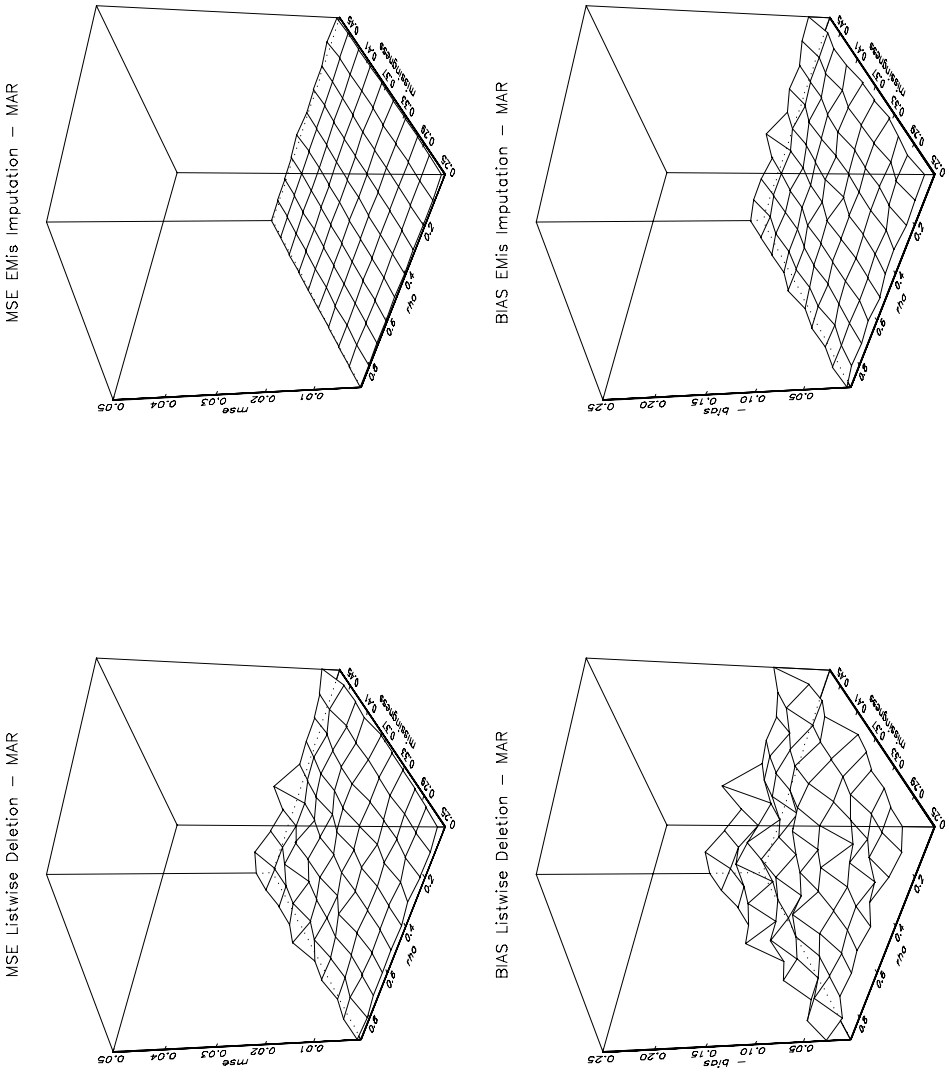

Figure 5a. Simultaneous Equation Bias and *MSE* - MAR

The top two panels display the mean squared error (*mse*) for the mean value of the distribution for a continuous dependent variable in a simultaneous equations system where the other dependent variable is also continuous. The bottom two panels display the bias for the mean value of this distribution. The panels on the left are based on *pairwise deletion*. The panels on the right are based on the *EM-is* imputation methodology. The missingness scheme is *Missing at Random* (MAR). Both dependent variables have missing values.

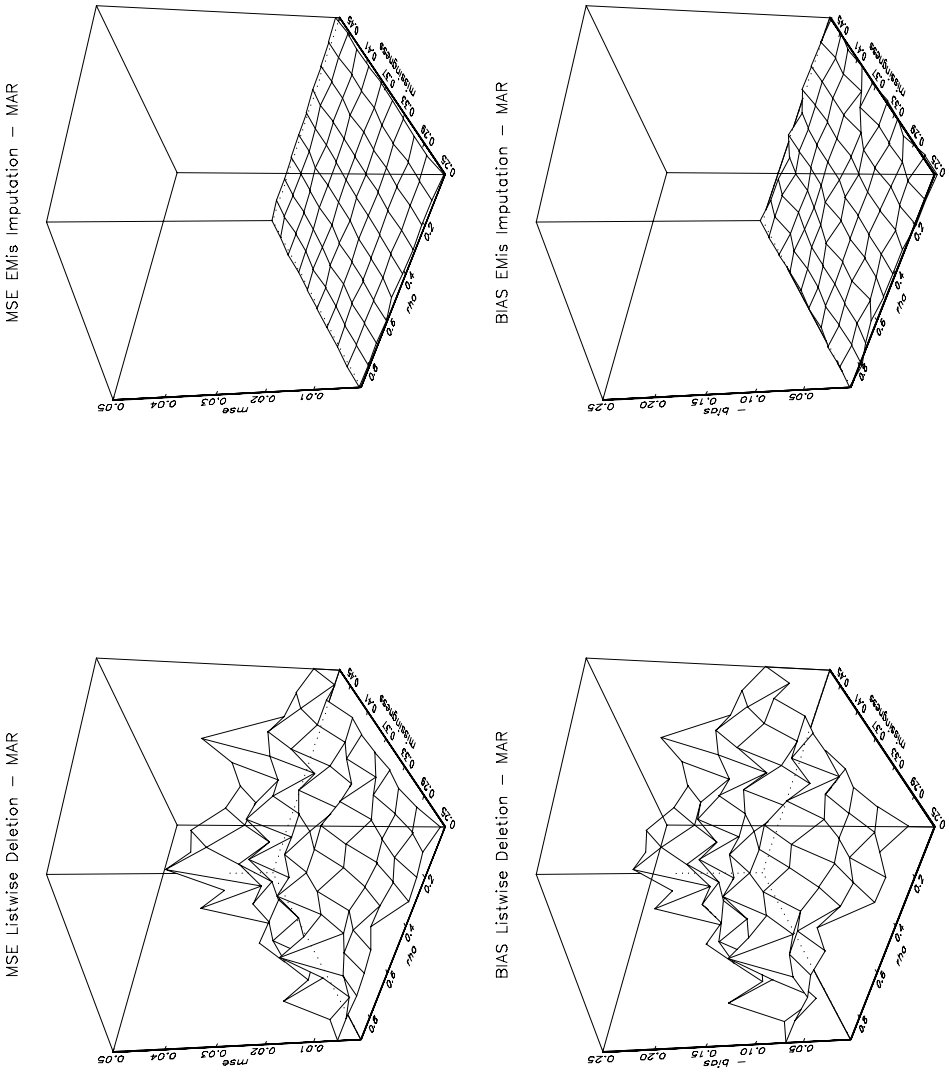

Figure 6. Simultaneous Equation Bias and *MSE* - NI

The top two panels display the mean squared error (*mse*) for the mean value of the distribution for a discrete (0/1) dependent variable in a simultaneous equations system where the other dependent variable is continuous. The bottom two panels display the bias for the mean value of this distribution. The panels on the left are based on *pairwise deletion*. The panels on the right are based on the *EM-is* imputation methodology. The missingness scheme is *Non-Ignorable* (NI). Both dependent variables have missing values.

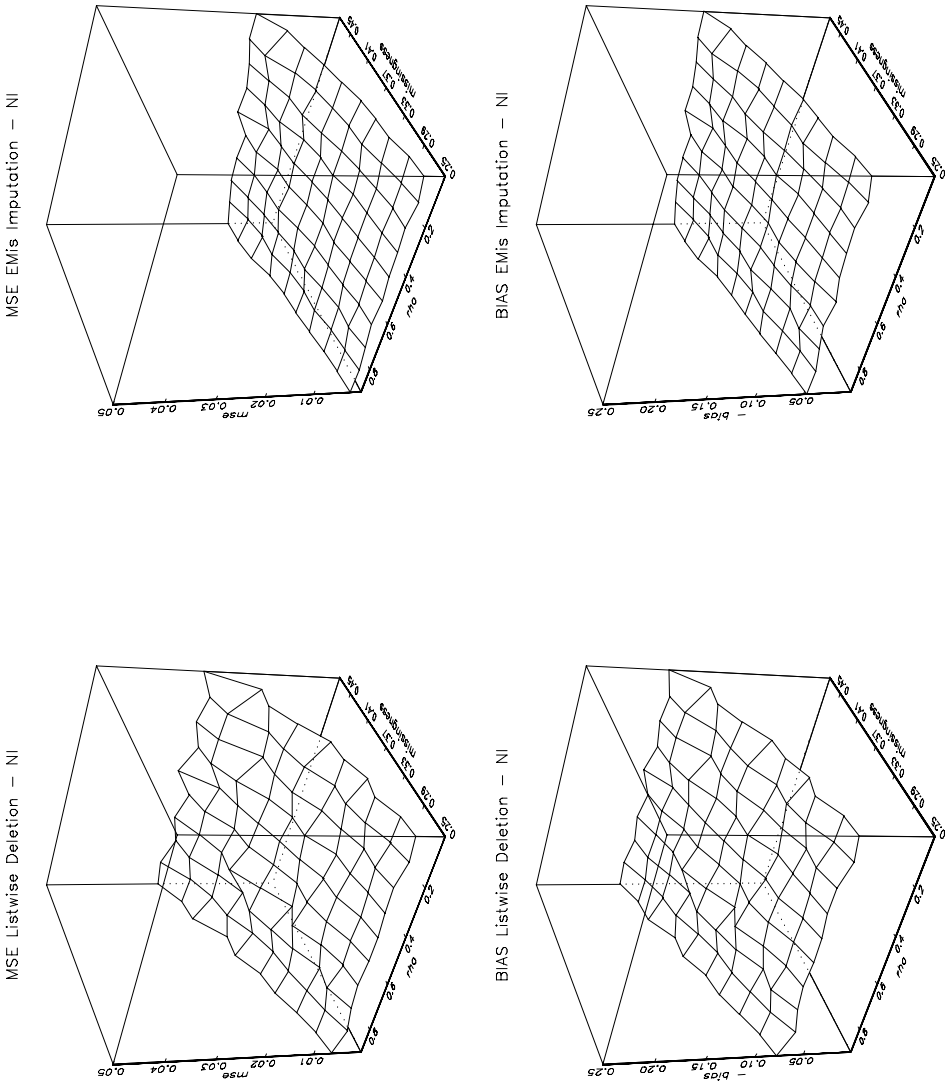

Figure 6a. Simultaneous Equation Bias and *MSE* - NI

The top two panels display the mean squared error (*mse*) for the mean value of the distribution for a continuous dependent variable in a simultaneous equations system where the other dependent variable is also continuous. The bottom two panels display the bias for the mean value of this distribution. The panels on the left are based on *pairwise deletion*. The panels on the right are based on the *EM-is* imputation methodology. The missingness scheme is *Non-Ignorable* (NI). Both dependent variables have missing values.

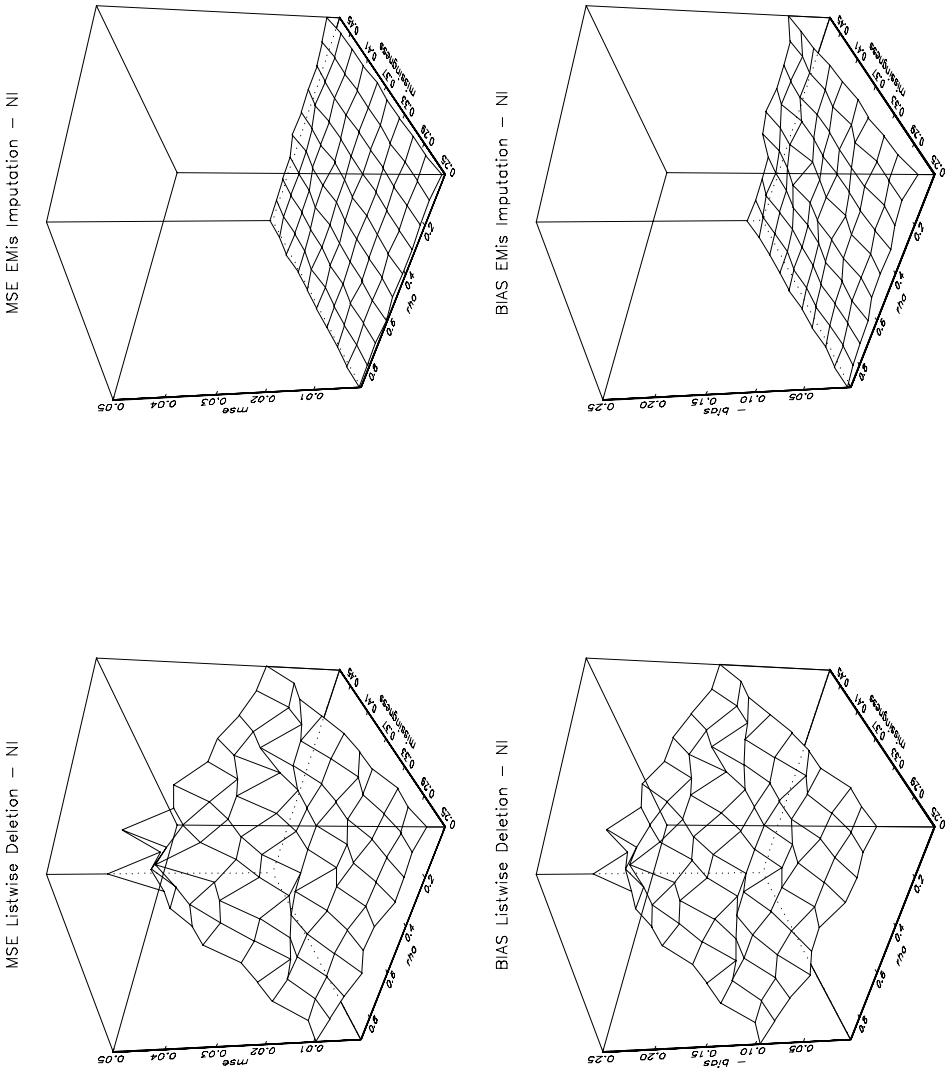

Figure 7. An Informal Tool to Determine MCAR or MAR

The panels on the left give histograms for the Z-score (ZSCR) explanatory variable, conditional on the missingness status for the commitment fee (COMFEE) dependent variable. The top two panels on the right give histograms for the all-in-spread (SPRD) explanatory variable, conditional on the missingness status for the secured status (SECURED) dependent variable. The bottom two panels on the right give histograms for the all-in-spread (SPRD) explanatory variable, conditional on the observed outcome of the secured status (SECURED) dependent variable.

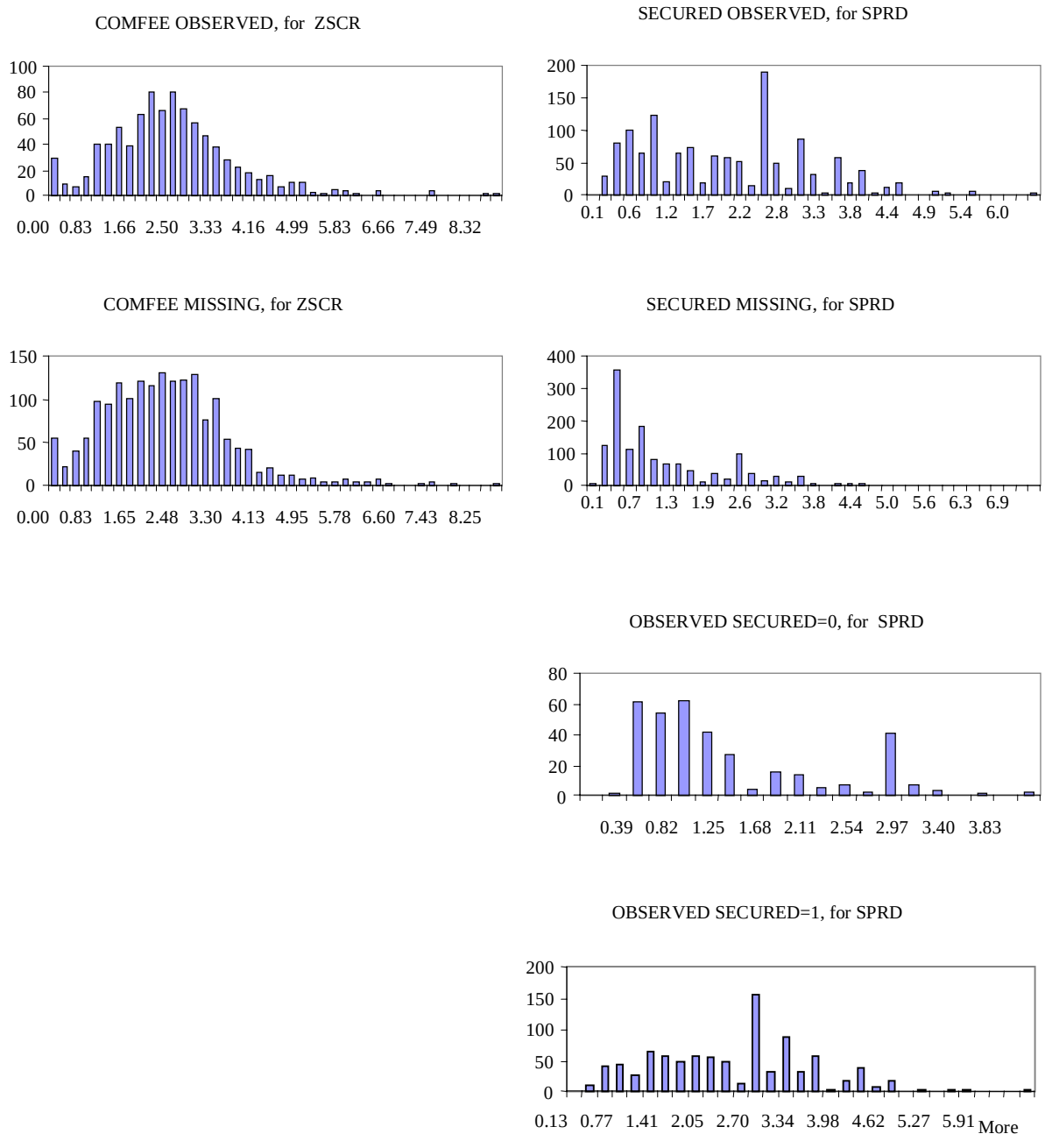

Figure 8. Predictability of Missing Values (Naive versus *EM-is*)

The panels on the left display the proportion of correct predictions for the discrete secured status dependent variable against the number of explanatory variables. The panels on the right display the proportion of correct predictions for the discrete commitment fee dependent variable against the number of explanatory variables. The top two panels are based on *equal probability imputation* (guess). The middle two panels are based on *mean imputation*. The bottom two panels are based on the *EM-is* imputation methodology. The missingness scheme is *Missing Completely at Random* (MCAR).