Using Multiple Imputation in the Analysis of Incomplete Observations in Finance

PAUL KOFMAN
University of Melbourne

IAN G. SHARPE
University of New South Wales

ABSTRACT

Incomplete observations are a common feature of financial applications that use survey response, annual report, and proprietary banking and security issue and pricing data. Finance researchers use a variety of procedures, including deleting offending observations and imputing ad hoc values, that potentially fail to deliver efficient and unbiased parameter estimates. This article examines the application of a statistical framework, multiple imputation methods, that minimizes incomplete data problems if the missingness satisfies certain criteria. When applied to two financial datasets involving severe data incompleteness, the imputation methods outperform the ad hoc approaches commonly used in the finance literature.

KEYWORDS: incomplete data, multiple imputation.

Incomplete data problems occur frequently in applied research in finance. While it is relatively easy to acknowledge the existence of such problems, it seems more complicated to assess 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. For missing data problems, the inherent risk of sample selection bias — due to, for example, survivorship — is well known. However, it is worthwhile to distinguish missing data (where no variables are available for some observations) from incomplete data (where some variables are available but others may be missing for some observations). Specific solutions are often available for missing data problems. For incomplete data problems, these solu-

We gratefully acknowledge support from the ARC grant schemes. Excellent research assistance was provided by Karen Lew. Comments by Neil Esho, Tony Hall, Joe Hirschberg, Guyonne Kalb, Julia Lane, Debarshi Nandy, and participants in seminars at Monash University, the University of Technology, Sydney, the University of New South Wales, the Queensland University of Technology, the Stockholm School of Economics, and at the 12th Annual PACAP/FMA Conference 2000, the 2000 FMA Annual Meeting, and the 8th World Congress of the Econometric Society 2000 are much appreciated. We are indebted to two anonymous referees and an (associate) editor for many valuable comments. Address correspondence to Paul Kofman, Department of Finance, University of Melbourne, Parkville, VIC 3010, Australia, or E-mail: pkofman@unimelb.edu.au.

Journal of Financial Econometrics, Vol. 1, No. 2, pp. 216–249

DOI: 10.1093/jjfinec/nbg013

tions are not apparent or may seem unnecessary. This article investigates whether incomplete data problems can indeed be ignored, and finds that this is not the case.

While some finance researchers may have altogether ignored (and not reported) the incomplete data problem, most have used one of the following three approaches.[1] The predominant approach — the *listwise deletion* method — is to exclude the observations with incomplete data from the study and only use complete records. This involves an efficiency cost due to lost observations. Moreover, finance researchers sometimes find that the observations where data are incomplete appear to have characteristics more attuned to a particular outcome of dependent or independent variables. In these circumstances, exclusion of the incomplete data introduces a systematic bias to the estimates. For example, in a Compustat-based study of the effect of earnings management on the performance of seasoned equity offerings Rangan (1998:105) reports being left with a final sample of 230 offerings after removing "149 firms that did not have sufficient data to compute discretionary accruals ... (and) six firms that did not have stock return or earnings data .... Issuers with insufficient data to estimate discretionary accruals are relatively young and small in terms of market capitalisation ... (and) by excluding them I tilt my sample toward older and larger firms." The impact of deleting 40% of the sample in terms of estimation inefficiency and parameter bias is potentially very significant.

A second common approach — the *omitted variable* method — is to exclude the variables with incomplete data from the analysis. For example, in a study of the costs and benefits of diversification, Servaes (1996:1210) notes that research and development (R&D) expenditures could not be included in the regressions because of severely incomplete data. King et al. (2001) show that this method also risks bias but not inefficiency. Also, variable omission will not be an option when the variables with missing values are the dependent variables in the analysis.

The third approach is to use some kind of imputation method where the missing values are given a value and then treated as observed. The typical 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[2] or using a binomial process to randomly allocate zeroes and ones to missing values. For continuous variables, researchers often use *mean imputation*, that is, replacing all missing values with the mean of the observed values. For example, Gompers (1995:1477) and Denis, Denis, and Sarin (1997:142) use industry means to proxy incomplete firm-level data. Yet another ad hoc value imputation method is to look for matching observations (those that are in every respect identical, but are completely observed). This occurs, for example, when proxy observations are used (or

[1] Many other methods (e.g., the hot deck imputation method) have been used sporadically in the finance literature. For a complete treatment of methods for dealing with incomplete data, refer to Little and Rubin (1987) and Schafer (1997).

[2] Barclay and Smith (1995; 627) impute zero values to missing values of commercial paper ratings; that is, to an indicator variable taking the value of unity if the firm is rated and zero otherwise.

even replacing the variable with missing values with an alternative variable). While the sample size is maximized, potential measurement error and estimation bias is introduced in each of the ad hoc approaches. Moreover, because missing values are now assumed to be known with certainty, regression results will underestimate the standard errors. The estimates of observed data have more uncertainty than the imputations, which have zero variance. A variant of the ad hoc imputation method, used in a recent study of the determinants of managerial ownership by Himmelberg, Hubbard, and Palia (1999), is to impute zero values for the incomplete (R&D and advertising expenditures) observations but include dummy variables in the regressions for the incomplete observations to test whether the nonreporting firms are discretely different from reporting firms. This approach may, in some circumstances, detect possible incomplete data imputation bias, but it does not impute values for missing values that are treated as best estimates for the missing values.

Assuming that the missingness pattern is not completely at random, the values for the observed variables may provide indirect evidence about the likely values of the variables with incomplete observations. 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 incomplete data.[3] Using complete observations, an auxiliary regression is run with the incomplete variable as regressand on a set of (more complete) regressors. The parameter estimates are then used to fit the missing values. However, the imputed values are really estimates. Ignoring the uncertainty of missing value prediction leads to standard errors that are too small.

The incomplete data problems are accentuated in simultaneous-equations studies of interrelated dependent variables. The incomplete data exclusion method — now known as the pairwise deletion method — applied commonly across all equations may dramatically reduce sample size and almost certainly introduce bias into estimates. The regression imputation method can also be adopted and is intuitively even more appealing given that there are more "information channels" than in the single equation context. However, the implications of regression imputation are not straightforward. Whereas pairwise deletion in simultaneous--equations models may lead to inconsistency in the covariance matrix and biased estimates, regression imputation may also lead to biased estimates.

The current article examines the application of a statistical framework that minimizes the problems associated with missing values in data that are independently and identically distributed (i.i.d.). Based on either maximum-likelihood or bayesian estimation methods, it is then often possible to obtain unbiased imputed values for the incomplete observations with correct standard errors.

The statistical literature on incomplete data problems dates back to the mid-1970s with Rubin's (1976, 1978) and Dempster, Laird, and Rubin's (1977)

[3] In Section 2 it is shown that such an approach is one step toward the preferred imputation method.

articles. These articles proposed an iterative maximum likelihood-based procedure to impute "most likely" values for the incomplete data. Dempster, Laird, and Rubin first suggested the by-now standard expectation maximization (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 algorithm is also based on an iterative (posterior density) maximization procedure where the imputation step imputes missing values based on an initial random draw of parameter values and the posterior 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 preferred method to deal with incomplete data problems.

Rubin (1978) and more recently Schafer (1997) developed an extension to the EM algorithm that solves its basic shortcoming. This EM importance sampling (EM-is) method adds a bayesian flavor to the classical EM approach. That is, it allows for the uncertainty regarding the imputations of missing values and hence approaches the theoretical exactness of the IP method. Whereas it does not offer a distinct advantage over the IP method, it offers an alternative to the bayesian methodology that some users may feel more comfortable with.

These imputation methods may be applied to many incomplete data problems in finance where the data satisfies the i.i.d. assumption. This is generally the case for cross-section studies commonly encountered in applied corporate finance and banking studies. However, the methods may also be applied in some time-series or panel data applications, as long as the non-i.i.d. characteristic satisfies certain specifications. For example, the EM-is method can impute incomplete time-series or panel data with an autoregressive lag structure (see Honaker et al., 2000). Moreover, Trevor and Morgan (1999) apply the basic EM algorithm to obtain unbiased estimates for generalized autoregressive conditionally heteroscedastic (GARCH) models with "missing" values (due to censoring).

The remainder of the article is organized as follows. In order to determine the extent and nature of the incomplete data problem in applied finance, Section 1 reports the results of a survey of articles published over a five-year period in five international finance journals. The survey reveals that incomplete data are a significant problem in empirical finance, particularly in cross-section studies likely to satisfy the i.i.d. assumption, and that financial researchers typically treat such data by discarding the incomplete observations or by using ad hoc imputation methods. Depending on the type of data incompleteness, these commonly used approaches to incomplete data may result in biased estimators with incorrect standard errors. In Section 2 we provide an overview of the multiple imputation methods that offer the potential to mitigate these problems in i.i.d. data. We identify three distinct types of data incompleteness, discuss their implications for the multiple imputation methods, and describe how the methods may be implemented. In an appendix we evaluate the statistical properties of the multiple imputation techniques within modeling frameworks often encountered in empirical finance. In addition to the case where incompleteness is in a continuous variable within single equation estimation, we also consider incompleteness in discrete choice variables that are frequently encountered in empirical finance. Moreover, with the increasing interest in modeling interrelated financial decisions within simultaneous-equations models, we also consider the presence of incomplete data in these systems.[4] The results of these Monte Carlo experiments reveal that, for certain types of data incompleteness, the imputation approaches generally produce estimates with less bias and greater efficiency than the approaches commonly used in the finance literature. In Section 3 we provide two financial applications of the multiple imputation techniques. The first involves panel data estimation of a relatively simple cointegration relation between stock price and dividend for 241 companies over the 1825–1870 period, with incomplete data throughout the sample. In this application we find that both listwise deletion and ad hoc imputation each produce significantly biased regression estimates vis-à-vis the IP estimator. The second application involves logit estimation of a model explaining the determination of the secured status of a cross section of revolving credit agreements made by U.S. banks where the dependent variable is characterized by a high degree of missingness. In this illustration, listwise deletion performs relatively well, while ad hoc imputation produces seriously misleading inferences vis-à-vis the multiple imputation technique.

1 SURVEY

An examination of articles published in five major journals of banking and finance,[5] as summarized in Table 1, suggests that incomplete observations are a common feature of many financial applications. In 257 articles of a total of 946 empirical studies (or 27%), authors explicitly recognized the presence of incomplete data. Moreover, it is likely that Table 1 somewhat underestimates the extent of the incomplete data problem in finance. Complicating the investigation is the fact that data descriptions are nonstandard, sometimes incomplete, or uninformative with regard to sources, availability, completeness, and transformations

[4] Dagenais (1976) proposes a procedure to estimate simultaneous-equations models with incomplete observations for continuous dependent and independent variables. His two-stage estimation method involves an auxiliary regression with extraneous variables to impute the missing values. It does not exploit the structure in the simultaneous-equations model to improve the imputations.

[5] The journals included the *Journal of Banking and Finance*, vols. 19–23; the *Journal of Finance*, vols. 50–54; the *Journal of Financial Economics*, vols. 37–52; the *Journal of Financial and Quantitative Analysis*, vols. 30–34; and the *Review of Financial Studies*, vols. 8–12.

Table 1 Literature survey of incomplete data in finance: 1995–1999.

Journal	Articles acknowledging missing values	Missing values occur in: Independent variable	Dependent variable	Both	Missing values occur in a: Cross-sectional analysis	Time-series analysis
JBF	67 (397) [270]	34	14	19	50 (215)	40 (220)
JF	98 (365) [292]	69	16	13	85 (246)	37 (252)
JFE	53 (222) [201]	34	8	11	51 (189)	20 (178)
JFQA	20 (135) [95]	13	3	4	14 (77)	9 (84)
RFS	19 (176) [88]	10	5	4	13 (70)	17 (79)
All	257 (1295) [946]	160	46	51	213 (797)	123 (813)

This table presents information on the number of articles that acknowledge the presence of an incomplete data problem, the type of variable where the problem occurs, and the type of empirical application in which the problem occurs. The total number of articles appearing in the journal is given in parentheses. The number of empirical articles appearing in the journal is given in square brackets. The journals include the *Journal of Banking and Finance* (JBF), vols. 19–23, the *Journal of Finance* (JF), vols. 50–54, the *Journal of Financial Economics* (JFE), vols. 37–52, the *Journal of Financial and Quantitative Analysis* (JFQA), vols. 30–34, and the *Review of Financial Studies* (RFS), vols. 8–12.

applied. Consequently we were only able to determine the proportion of incomplete data in 72 (or 28%) of the 257 articles acknowledging incomplete data. The proportion of data that is reported as missing in these articles is significant, varying from 0.1% to 81.1%, with a mean of 23.3% and median of 15.7%. Thus both the frequency and extent of incomplete data in published finance articles is high.

Also of interest in Table 1 is the analysis of whether the missingness occurs in the dependent or independent variables and in time-series or cross-sectional analysis. Missingness is more frequently reported in relation to independent variables than dependent variables. The less frequent reporting of missing dependent variable values may reflect the practice of some authors of reporting that their sample includes the number of available observations of the dependent variable.[6]

The numbers in the last two columns of Table 1 are the number of articles with incomplete data in cross-sectional and time-series applications, respectively. The

[6] There may also be significant self-selection. Whereas it may be considered straightforward to estimate a model with some missing values in the explanatory variables, a similar missingness among the dependent variables may often seem more complicated.

Table 2 Treatment of incomplete data in finance: 1995–1999.

Journal	Articles acknowledging missing values	Missing value treatment			
		Listwise deletion	Regression imputation	Ad hoc imputation	Proxy imputation
JBF	67	56	5	5	3
JF	98	77	6	9	7
JFE	53	44	2	3	5
JFQA	20	18	—	1	2
RFS	19	10	3	7	1
All	257	205	16	25	18

This table presents information regarding the treatment of an incomplete data problem in those articles that acknowledge the presence of incomplete data. The journals include the *Journal of Banking and Finance* (JBF), vols. 19–23, the *Journal of Finance* (JF), vols. 50–54, the *Journal of Financial Economics* (JFE), vols. 37–52, the *Journal of Financial and Quantitative Analysis* (JFQA), vols. 30–34, and the *Review of Financial Studies* (RFS), vols. 8–12.

article is included in both columns if a combination of cross-sectional and time-series data was used.[7] The numbers in parentheses in these columns are the total number of articles of the analysis type.[8] There is some evidence in Table 1 that incomplete data problems are more common in cross-sectional studies, with 27% of cross-sectional analyses acknowledging incomplete data, compared to 15% of time-series analyses.[9]

Table 2 summarizes the methodology applied when the researcher was confronted with an incomplete data problem. In most of the missing value cases, the solution adopted in the article was listwise deletion (in 205 cases), while regression imputation, ad hoc imputation, and proxy imputation were used less frequently.[10]

This survey of the applied finance literature suggests that incomplete data are a relatively common problem in empirical finance, and particularly in cross-sectional studies likely to satisfy the i.i.d. requirement necessary for the application of the imputation techniques. Moreover, researchers have generally adopted

[7] Note that even if the empirical analysis was based on a combination of time-series and cross-sectional data, it often occurred that the incomplete data problem was only relevant for the time-series or the cross-sectional analysis.

[8] Event studies (or asset pricing models) use a mixture of time-series analysis to estimate excess returns (to estimate asset betas) and cross-sectional analysis to estimate event parameters (to estimate premia). Hence these studies are counted under both columns.

[9] An alternative explanation for the phenomenon could be that some researchers deem it more acceptable for cross-sectional studies to have incomplete data than for time-series studies.

[10] Note, however, that in a substantial number of cases, this information was derived indirectly from the data description.

approaches to deal with incomplete data problems that are potentially biased and/or inefficient. Hence the remainder of the article describes the multiple imputation methods, examines their statistical properties vis-à-vis the common approaches in finance, and illustrates their potential application.

2 EM AND IP IMPUTATION ALGORITHMS

2.1 Type of Missingness

Excellent statistical treatments of incomplete data imputation methodologies can be found in Rubin (1987) and Schafer (1997). Insight into the consequences of missing values and potential solutions for statistical analysis with missing values requires an understanding of why and how missing values occur. The occurrence of missing values can be captured by three distinct missingness schemes, distinguished by whether the source of missingness is completely independent from the dataset or any other source, internal to the dataset, or external to the dataset. Suppose x is a completely observed variable for an incomplete variable y. Missing values in y are now said to be (see Rubin, 1976):

- Missing completely at random (MCAR)—
  when the missingness in y is independent of both x and y or indeed of any other source of information. The missing values are then missing at random, while the observed values are observed at random.
- Missing at random (MAR)—
  when the missingness in y depends on x but not on y. Missing values are still missing at random, but the observed values are no longer observed at random.
- Non ignorable (NI)—
  when the missingness in y depends only on y and cannot be explained by x.

To illustrate the distinction in missingness types, consider the following application based on Pulvino's (1998) study of asset fire sales. 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 excluded, however, due to some missing values in the transaction prices. Let us assume that these post-1991 transaction prices are missing simply because of data handling. Given that missingness only appears after 1991, this seems unlikely. However, it is possible that a new data manager was appointed in 1991. If it can reasonably be assumed that this new "sloppy" data manager makes these mistakes at random, then the missingness type is MCAR. Instead, consider the possibility 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 presumably low transaction price gets reported. The missingness would now be of type MAR. The negative relationship between age and transaction price can now be extrapolated for the older aircraft to predict the missing transaction prices. However, for this particular data series Pulvino (1998:947) notes that "... [post] 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 (very high or very low) transaction prices. Thus missingness in the dependent variable is now a function of the value of the dependent variable itself. This is known as type NI.

When the missingness is determined outside the dataset and we do not have the necessary information as to how the selection occurred, it is almost impossible to infer unbiased values for the incomplete data from the observed data. This suggests that Pulvino did the best he possibly could, given NI missingness, by excluding the observations with missing values. However, as long as the missingness scheme is not NI, likelihood-based imputation methods should be used to generate unbiased and efficient estimates for the complete data statistical model. Thus if the MCAR or MAR assumption is reasonable, then among the completely observed x, the distribution of y is the same for y_{obs} (the observed dependent variable values) as it is for y_{mis} (the missing dependent variable values). This implies that the relationship between x and y for the observed values can be extrapolated to the missing values, for which we observe the x values. This is known as the ignorability assumption, that is, missingness is assumed to be either of type MCAR or type MAR. In subsection 2.3.1 we discuss a test by Little (1988) for assessing the MCAR assumption. If this test is rejected by the data, we then conclude that either the MAR or NI assumptions apply. Of course, given that NI missingness depends on unobservables, it is impossible to test the ignorability assumption against the NI assumption.

Recent articles by Rotnitzky, Robins, and Scharfstein (1998) and Horowitz and Manski (2000) deal with nonignorable (nonrandomly) incomplete 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 at face value. Ultimately, whether one can reasonably assume ignorability of the missingness model depends on the specific application.

For now, assume that the data satisfies this ignorability assumption. Consider the data matrix Z (of dimension $n \times p$) consisting of dependent and independent variables, some of which are not completely observed. Matrix Z can be partitioned according to missingness status, such that $Z = (Z_{obs}, Z_{mis})$. Assuming the observations are independently and identically distributed (*i.i.d.*), the probability density function (PDF) of the complete data can be written as the product of the n densities

$f(\cdot)$ for the individual observations z_i:

$$f(Z \mid \theta) = \prod_{i=1}^{n} f(z_i \mid \theta) \tag{1}$$

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 Z, 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 $n \times p$ as the data matrix can be introduced to indicate which part of Z 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 values, the missing values, 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 PDF of this missingness matrix Π simplifies to

$$f(\Pi | Z_{obs}, Z_{mis}, \vartheta) = f(\Pi). \tag{2}$$

As the probability distribution does not depend on any available information in the data (missing or observed), missingness is truly random. On the other hand, when the missingness type is assumed to be MAR, the PDF of this missingness matrix Π simplifies to

$$f(\Pi | Z_{obs}, Z_{mis}, \vartheta) = f(\Pi | Z_{obs}, \vartheta) \tag{3}$$

While the probability distribution is independent of the incomplete data, it does depend on the observed data, and if the incomplete data are MAR, then the observed data help explain incomplete data. Combining the simplification in Equation (3) with Equation (1) and assuming that θ and ϑ are distinct nuisance parameters (Schafer, 1997), then

$$f(\Pi, Z_{obs} \mid \theta, \vartheta) = f(\Pi \mid Z_{obs}, \vartheta) f(Z_{obs}, \theta) \tag{4}$$

The joint assumptions of MAR (or MCAR) and distinctness allow ignorability of the missingness model (Rubin, 1976). Equation (4) then implies that likelihood estimation of θ (the parameters of interest) is unaffected by the exact model for missingness, but not the missingness type. Nor does it imply that the missing values are of no consequence to inference on θ. All the necessary information to complete the data (i.e., to simulate the missing values) is contained in the observed data.

2.2 Imputation Methods

A difficulty in multivariate data with arbitrary missingness patterns is that the data likelihood, as derived in Rubin (1974), is very complicated. As there are no closed-form solutions, and numerical computation is not straightforward, iterative procedures have an intuitive appeal because they exploit 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 simulating missing values. At the same time the missing values have relevant information with regard to the (power of the) parameter estimates. This interaction has been captured by two iterative techniques known as the bayesian IP method and the classical EM algorithm.

2.2.1 The IP method The *imputation posterior* methodology, discussed in Tanner and Wong (1987), is an iterative bayesian procedure,[11] imputing values for the incomplete observations and making inferences about unknown parameters in a stochastic manner. IP initially simulates missing values 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 simulated (imputed) data. Thus we begin by selecting starting values ($\tilde{\theta}^{(0)}$) and sample from

$$\tilde{Z}_{mis}^{(i)} \sim f\left(Z_{mis} \,|\, Z_{obs}, \tilde{\theta}^{(i-1)}\right), \tag{5}$$

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

$$\tilde{\theta}^{(i)} \sim f\left(\theta \,|\, Z_{obs}, \tilde{Z}_{mis}^{(i)}\right), \tag{6}$$

which is the posterior (P) step. Iteratively sampling for $i = 1, 2, \ldots, N$ while updating the conditioning variables in Equations (5) and (6) produces a sample of N parameter sets (of parameter values) which converge in distribution (asymptotically) to the posterior distribution $f(\theta|Z)$. Features of the posterior distribution, including its marginal densities and probability intervals, can be extracted from this sample.

An advantage of the IP method is that the parameter distribution converges to a posterior distribution averaging over the missing values (Schafer, 1997). The parameter distribution and missing values distribution converge to an (exact) predictive distribution for θ and Z_{mis}, respectively. Assessment of convergence is relatively straightforward based on inspection of the parameter autocorrelation plots.

2.2.2 The EM method An alternative method is provided by the expectation maximization algorithm, first introduced by Dempster, Laird, and Rubin (1977). Without any assumptions on the type of missingness, the distribution of the (incompletely observed) data Z can be split into an observed data distribution and an incomplete data distribution, as follows:

$$f(Z \,|\, \theta) = f(Z_{obs} \,|\, \theta) f(Z_{mis} \,|\, Z_{obs}, \theta) \tag{7}$$

The data likelihood for the parameters of interest can then be expressed as

$$L(\theta \,|\, Z) = L(\theta \,|\, Z_{obs}) + \ln f(Z_{mis} \,|\, Z_{obs}, \theta) + \tau, \tag{8}$$

[11] The IP algorithm is also known as *data augmentation*, one of a series of Markov chain Monte Carlo methods developed to generate pseudo-random draws from probability distributions.

with τ, a constant. The second term on the right-hand side of Equation (8) is a predictive distribution of the incomplete data given the observed data and the parameters. Given that the missing values are not observed, this term cannot be computed. Instead, by replacing θ with $\tilde{\theta}^{(i)}$, which for $i = 0$ is an initial estimate of the unknown parameters, the EM algorithm computes the missing values and parameters iteratively. 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 Z, given the observed data Z_{obs} and the initial parameter estimate $\tilde{\theta}^{(0)}$. It then computes expected values for the missing values $\tilde{Z}^{(1)}_{mis}$. The M step then takes the estimated complete data $\tilde{Z}^{(1)}$ 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 values $\tilde{Z}^{(2)}_{mis}$) until convergence is achieved based on $\tilde{\theta}^{(i)} \approx \tilde{\theta}^{(i-1)}$. Convergence is based on the parameter estimates' contribution to the likelihood and occurs without any assumptions on the derivatives or the starting values. It will also occur for small sample sizes (Dempster, Laird, and Rubin, 1977). 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 Z_{mis}. Both the EM imputations and the EM parameters are single values (maximum posterior), instead of a complete distribution. The method therefore ignores estimation uncertainty and consequently underestimates the standard errors (Schafer, 1997).

Despite its computational simplicity and good convergence properties, very few finance applications have utilized the EM methodology to simulate missing values. One application is by Trevor and Morgan (1999), who use the algorithm to estimate a time-varying volatility model for censored asset return observations. Another interesting application is in the estimation of parameters in Markov-switching models, as in Hamilton (1994) and Krolzig (1997). This type of nonlinear time-series model has recently been used in a financial application by Perez-Quiros and Timmermann (2000). However, the missingness in these models is of a different nature than the missingness we consider in this article. The Trevor and Morgan application is an example of NI missingness for which a specific solution (estimation methodology) exists. The Markov-switching applications simulate a latent variable, that is, a variable which is unobserved for all observations. Simulating values for this latent variable requires strong assumptions on the generating stochastic process.

2.2.3 The EM-is method Several extensions of the EM algorithm are now available, each introducing imputation uncertainty into the EM algorithm through sampling procedures. For space reasons, in this article we present only the EM-is method, as it is generally the preferred approach in the literature.[12] EM-is first applies the standard EM algorithm to generate the maximum posterior parameter

[12] Alternatives include the EM sampling procedure of Tanner (1996) and the stochastic EM algorithm in Ip (1994).

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 $V(\tilde{\theta}^*)$ subject to an acceptance-rejection algorithm that keeps parameter draws $\hat{\theta}$ with probability proportional to the importance ratio,

$$IR = \frac{L(\hat{\theta}|Z_{obs})}{N(\hat{\theta}|\tilde{\theta}^*, V(\tilde{\theta}^*))}, \tag{9}$$

and discards the rest. The IR is a ratio of the actual posterior distribution, $L(\cdot)$, to the asymptotic normal approximation, $N(\cdot)$. This implies that the likelihood is evaluated more frequently at the important segments of the range than otherwise. This sampling procedure is repeated m times (after the initial EM step and variance computation) which generates a sampling distribution for Z_{mis}.

The EM-is 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 easily determined. While Schafer (1997) suggests that it might not do so well for nonnormal likelihoods, the data can often be transformed so that they are approximately multivariate normal.

2.2.4 A Monte Carlo comparison of imputation methods To assess the relative size of the bias and/or inefficiency caused by these missing value imputation methodologies vis-à-vis simply discarding observations with missing values, a series of Monte Carlo experiments were designed. One set considers a single-equation regression model with missingness in the independent variable. The other set of experiments considers a simultaneous-equations model with missingness in the (related) dependent variables. Details of the design of these experiments and the results can be found in the appendix.

The single-equation results indicate that the IP and EM-is imputation methods outperform the common procedure of deleting incomplete observations, in terms of both reducing bias and improving the efficiency of the estimates. Note that the single-equation experiments were designed for missingness in the independent variable. For missingness in the dependent variable, the bias in the parameter estimates disappears for listwise deletion as well as for the IP and EM-is imputation methods, but the relative efficiency gain of the IP and EM-is imputation methods remains.

The comparative advantage of IP and EM-is imputation methods over pairwise deletion, though still considerable, is somewhat less for the simultaneous-equations model. For MCAR missingness and continuous dependent variables, both methods produce relatively unbiased estimates, though the IP and EM-is methods are considerably more efficient. Again, it is in the MAR experiments where the IP and EM-is methods produce markedly superior estimates with greater efficiency and less bias than pairwise deletion. When one of the dependent variables is discrete, with MCAR missingness, pairwise deletion is slightly less biased, but less efficient, than the IP and EM-is methods. However, for MAR missingness, the IP and EM-is methods reassert their relative advantage with less biased and markedly more efficient estimates than pairwise deletion. Also note that the difference in bias and efficiency of the IP and EM-is results is negligible. The choice between the two multiple imputation methods will ultimately depend on the user's preference for bayesian or classical inference. The improved user friendliness of specialist IP multiple imputation software (with convenient tools to judge convergence) combined with improved computational efficiency have made this method much more accessible.

2.3 Implementation

2.3.1 Pretesting for MCAR Most studies surveyed in Section 1 make the implicit assumption that missing values are missing completely at random. If this assumption holds, the parameter estimates will at least be unbiased (though not necessarily efficient). For each variable with missing values, the dataset is split into a subset with observed values and a subset with missing values for the incomplete variable. The two means of the observed values for the other (completely observed) variables are then compared by a two-sample *t*-test. An alternative, and potentially more insightful, procedure is to plot the empirical distributions of the other variables for the two subsets. This would provide insight into other comparative moments than just the means. At any rate, both procedures are cumbersome since they yield at least $p(p-1)$ *t*-tests for datasets with missing values on p variables. Little (1988) therefore proposes a single test statistic for testing MCAR, which is easy to compute (our tests are computed in Excel, but the test is also implemented in SPSS and SYSTAT).

The data matrix z is an $(N \times p)$ matrix of N observations on p variables. $\hat{\mu}, \tilde{\Sigma}$ are maximum likelihood estimates of the population mean vector $(1 \times p)$ and the population covariance matrix $(p \times p)$ of z. The z matrix can be split into K submatrices (S_k) of distinct missingness patterns in the data, including the complete observations as a separate pattern. Each submatrix k has m_k observations and p_k observed variables. $\bar{z}_{obs,k} = m_k^{-1} \sum_{i \in S_k} y_{obs,i}$ is a $(1 \times p_k)$ vector of means of observed variables for pattern k. A matrix D_k $(p \times p_k)$ indicates which variables are observed for pattern k. It has one column for each observed variable, and has $p-1$ zeroes and a single one for the observed variable. The population estimates for each submatrix are then $\hat{\mu}_{obs,k} = \hat{\mu} \times D_k$, and $\tilde{\Sigma}_{obs,k} = D_k' \tilde{\Sigma} D_k$. When Σ is not known, and has to be estimated, Little derives the following test statistic,

$$d^2 = \sum_{k=1}^{K} m_k (\bar{z}_{obs,k} - \hat{\mu}_{obs,k}) \tilde{\Sigma}_{obs,k}^{-1} (\bar{z}_{obs,k} - \hat{\mu}_{obs,k})', \tag{10}$$

which combines the individual means test into a single statistic. This test is asymptotically chi-squared distributed with $\sum p_k - p$ degrees of freedom. The computation of $\hat{\mu}, \tilde{\Sigma}$ is iterative and can be obtained by running the simple EM imputation model described in subsection 2.2.2. MCAR will be rejected for large values of the d^2 test statistic. A rejection of MCAR missingness implies that parameter estimates obtained from listwise (or pairwise) deletion of incomplete observations would be biased. If MCAR missingness cannot be rejected, this would imply unbiased parameter estimates with listwise (or pairwise) deletion, but still inefficient estimates.

2.3.2 Estimation Rather than estimate the incomplete data likelihood function directly, the iterative IP and EM-is 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 missing values by their simulated values, conditioning upon all extraneous information available. That is, implementation requires the researcher to include as many variables as possible in the imputation model. The imputation model should be at least of the same dimension as the analysis model because a smaller imputation model may result in omitted variable bias. Consider, for example, that the researcher wants to avoid overspecification of the model and discards those variables that have a minor contribution to the overall explanatory power. Whereas the omitted variables may have a limited impact on the overall parameter estimates, they may have critical explanatory power for the missing values. Excluding variables from the imputation model may even cause NI missingness when including all available variables would have satisfied MAR missingness. This suggests that the imputation model may well 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. (2001) argue that the risk of overspecification is not an issue of concern at the imputation stage.

This distinction between imputation and analysis models is reflected in the two possible implementations of the IP and EM-is imputation methods (Rubin, 1978): (i) parameter simulation to simulate m parameter values $\tilde{\theta}^{*(1)},\tilde{\theta}^{*(2)},\ldots,\tilde{\theta}^{*(m)}$ from the observed data posterior density for the parameters $f(\theta/Z_{obs})$;[13] or (ii) multiple imputation to simulate m missing values $\tilde{Z}^{*(1)}_{mis}$, $\tilde{Z}^{*(2)}_{mis},\ldots,\tilde{Z}^{*(m)}_{mis}$ from the observed data posterior density for the missing values $f(Z_{mis}|Z_{obs})$. In parameter simulation, the analysis model will be identical to the imputation model. Thus the iteratively optimized parameter simulations, $\tilde{\theta}^{*(i)}$, will be those of direct interest to the researcher. In multiple imputation, the analysis model may be of smaller dimension than the imputation model. The parameter simulations are then of no direct interest to the researcher. Instead, the Z_{mis} are replaced successively by simulated values $\tilde{Z}^{*(i)}_{mis}$ and each of the m complete datasets are subsequently analyzed by standard methods. 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

[13] Parameter simulation seems attractive since it directly provides the required parameter estimates. These could be functions of $\tilde{\theta}^{*(i)}$. However, as explained above, the imputation model does not necessarily match the (typically 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.

interest. Unlike parameter simulation, 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 values can be imputed linearly (Rubin, 1976),

$$\tilde{z}_{ij}^{*} = z_{i,-j}\tilde{\lambda}^{*} + \tilde{\varepsilon}_{i}, \tag{11}$$

where $\tilde{z}_{ij}^{*}$ is a simulated value for missing value i and variable j, $z_{i,-j}$ 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. Thus multiple imputation requires m independent draws for the missing values from a posterior predictive distribution for θ.

2.3.3 Inference Having computed the m different imputed datasets,[14] inference can then be drawn with respect to the parameters b, and their variances s_b for the analysis model (Rubin, 1987). The successive estimation of the analysis model generates m equally likely estimates $(b_1, \ldots, b_m; s_1, \ldots, s_m)$ for the parameters, b, and their variances, s, respectively. The combined multiple imputation estimate is then

$$\bar{b} = \frac{1}{m}\sum_{i=1}^{m} b_i. \tag{12}$$

The combined variance (and standard error) estimates for these parameter estimates have two components. The within-imputation component measures the average of the m variances obtained within each model estimation. The between-imputation component measures the standard error across the m parameter estimates, that is, the error caused by imputation of unknown missing values. The within-imputation variance $\bar{s}$ and between-imputation variance s_m are given by

$$\begin{aligned} \bar{s} &= \frac{1}{m}\sum_{i=1}^{m} s_i \\ s_m &= \frac{1}{m-1}\sum_{i=1}^{m}(b_i - \bar{b})^2 \end{aligned} \tag{13}$$

[14] The imputation models for this article have been estimated using the Gauss programs readily available on the Internet at http:\\Gking.Harvard.edu and are explained in Honaker et al. (2000). S-PLUS programs for IP and EM imputation are available on the Internet at http:\\www.stat.psu.edu/~jls and are explained in Schafer (1997). Schafer also provides a stand-alone software package, NORM, that offers user-friendly diagnostics to interpret convergence of the IP method. Standard packages like Stata and SPSS have some missing value imputation options, but they are restricted to regression imputation, mean imputation, or standard *EM*. Stata intends to incorporate multiple imputation in its next release. SAS has recently included new procedures (in its SAS/Stat Product, release 8.1) for creating multiple imputations for incomplete multivariate data. The analysis models for this article have been estimated using Gauss programs written by the authors of this article, and are available on request.

and total variance s_b, standard error $se_{\overline{b}}$, and confidence interval $ci_{\overline{b}}$ are given by

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

$$se_{\overline{b}} = \sqrt{s_b} \tag{14}$$

$$ci_{\overline{b}} = \overline{b} \pm t_{df} se_{\overline{b}} \quad \text{with } df = (m-1)\left(1 + \frac{\bar{s}}{((1/m)+1)s_m}\right)^2.$$

Uncertainty with regard to the parameter estimate decreases with the number of imputations, directly in the total variance s_b, and indirectly through the degrees of freedom. Thus uncertainty increases with the ratio

$$r_b = \frac{((1/m)+1)s_m}{\bar{s}}, \tag{15}$$

which measures the relative weight of between-imputation variance to within-imputation variance (Rubin, 1987) in total variance s_b. This ratio can then be used to determine π_b, the fraction of missing information with regard to parameter estimate b,

$$\pi_b = \frac{r_b + 2/(df+3)}{r_b + 1}. \tag{16}$$

The efficiency for a single parameter estimate b based on m imputations is then approximately

$$\frac{1}{1 + (\pi_b/m)}, \tag{17}$$

which approaches 1 (= 100%) for an infinite number of imputations.[15] Empirically one would rarely need more than 10 imputations, as increasing the number of imputations to $m = 10$ already increases the efficiency to 94%. Further imputations will only contribute minor efficiency gains.

3 ILLUSTRATIVE EXAMPLES

In this section we illustrate the application of multiple imputation techniques on two financial datasets. The first is a historical panel database of share prices and dividends on New York Stock Exchange (NYSE) stocks over the 1825–70 period. Typical of many historical databases, incomplete data occurs frequently throughout this sample. The second is a cross-sectional database that has been used to study the determinants of bank loan contract terms. With voluntary reporting of

[15] Schafer (1997) inverts this statement using $\sqrt{1 + \pi_b/m}$, which for $m = 10$ and $\pi_b = 0.65$ implies a standard error for the parameter estimate which is 1.032 times as large as the standard error based on an infinite number of imputations.

loan details, the observations are frequently incomplete. In each of the illustrations we compare results using the two commonly used missing value techniques in empirical finance, listwise deletion and ad hoc imputation, with the IP multiple imputation method.

3.1 Cointegration Between Prices and Dividends: Panel Data

Our first application is based on a new historical database compiled by the Old New York Stock Exchange Research Project team of the International Center for Finance at Yale University. The database comprises monthly stock prices and annual dividend data over the 19th century on more than 600 individual securities. There are a few missing months in the price data. The problem is more severe for the dividend data. Goetzmann, Ibbotson, and Peng (2001:7) report that, "Since we do not know whether these journals always reported dividends . . . , we do not know whether exclusions of dividends meant that they were not paid, or whether we failed to find them. . . . the number of stocks for which [we] have an unbroken series . . . is small."

We select our samples from the "NYSE monthly price 1815–1925" Excel spreadsheet and the "NYSE annual dividend 1825–1870" Excel spreadsheet. We use the lowest common frequency, annual data, and select the December prices and annual dividends over the 1825–70 period. After matching the dividend and price data files, we end up with an unbalanced panel sample of 241 firms with an average of 19 annual observations per firm (out of a maximum of 46). The number of firms is substantially less than the 600 firms identified by Goetzmann, Ibbotson, and Peng (2001). We applied a few filters that explain the difference. First, there are only 515 firms for which there are dividend observations. Second, we need at least three time periods per firm in order to estimate the firm time-series regressions. Third, we only considered firm data that were based on common equity. We then created an industry dummy by coding the industry indicators in the original files numerically (from 1 to 7). Ultimately we have a total of 4625 observations of which 1419 are completely observed, 835 observations where prices are missing, 1336 observations where dividends are missing, and 1035 observations where both price and dividend are missing. Normally we would exclude the latter observations, but we expect a high degree of persistence in the price and dividend time series and therefore include lagged price variables and lagged dividend variables in our imputation model. This satisfies the requirement that the observations are incomplete, and not completely missing. As can be seen in Figure 1, the missingness occurs fairly evenly throughout the 46-year period. The number of observations (firms) per year increases steadily from 39 in 1825 to a peak of 163 in 1853, after which it declines again to 100 in 1870. Average missingness is 69%, with one exceptional year (1867) of 100% missingness. For that year, there are no price observations for any firm, but there are dividend observations.

Goetzmann, Ibbotson, and Peng (2001) use the historical price-dividend dataset to estimate the power of past returns and dividend yields to forecast future

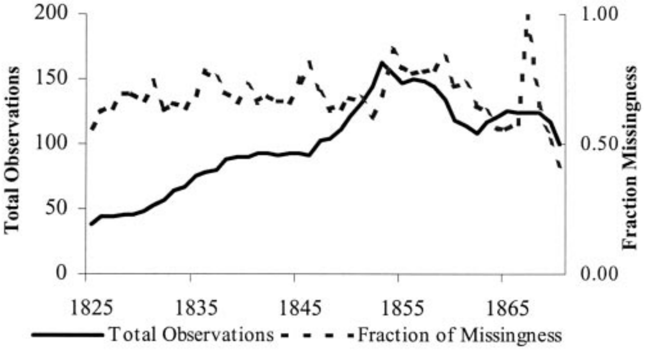

Figure 1 Panel size and missingness, NYSE 1825–1870.

long-horizon returns.[16] The use of first differenced series (returns) in combination with the levels of the series (dividend yields) is fairly common in this asset pricing literature. Such models are, however, only correctly specified if a cointegrating relationship exists between the levels of prices and dividends. A standard procedure is therefore to pretest for a cointegrating relation before estimating the error correction model, as in Goetzmann, Ibbotson, and Peng (2001). We therefore estimate the loglinear relationship between asset prices and dividend payments over the 1825–70 period. To allow for variation across individual firms, we specify the following fixed effects panel data regression:

$$\ln(P_{it}) = \beta_0 + \beta_{0i} + \beta_{0t} + \beta_1 \ln(D_{it}) + u_{it}, \tag{18}$$

where variation across firms (β_{0i}) and across time (β_{0t}) changes the intercepts of the conditional mean function. Assuming that prices and dividends contain a unit root (are nonstationary), Equation (18) represents a cointegrating relation only if the residuals u_{it} are found to be stationary. After estimating Equation (18), we test for cointegration in panel data based on the average of the Phillips and Ouliaris (1990) $\hat{Z}_{it}$ cointegration test statistics for each firm i. The applicability of this test in panel data is explained in McCoskey and Kao (1998).

Given the typical nonstationarity of the (logarithmic) price and dividend series, inclusion of these variables violates the assumptions of the imputation model. The IP imputation model thus uses first differences of these nonstationary variables. To exploit the possible autocorrelation in these first differenced series, we include the one-period lagged first differenced series. Of course, the lagged variables have a similar degree of missingness as the original variables, but due to the lag, this missingness occurs in different observations. Finally, we also include an industry dummy variable.

[16] This summary of their analysis does some injustice to Goetzmann, Ibbotson, and Peng (2001). Their study provides a comprehensive analysis of the historical NYSE dataset. From individual stock price and dividend series they create the NYSE historical stock index and compute the equity premium over the century preceding the Center for Research in Security Prices (CRSP). Their subsequent analysis of the predictability of long-horizon returns is based on this calculated price-weighted index and index dividend yield.

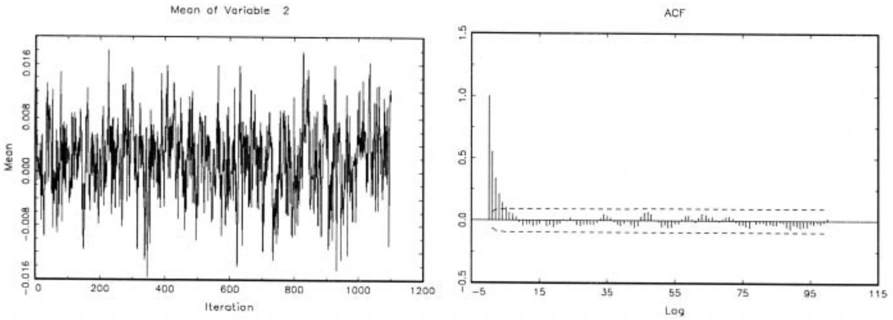

Figure 2 Convergence and autocorrelation in the IP imputation model. The graphs show the mean of the logarithmic price change imputation iterates and the autocorrelation function for the imputations.

Table 3 Cointegrating relation estimates.

	Parameter estimates (*t*-values)		
Independent variables	Pairwise deletion	Ad hoc imputation	Imputation posterior (IP)
CONS	4.522 (209.39)**	4.480 (476.05)**	4.039 (96.74)**
LDIV	0.075 (7.06)**	0.060 (8.31)**	0.259 (11.29)**
PO-ADF	−2.503	−2.303	−3.736*
Observations	1419	2755	4625

Columns give the parameter estimates with *t*-values in parentheses of the logarithm of price versus logarithm of dividend panel data estimates using pairwise deletion, ad hoc imputation, and the IP multiple imputation methods. Explanatory variables are constant (CONS) and the logarithm of dividend (LDIV). PO-ADF gives the Phillips and Ouliaris (1990) cointegration test statistic for panel data. Data source is the Old New York Stock Exchange Research Project at the International Center for Finance at Yale University.

*, ** Statistical significance at the 95% and 99% confidence levels, respectively.

Initially the Little (1988) test for MCAR missingness was applied to the data, resulting in a d^2 statistic of 338.04, which is highly significant (at a 99% critical value of $\chi_8^2 = 20.09$), allowing us to reject the null of MCAR missingness.[17] We then apply the IP imputation methodology and impute $m = 10$ complete datasets from which we then "reconstitute" the logarithmic price and dividend series. Convergence of the IP iterations is illustrated for the logarithmic price change variable in Figure 2. Quite clearly, the mean converges very quickly. The autocorrelation function illustrates persistent autocorrelation at lags 1–6, but insignificant at higher lags.

In Table 3 we report three sets of panel data regression estimates. The first column gives the pairwise deletion estimates based on $N = 1419$ observations. The

[17] One referee suggested a further test to get a sense of whether the pattern of missingness is consistent with MCAR. A principal components analysis of the missingness matrix Π would allow one to investigate whether there is a component that captures where the missingness is generated.

ad hoc imputation estimates in the second column are based on a sample of $N = 2755$ observations with zero values imputed for missing values of the dividend variable and listwise deletion on the missing values for the price variable. The third column gives the multiple imputation IP estimates based on the $m = 10$ imputed datasets. Using Equations (16) and (17), the efficiency of the individual IP parameter estimates lies between 93% and 94%. Whereas the "common" constant (across firms and time) parameter estimate is fairly close, the log dividend parameter is significantly different between IP multiple imputation on the one hand and pairwise deletion and ad hoc imputation on the other hand.

Of course, we can only reliably interpret these results if Equation (18) is indeed a cointegrating relationship. For that purpose we also report the average of the cross-sectional Phillips and Ouliaris (1990) $\hat{Z}_{it}$ cointegration test statistics. The results are striking. We clearly reject cointegration at the 99% level for both pairwise deletion and ad hoc imputation estimations, which suggests that the parameter estimates are spurious. For the multiple-imputation IP estimations, we cannot reject cointegration (99% critical value = −3.387). The between-imputation standard deviation $\sqrt{s_m}$ (Equation (13)) for the IP cointegration test is only 0.06 and indicates the robustness of this cointegration result. This suggests that a relation between returns and dividend yields that ignores the error correction component in levels is misspecified and justifies the model in Goetzmann, Ibbotson, and Peng (2001).

3.2 Bank Loan Secured Status: Cross-Sectional Data

In the banking and corporate finance literature there is considerable interest in the determinants of loan/debt contract terms including the secured status, seniority, maturity, and pricing of the loan. Of particular interest is the question of whether the conflict of interest between the lenders to the firm and its owners can be mitigated by the choice of contract terms. As many of these studies use incremental cross-sectional data of loan or debt issues, where the reporting of contract terms is not compulsory, the data is frequently incomplete.

To illustrate the effect of alternative treatments of missingness we use data from Dennis, Nandy, and Sharpe's (2000) study of the determinants of contract terms in bank revolving credit agreements, often referred to as "revolvers." We test the hypothesis that firms with extensive growth opportunities are more likely to use secured rather than unsecured debt as a means of mitigating share-holder/debt-holder conflicts. The dependent variable is a discrete choice variable, *SECURED*, that takes the value of unity when the revolver is secured and zero otherwise. The proxy for the borrower's growth opportunities is the ratio of the borrower's market value to its accounting book value, *MKBK*. We control for the maturity of the loan, *MATY*, the borrower's leverage, *LEVG*, unexpected earnings, *UNEXEAR*, credit quality proxied by Z-Score, firm size, *FSIZE*, and loan concentration, *LOANCON*.

The data consists of 2634 revolvers, but data for the dependent variable, *SECURED*, is only available for 1303 observations (51% missing). For the IP

Table 4 Secured status estimates.

Independent variables	Parameter estimates (t-values) Listwise deletion	 Ad hoc imputation	 Imputation posterior
CONS	0.501 (8.92)**	0.309 (6.16)**	0.529 (7.13)**
MATY	0.005 (0.71)	0.001 (0.22)	0.007 (0.69)
MKBK	0.026 (1.99)*	0.056 (4.67)**	0.024 (1.78)
LEVG	0.943 (8.86)**	0.829 (9.98)**	0.878 (7.19)**
UNEXEAR	0.060 (1.30)	0.042 (1.60)	0.049 (1.16)
Z-SCORE	−0.029 (−2.84)**	−0.045 (−4.84)**	−0.029 (−2.42)*
FSIZE	−0.082 (−10.34)**	−0.092 (−13.56)**	−0.086 (−8.48)**
LOANCON	0.070 (3.54)**	0.110 (6.04)**	0.065 (3.24)**

Columns give the parameter estimates with t-values in parentheses of the revolver secured status estimates using listwise deletion, ad hoc imputation, and the IP methods. Explanatory variables are constant (CONS), maturity (MATY), market-to-book value (MKBK), leverage (LEVG), unexpected earnings (UNEXEAR), credit quality (Z-SCORE), firm size (FSIZE), and loan concentration (LOANCON). Data source is Dennis, Nandy, and Sharpe (2000).

*, ** Statistical significance at the 95% and 99% confidence levels, respectively.

estimates, the imputation model includes the variables in the analysis model plus the following: commitment fee (*COMM*), London interbank offer rate (*LIBOR*), earnings volatility (*VAR*), syndication indicator variable (*SYND*), multiple tranche indicator variable (*MULT*), presence of term loan in deal indicator variable (*TMLN*), loan purpose indicator variables (*PNRM*, *PACQ*, and *PMSC*), government bond interest rate volatility (*SDGB*), tax paid to total assets ratio (*TAXA*), all-in-spread (*SPRD*), asset maturity (*AMAT*), squared Z-Score (*ZSQD*), and a post-1993 indicator variable (all variables drawn from Dennis, Nandy, and Sharpe [2000]). All the variables are complete except the commitment fee, which has 33% missingness.

Initially the Little (1988) test for MCAR missingness was applied to the data resulting in a d^2-statistic of 830.30, which is highly significant (at a 99% critical value of $\chi^2_{65} = 94.4$), allowing us to reject the null of MCAR missingness. In Table 4 we report three sets of logit regression estimates. The first is the listwise deletion estimate based on $N = 1303$ observations, while the ad hoc imputation estimate in the second column uses the full sample of $N = 2634$ observations, but with zero values imputed for missing values of the secured status variable. The third column reports results using the multiple imputation IP method with $m = 10$ imputed datasets. Using Equations (16) and (17), the efficiency of the individual IP parameter estimates lie between 94% and 98%.

With the missingness confined to the dependent variable, we find, not surprisingly, that the listwise deletion coefficients are relatively close to those of the IP model. The proxy for growth opportunities has a positive coefficient in both methods, consistent with the prediction from agency cost theory. However, the t-statistics of the listwise deletion estimates are consistently greater in absolute value than those of the multiple imputation estimates. This feature is critically important in drawing inferences regarding our central hypothesis. Thus the null that firms with greater growth opportunities are just as likely to borrow on a secured than unsecured basis can be rejected at the 95% confidence level for the listwise deletion estimates, whereas the IP coefficient estimate would not be considered significant at traditional confidence levels in empirical finance.

In the ad hoc estimates the effect on the *t*-statistics is further magnified as a result of the increased sample together with the implicit assumption that the incomplete observations are known to be zero (unsecured loans) with perfect certainty. The estimated coefficient in the ad hoc method is again positive, consistent with the theory, but is twice the size of the corresponding listwise deletion and IP estimates and with a significantly inflated *t*-statistic of 4.67 compared to 1.78 for the IP method. Thus for this illustration the ad hoc method very considerably overstates the statistical significance of the estimated results.

4 CONCLUSION

The practice observed in many financial applications of dropping observations with incomplete information is, at best, inefficient and, at worst, influences inference. The approach advocated in this article is based on incomplete data imputation. Vastly improved imputation methods are becoming available in statistical software packages for incomplete datasets, but their application has so far largely been restricted to the statistics literature. This is lamentable given the scope for application in the finance literature, where incomplete data is common and where many cross-sectional studies and some time-series/panel data studies satisfy the necessary assumptions of the multiple imputation methods. Even in cases where the incomplete data are missing completely at random (MCAR), more efficient parameter estimates can be gained by performing complete-data analysis. However, when the missing values can be related to the (complete) explanatory variables (i.e., the MAR scheme), formal imputation methods become imperative (particularly when the level of missingness in the data exceeds 5%). Through Monte Carlo experiments for single- and simultaneous-equations models with either continuous or discrete incomplete variables, the superiority of the IP and EM-is multiple imputation methods, vis-à-vis excluding incomplete observations, is demonstrated for MAR missingness in terms of reduced bias and improved efficiency. Moreover, when the techniques are applied to two financial applications, the imputation methods strongly outperform the listwise (pairwise) deletion and ad hoc imputation approaches commonly used in the finance literature in terms of reducing bias, correctly identifying statistically significant parameter estimates, and in predicting missing values.

The multiple imputation techniques may be applied in areas within finance where there is a significant degree of data incompleteness and where the data are likely to meet the MAR (MCAR) requirement. For example, studies using survey data relating to investment preferences, financial decision making, and small business finance typically have high nonresponse rates to some questions. Moreover, frequently used databases in corporate finance and banking studies, such as Compustat, Global Vantage, Compact Disclosure, Dealscan, and SDC Platinum, have relatively high missingness for series where the source materials, whether they be annual reports or security registration statements, involve differential disclosure across the reporting identities. Thus annual data on R&D expenditures, discretionary accruals, and nondebt tax shields is frequently incomplete, while loan/debt contract information on the secured status, maturity, elements of loan pricing, covenants, and aspects of underwriting arrangements are often missing. Similarly market-based measures of firm value and risk are frequently incomplete because of non- or infrequent trading of the underlying securities. In many of these examples, the incomplete data is likely to be MAR, with the occurrence of missingness associated with values of relatively complete variables such as firm size, SIC industry codes, leverage ratios, location, etc. For these cases, the multiple imputation methods offer the potential of significantly improved estimates with less bias and greater efficiency.

The multiple imputation methods examined in this article are computationally fast, are simple to apply, and are intuitively appealing. However, these methods are not magic and cannot create data where no observation (besides the missing values) exists. Sometimes, specific solutions exist for these missing data problems, like latent variable estimation in Markov-switching models and the survivorship bias treatment. These specific solutions are clearly preferred to the multiple imputation methods that have little to offer when confronted with NI missingness. 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. Survivorship bias and censored observations are clear examples of NI missingness. Less obvious cases might include surveys that require disclosure of sensitive information, like executive compensation studies. In the examples used in this study, the ignorability assumption (MAR or MCAR) seems reasonable.

Careful application of the incomplete data imputation methods discussed in this article 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 of datasets previously discarded because of severely incomplete data. The application of the multiple imputation method to the historical NYSE database illustrates the feasibility of such studies.

APPENDIX: A MONTE CARLO COMPARISON OF IMPUTATION METHODS

This appendix contains a description of the design of a set of Monte Carlo experiments examining the efficiency and bias of the imputation methodologies vis-à-vis listwise/pairwise deletion (see Table A.1). The experiments relate to both single-equation and simultaneous-equations models with both continuous and discrete variables with incomplete observations.

Table A.1 A Monte Carlo comparison of imputation methods.

Tables A.2 to A.5 present the mean bias and mean percentage inefficiency of the Listwise (Pairwise) deletion, IP, and EM-is estimators for different parameter values, β_2, α_{11} $(=\alpha_{12})$, and different degrees of missingness, κ. Mean bias is measured as the Monte Carlo average of the sum of the absolute bias in the parameter estimates. Mean percentage inefficiency is measured as the Monte Carlo average of the inefficiency of the parameter estimates.

Table A.2 presents the mean bias and mean percentage inefficiency results for the parameter estimates of the single-equation model:

$$y = \beta_1 + \beta_2 x + e,$$

where y and x are continuous variables. Missingness occurs in the x variable.

Table A.2 presents the mean bias and mean percentage inefficiency results for the parameter estimates of the single-equation model:

$$y = \beta_1 + \beta_2 x + e,$$

where y is a continuous variable and x is a discrete explanatory variable. Missingness occurs in the x variable.

Table A.3 presents the mean bias and mean percentage inefficiency results for the parameter estimates of the simultaneous-equations model:

$$y_1 = \alpha_{01} + \alpha_{11} y_2 + \alpha_{21} x_1 + \varepsilon_1$$
$$y_2 = \alpha_{02} + \alpha_{12} y_1 + \alpha_{22} x_2 + \varepsilon_2,$$

where y_1, y_2, x_1, and x_2 are all continuous variables. Missingness occurs in the y_1, y_2 variables.

Table A.4 presents the mean bias and mean percentage inefficiency results for the parameter estimates of the simultaneous-equations model:

$$y_1 = \alpha_{01} + \alpha_{11} y_2 + \alpha_{21} x_1 + \varepsilon_1$$
$$y_2 = \alpha_{02} + \alpha_{12} y_1 + \alpha_{22} x_2 + \varepsilon_2,$$

where y_1 is a continuous dependent variable, y_2 is a discrete dependent variable, and x_1 and x_2 are continuous explanatory variables. Missingness occurs in the y_1, y_2 variables.

A.1 A Single-Equation Model

Consider first a complete dataset based on the single-equation model $y = \beta_1 + \beta_2 x + e$. One thousand values for a single explanatory variable are randomly drawn from a standard normal distribution and, for known regression parameters $\beta_1 = 0$ and $\beta_2 = 0$, 0.2, 0.4, 0.6, and 0.8, and independent and standard normally distributed (*i.i.d.*) innovations e, 1000 generalizations of y are computed. The Monte Carlo experiments are conducted for two different types of variable x, continuous and discrete (0/1), respectively. The discrete variable is obtained by rounding the cumulative density evaluated at the randomly drawn continuous value. The latter is particularly relevant for the application in Section 3.

Two missing value datasets were then created from this complete dataset. The first is based on the MCAR scheme, where a fraction κ (0.05, 0.25, 0.45) of the x 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 the fraction κ of the x values are eliminated based on a missingness function $g(y)$. Thus missingness in x depends on the values of variable y.

The models are then estimated based on listwise deletion, the IP and the EM-is imputation methods, respectively. For the purpose of comparison, only the preferred imputation methodologies (IP and EM-is) and the most common naive (listwise/pairwise deletion) imputation methodology are presented here. Results for the basic EM algorithm are available from the authors upon request. For each combination of missingness and β_2, the multiple imputation analysis is repeated 1000 times, each consisting of $m = 10$ imputations. For the IP method a "burn-in" period of 500 iterations was chosen, after which an imputation is "taken" at intervals of 100 iterations. Examination of the autocorrelation plots illustrated that convergence occurs quickly for these controlled experiments. Bias and inefficiency are computed as Monte Carlo means from the simulated parameter estimates for the listwise deletion, the IP and EM-is "completed" regression of variable y on variable x. The mean absolute bias is measured as

$$\text{bias} = \frac{1}{R}\sum_{r}\sum_{i=1,2} \left|(\hat{\beta}_i)_r - (\hat{\beta}_i^{TRUE})_r\right|, \tag{A.1}$$

where $r = 1, \ldots, R$ indicates the Monte Carlo run where $R = 1000$. The mean percentage inefficiency is measured as

$$\% - \text{inefficiency} = \frac{1}{2R}\sum_{r}\sum_{i=1,2} 100^{*}\left(\frac{se(\hat{\beta}_i)_r - se(\hat{\beta}_i^{TRUE})_r}{se(\hat{\beta}_i^{TRUE})_r}\right). \tag{A.2}$$

Note that the mean absolute bias measures the sum total bias over all parameter estimates, but mean percentage inefficiency measures the average inefficiency per parameter estimate.

The results are summarized in Table A.2, for a continuous variable x, while the discrete variable results are given in Table A.3. For the continuous variable and MCAR missingness, listwise deletion, IP, and EM-is methods produce estimates

Table A.2 Single equation, continuous variable.

		Listwise deletion estimator			IP estimator			EM-is estimator		
Missingness type	β_2	$\kappa = 0.05$	$\kappa = 0.25$	$\kappa = 0.45$	$\kappa = 0.05$	$\kappa = 0.25$	$\kappa = 0.45$	$\kappa = 0.05$	$\kappa = 0.25$	$\kappa = 0.45$
Mean percentage inefficiency										
MCAR	0	2.20	17.87	36.06	1.01	9.88	17.78	1.32	9.94	18.29
	0.2	3.49	15.27	32.86	2.11	7.95	14.64	2.21	8.02	14.00
	0.4	2.21	14.74	36.11	1.00	6.24	17.48	1.21	6.79	16.65
	0.6	1.92	17.76	33.67	0.45	10.61	15.93	0.41	9.86	17.68
	0.8	2.74	16.23	33.81	1.79	11.98	22.44	1.78	10.82	18.65
MAR	0	5.24	32.88	77.53	0.02	0.39	1.47	0.02	0.37	1.59
	0.2	5.14	31.55	73.25	0.29	1.18	3.61	0.26	1.29	3.47
	0.4	4.58	27.93	64.06	0.58	3.86	9.59	0.56	3.57	7.95
	0.6	3.76	22.71	48.59	1.20	7.19	16.62	1.27	6.68	15.76
	0.8	2.56	15.44	33.41	1.27	8.60	21.14	1.39	8.19	19.13
Mean absolute bias										
MCAR	0	0.0008	0.0028	0.0010	0.0000	0.0007	0.0027	0.0003	0.0023	0.0001
	0.2	0.0004	0.0017	0.0032	0.0006	0.0012	0.0051	0.0004	0.0014	0.0033
	0.4	0.0008	0.0026	0.0039	0.0013	0.0024	0.0080	0.0016	0.0023	0.0086
	0.6	0.0015	0.0022	0.0041	0.0027	0.0041	0.0089	0.0028	0.0050	0.0083
	0.8	0.0011	0.0020	0.0015	0.0011	0.0029	0.0066	0.0014	0.0030	0.0078
MAR	0	0.0002	0.0050	0.0014	0.0001	0.0002	0.0010	0.0001	0.0004	0.0010
	0.2	0.0057	0.0305	0.0686	0.0005	0.0007	0.0021	0.0006	0.0004	0.0023
	0.4	0.0089	0.0547	0.1164	0.0004	0.0007	0.0012	0.0005	0.0009	0.0010
	0.6	0.0110	0.0563	0.1040	0.0010	0.0041	0.0025	0.0007	0.0033	0.0046
	0.8	0.0087	0.0431	0.0843	0.0011	0.0055	0.0113	0.0013	0.0055	0.0127

Table A.3 Single equation, discrete variable.

		Listwise deletion estimator			IP estimator			EM-is estimator		
Missingness type	β_2	$\kappa = 0.05$	$\kappa = 0.25$	$\kappa = 0.45$	$\kappa = 0.05$	$\kappa = 0.25$	$\kappa = 0.45$	$\kappa = 0.05$	$\kappa = 0.25$	$\kappa = 0.45$
Mean percentage inefficiency										
MCAR	0	2.55	16.30	35.23	2.55	13.46	24.60	2.39	10.07	20.02
	0.2	2.37	14.88	34.06	1.79	10.93	22.81	2.25	12.13	15.17
	0.4	2.51	14.57	35.24	1.98	11.36	22.02	2.15	9.62	20.69
	0.6	2.93	15.19	34.10	2.98	11.65	25.02	2.79	11.37	22.59
	0.8	2.76	15.76	33.58	3.84	15.34	29.68	3.22	15.25	30.87
MAR	0	5.30	33.04	77.06	0.05	0.55	2.11	0.05	0.46	2.26
	0.2	5.23	32.32	75.24	0.05	0.47	2.66	0.06	0.49	2.42
	0.4	4.97	30.22	70.48	0.10	0.87	3.24	0.11	0.95	3.81
	0.6	4.42	26.56	58.20	0.25	1.69	5.54	0.22	1.86	5.23
	0.8	3.46	18.38	34.10	0.46	2.98	10.61	0.46	3.31	8.93
Mean absolute bias										
MCAR	0	0.0024	0.0068	0.0034	0.0012	0.0049	0.0012	0.0009	0.0077	0.0005
	0.2	0.0022	0.0021	0.0023	0.0029	0.0215	0.0382	0.0032	0.0220	0.0309
	0.4	0.0034	0.0073	0.0339	0.0098	0.0393	0.0496	0.0092	0.0398	0.0466
	0.6	0.0049	0.0052	0.0064	0.0156	0.0583	0.1004	0.0152	0.0550	0.0974
	0.8	0.0018	0.0062	0.0200	0.0167	0.0728	0.1458	0.0166	0.0717	0.1450
MAR	0	0.0005	0.0024	0.0085	0.0003	0.0002	0.0000	0.0003	0.0008	0.0004
	0.2	0.0118	0.0795	0.1650	0.0002	0.0001	0.0071	0.0003	0.0001	0.0075
	0.4	0.0249	0.1552	0.3433	0.0000	0.0010	0.0081	0.0001	0.0006	0.0080
	0.6	0.0373	0.2281	0.5092	0.0003	0.0027	0.0167	0.0002	0.0028	0.0149
	0.8	0.0500	0.2900	0.6555	0.0001	0.0088	0.0298	0.0002	0.0094	0.0278

with little bias. However, all three estimators are somewhat inefficient at high levels of missingness, though listwise deletion is always more inefficient (less efficient) than IP and EM-is methods. On the other hand, the superiority of IP and EM-is methods is more apparent in the case of MAR missingness, where they have negligible bias (maximum 0.0301 for IP) compared to listwise deletion (maximum 0.2488) and exhibit much less inefficiency (maximum of 20% for EM-is and 77% for listwise deletion). Although the direction of bias is not evident from the table, unreported results indicate that the constant term is biased downward and the slope parameter is biased upward, suggesting a stronger relationship between x and y than the true relationship. Moreover, the bias in listwise deletion increases directly with the level of missingness. While inefficiency increases with missingness for all three methods, inefficiency is negatively related to the regression slope (β_2) for listwise deletion but is positively related to this regression slope for the IP and EM-is methods.

The discrete variable results in Table A.3 mirror those of the continuous variable, with IP and EM-is methods producing large improvements in both bias and efficiency for MAR missingness. However, when missingness is MCAR there is a bias/efficiency trade-off with IP and EM-is methods a little more biased, though more efficient, than listwise deletion.

A.2 A Simultaneous-Equations Model

Next, the single-equation Monte Carlo experiments are extended to a simultaneous-equations model of the form

$$\begin{aligned} y_1 &= \alpha_{01} + \alpha_{11} y_2 + \alpha_{21} x_1 + \varepsilon_1 \\ y_2 &= \alpha_{02} + \alpha_{12} y_1 + \alpha_{22} x_2 + \varepsilon_2. \end{aligned} \tag{A.3}$$

Estimating Equation (A.3) directly by applying ordinary least squares (OLS) would generate inconsistent estimates in view of the correlation between the stochastic explanatory variable y_2 (y_1) and the stochastic disturbance term ε_1 (ε_2). Instead, when a proxy for y_2 (y_1) can be found that is similar — in the sense of retaining its correlation with y_1 (y_2) — but uncorrelated with ε_1 (ε_2), then OLS can be applied to estimate Equation (A.3). One way to find such a proxy is using two-stage least squares (TSLS). This is somewhat more complicated when one dependent variable is continuous and the other is discrete. The two-stage procedure of Nelson and Olson (1978), also used by Dennis, Nandy, and Sharpe (2000), is used in here. First, the reduced-form equations are estimated and instrumental variables for y_1 and y_2 are formed. Then the structural equations are estimated using these instruments ($\hat{y}_1$,$\hat{y}_2$) as regressors for y_1, y_2 on the right-hand side of Equation (A.3). In the Monte Carlo experiment, values for y_1 and y_2 are generated for x_1, x_2, u_1, u_2 i.i.d. $N(0,1)$, $\varepsilon_1 = \alpha_{11} u_2 + u_1$, and $\varepsilon_2 = \alpha_{12} u_1 + u_2$. The stochastic explanatory regressors are generated based on the reduced form for Equation (A.3). The parameter values are chosen such that $\alpha_{01} = \alpha_{02} = 0.1$, $\alpha_{11} = \alpha_{12} = 0$, 0.2, 0.4, 0.6, and 0.8, $\alpha_{21} = 0.3$, and $\alpha_{22} = 0.7$. The experiment is repeated for the case where one dependent variable (y_2) is a discrete (0/1) variable. Now, TSLS estimation is based on a combination of OLS and probit.

A fraction κ (0.05, 0.25, 0.45) of the dependent variables values are eliminated by MCAR and MAR, respectively. Missingness occurs in both dependent variables, but not in the explanatory variables. Missingness in y_1 (y_2) depends on the values for x_1 (x_2). The models that are reported below are based on pairwise deletion and the IP and EM-is methods. As before, simulations are based on a range of missingness fractions as well as a range of structural parameter values relating the dependent variables (i.e., α_{11} and α_{12}). The IP method uses a "burn-in" period of 500 iterations. The simultaneous-equations model estimation procedure is replicated 1000 times, and the results illustrating the mean absolute bias and mean percentage inefficiency in the parameter estimates are summarized in Table A.4 for the case where both dependent variables are continuous and in Table A.5 for the case where one dependent variable is continuous and the other discrete. The bias and efficiency measures are computed as the Monte Carlo means for the simulated TSLS parameter estimates of the pairwise deletion, IP, and EM-is methods. The mean absolute bias is measured as

$$\text{bias} = \frac{1}{R}\sum_{r}\sum_{j=1}^{2}\sum_{i=0}^{2}\left|(\hat{\alpha}_{ij})_r - (\hat{\alpha}_{ij}^{TRUE})_r\right|, \tag{A.4}$$

where $r = 1, \ldots, R$ indicates the Monte Carlo run, while mean percentage inefficiency is given by

$$\% - \text{inefficiency} = \frac{1}{6R}\sum_{r}\sum_{j=1}^{2}\sum_{i=0}^{2} 100 * \left(\frac{se(\hat{\alpha}_{ij})_r - se(\hat{\alpha}_{ij}^{TRUE})_r}{se(\hat{\alpha}_{ij}^{TRUE})_r}\right). \tag{A.5}$$

As before, mean absolute bias measures the sum total bias over all parameter estimates, but mean percentage inefficiency measures the average inefficiency per parameter estimate.

For MCAR missingness and continuous dependent variables, the results in Table A.4 mirror the single-equation results. Both methods produce relatively unbiased estimates, though the IP and EM-is estimators are considerably more efficient. Again, it is in the MAR experiments where IP and EM-is methods produce markedly superior estimates with greater efficiency and generally less bias than pairwise deletion. However, the comparative advantage of the IP and EM-is methods over pairwise deletion is somewhat less in the simultaneous-equations model. The coefficients of the exogeneous, x, variables are biased upward, as in the single-equation results, while those of the endogenous variables are biased downward. Thus the pairwise deletion method suggests a stronger relationship between the x and y variables, and weaker evidence of simultaneity, than in the true relationship.

When one of the dependent variables is discrete, the results as reported in Table A.5 are similar to the single-equation experiments. With MCAR missingness, pairwise deletion is slightly less biased but generally less efficient than IP

Table A.4 Simultaneous equations, all variables continuous.

		Pairwise deletion estimator			IP estimator			EM-is estimator		
Missingness type	$\alpha_{11} = \alpha_{12}$	$\kappa = 0.05$	$\kappa = 0.25$	$\kappa = 0.45$	$\kappa = 0.05$	$\kappa = 0.25$	$\kappa = 0.45$	$\kappa = 0.05$	$\kappa = 0.25$	$\kappa = 0.45$
Mean percentage inefficiency										
MCAR	0	2.71	16.99	33.04	1.59	9.07	14.30	1.66	7.64	14.39
	0.2	2.11	16.60	34.94	1.33	8.92	15.14	1.11	8.61	15.36
	0.4	2.15	14.29	35.87	1.48	8.77	16.40	1.50	8.40	16.66
	0.6	2.82	15.08	35.31	2.07	9.30	19.41	2.06	9.15	19.19
	0.8	2.79	14.00	39.74	2.30	11.35	22.32	2.26	11.21	26.05
MAR	0	2.70	16.79	40.17	1.20	7.88	12.75	1.44	6.96	12.36
	0.2	2.69	16.56	45.79	1.40	7.19	17.59	1.51	7.54	15.77
	0.4	2.85	15.53	41.19	1.69	7.97	15.10	1.83	7.63	15.68
	0.6	2.69	15.35	36.79	1.61	9.77	17.52	1.77	10.11	17.01
	0.8	2.49	13.46	35.80	2.19	10.95	22.47	2.30	10.52	20.97
Mean absolute bias										
MCAR	0	0.0021	0.0074	0.0017	0.0013	0.0058	0.0040	0.0014	0.0058	0.0028
	0.2	0.0013	0.0020	0.0105	0.0011	0.0025	0.0051	0.0014	0.0017	0.0048
	0.4	0.0020	0.0042	0.0041	0.0009	0.0010	0.0062	0.0013	0.0020	0.0085
	0.6	0.0014	0.0065	0.0076	0.0006	0.0017	0.0042	0.0006	0.0022	0.0020
	0.8	0.0015	0.0101	0.0020	0.0008	0.0030	0.0028	0.0009	0.0040	0.0027
MCAR	0	0.0038	0.0244	0.0602	0.0016	0.0048	0.0098	0.0019	0.0049	0.0107
	0.2	0.0040	0.0253	0.0598	0.0011	0.0031	0.0132	0.0010	0.0032	0.0131
	0.4	0.0031	0.0248	0.0594	0.0016	0.0035	0.0101	0.0011	0.0051	0.0101
	0.6	0.0028	0.0225	0.0588	0.0005	0.0041	0.0147	0.0008	0.0039	0.0153
	0.8	0.0008	0.0040	0.0152	0.0002	0.0059	0.0160	0.0005	0.0045	0.0167

Table A.5 Simultaneous equations, one continuous and one discrete dependent variable.

		Pairwise deletion estimator			IP estimator			EM-is estimator		
Missingness type	$\alpha_{11} = \alpha_{12}$	$\kappa = 0.05$	$\kappa = 0.25$	$\kappa = 0.45$	$\kappa = 0.05$	$\kappa = 0.25$	$\kappa = 0.45$	$\kappa = 0.05$	$\kappa = 0.25$	$\kappa = 0.45$
Mean percentage inefficiency										
MCAR	0	3.18	15.37	37.16	2.12	7.90	15.74	1.94	7.57	14.12
	0.2	2.95	15.58	32.70	2.20	7.94	13.53	1.82	8.32	14.95
	0.4	2.13	14.43	39.92	1.53	7.31	18.65	1.56	8.93	17.27
	0.6	2.59	16.04	36.94	2.97	12.50	22.54	2.56	13.01	22.28
	0.8	2.45	15.99	34.37	6.23	28.50	33.96	6.37	28.61	31.19
MAR	0	2.32	15.43	39.99	1.63	7.18	14.07	1.66	7.12	13.46
	0.2	2.44	15.51	40.96	1.64	7.97	15.20	1.50	8.02	14.81
	0.4	2.40	15.94	41.27	1.81	8.57	16.06	1.88	8.70	17.01
	0.6	2.70	17.17	44.39	2.23	11.42	20.29	2.69	10.66	22.11
	0.8	3.56	27.30	57.91	4.78	22.79	36.96	5.11	22.62	35.15
Mean absolute bias										
MCAR	0	0.0010	0.0047	0.0067	0.0019	0.0099	0.0148	0.0026	0.0100	0.0163
	0.2	0.0033	0.0086	0.0088	0.0050	0.0154	0.0184	0.0050	0.0151	0.0179
	0.4	0.0012	0.0059	0.0122	0.0063	0.0207	0.0302	0.0051	0.0217	0.0317
	0.6	0.0008	0.0066	0.0121	0.0066	0.0278	0.0513	0.0060	0.0273	0.0503
	0.8	0.0031	0.0037	0.0060	0.0167	0.0925	0.1317	0.0149	0.0929	0.1356
MAR	0	0.0150	0.0309	0.0648	0.0043	0.0128	0.0187	0.0040	0.0123	0.0205
	0.2	0.0119	0.0342	0.0708	0.0027	0.0161	0.0286	0.0036	0.0161	0.0262
	0.4	0.0126	0.0446	0.0741	0.0047	0.0210	0.0435	0.0048	0.0208	0.0447
	0.6	0.0142	0.0511	0.0843	0.0064	0.0276	0.0705	0.0067	0.0278	0.0727
	0.8	0.0206	0.0701	0.0980	0.0148	0.0842	0.2016	0.0141	0.0829	0.2035

and EM-is methods. However, for MAR, the IP and EM-is methods reassert their relative advantage with less biased and markedly more efficient estimates than pairwise deletion. As for the case with two continuous dependent variables, the pairwise deletion estimates generally improve in the simultaneous-equations model relative to the single-equation estimates, while those based on IP or EM-is methods are generally more biased and less efficient than their single equation counterparts. Thus, while the IP and EM-is imputation methods outperform pairwise deletion in all but the discrete variable case (for bias only) in the presence of MCAR missingness, their relative advantage is reduced in the context of simultaneous-equations estimates.

Received June 19, 2001; revised June 10, 2002; accepted March 6, 2003

REFERENCES

Barclay, M. J., and C. W. Smith, Jr. (1995). "The Maturity Structure of Corporate Debt." *Journal of Finance 50*, 609–631.

Dagenais, M. G. (1976). "Incomplete Observations and Simultaneous-Equations Models." *Journal of Econometrics 4*, 231–241.

Dempster, A. P., N. M. Laird, and D. B. Rubin. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society 39B*, 1–22.

Denis, D. J., D. K. Denis, and A. Sarin (1997). "Agency Problems, Equity Ownership, and Corporate Diversification." *Journal of Finance 52*, 135–160.

Dennis, S., D. Nandy, and I. G. Sharpe (2000). "The determinants of Contract terms in Bank Revolving Credit Agreements." *Journal of Financial and Quantitative Analysis 35*, 87–110.

Goetzmann, W. N., R. G. Ibbotson, and L. Peng (2001). "A New Historical Database for the NYSE 1815 to 1925: Performance and Predictability." *Journal of Financial Markets 4*, 1–32.

Gompers, P. A. (1995). "Optimal Investment, Monitoring, and the Staging of Venture Capital." *Journal of Finance 50*, 1461–1489.

Hamilton, J. D. (1994). *Time Series Analysis*, Princeton, NJ: Princeton University Press.

Himmelberg, C. P., R. G. Hubbard, and D. Palia (1999). "Understanding the Determinants of Managerial Ownership and the Link Between Ownership and Performance." *Journal of Financial Economics 53*, 353–384.

Honaker, J., A. Joseph, G. King, and K. Scheve (2000). "Amelia: A Program for Missing Data (Gauss Version)," Harvard University, available at http://Gking.Harvard.edu/.

Horowitz, J. L., and C. F. Manski (2000). "Non-Parametric Analysis of Randomized Experiments with missing Covariate and Outcome Data." *Journal of the American Statistical Association 95*, 77–84.

Ip, E. H. S. (1994). "A Stochastic EM Estimator in the Presence of Missing Data — Theory and Applications." Technical Report 304, Stanford University.

King, G., J. Honaker, A. Joseph, and K. Scheve (2001). "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." *American Political Science Review 95*, 49–69.

Krolzig, H.-M. (1997). *Markov-Switching Vector Autoregressions—Modelling, Statistical Inference and Applications to Business Cycle Analysis*. Lecture Notes in Economics and Mathematics 454, Berlin: Springer-Verlag.

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

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

McCoskey, S., and C. Kao (1998). "A Residual-Based Test of the Null of Cointegration in Panel Data." *Econometric Reviews 17*, 57–84.

Nelson, F., and L. Olson, (1978). "Specification and Estimation of a Simultaneous-Equation Model with Limited Dependent Variables." *International Economic Review 19*, 695–709.

Perez-Quiros, G., and A. Timmermann (2000). "Firm Size and Cyclical Variations in Stock Returns." *Journal of Finance 55*, 1229–1262.

Phillips, P. C. B., and S. Ouliaris (1990). "Asymptotic Properties of Residual Based Tests for Cointegration." *Econometrica 58*, 165–193.

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

Rangan, S. (1998). "Earnings Management and the Performance of Seasoned Equity Offerings." *Journal of Financial Economics 50*, 101–122.

Rotnitzky, A., J. M. Robins, and D. O. Scharfstein (1998). "Semiparametric Regression for Repeated outcomes with Non-Ignorable Non-Response." *Journal of the American Statistical Association 93*, 1321–1339.

Rubin, D. B. (1974). "Characterizing the Estimation of Parameters in Incomplete Data Problems." *Journal of the American Statistical Association 69*, 467–474.

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

Rubin, D. B. (1978). "Multiple Imputations in Sample Surveys." ASA 1978 Proceedings of the Survey Research Methods Section, 20-34.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*, London: Chapman & Hall.

Servaes, H. (1996). "The Value of Diversification During the Conglomerate Merger Wave." *Journal of Finance 51*, 1201–1225.

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

Tanner, M. A. (1996). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distribution and Likelihood Functions*, 3rd ed. New York: Springer-Verlag.

Trevor, R., and I. Morgan (1999). "Limit Moves as Censored Observations of Equilibrium Futures price in GARCH Processes." *Journal of Business and Economic Statistics 17*, 397–408.