MARKETING SCIENCE INSTITUTE

MSI Reports

Standard-Scape: An Agent-based Model of Competition in Markets with Network Externalities (05-105)
Judy K. Frels, James A. Reggia, and Debra Heisler

How Firm Advertising Affects Employees' Trust, Organizational Identification, and Customer Focus (05-106)
Mary Wolfinbarger and Mary C. Gilly

Handling Missing Values in Marketing Data: A Comparison of Techniques (05-107)

James Lemieux and Leigh McAlister

Mapping the Domain of Subjective Value in Negotiation (05-108)
Jared R. Curhan, Hillary Anger Elfenbein, and Heng Xu

Consumer Testimonials as Self-Generated Advertisements: Evaluative Reconstruction Following Product Usage (05-109)
Terence A. Shimp, Stacy L. Wood, and Laura Smarandescu

2005

WORKING
PAPER
SERIES

ISSUE TWO

NO. 05-002

MARKETING SCIENCE INSTITUTE

MSI Reports

Executive Director
Leigh M. McAlister

Research Director
Ross Rizley

Editorial Director
Susan Keane

Publication Design
Laughlin/Winkler, Inc.

2005

WORKING PAPER SERIES

ISSUE TWO

NO. 05-002

The Marketing Science Institute supports academic research for the development—and practical translation—of leading-edge marketing knowledge on issues of importance to business performance. Topics are identified by the Board of Trustees, which represents MSI member corporations and the academic community. MSI supports academic studies on these issues and disseminates findings through conferences and workshops, as well as through its publications series.

Marketing Science Institute
1000 Massachusetts Avenue
Cambridge, MA
02138-5396

Phone: 617.491.2060
Fax: 617.491.2065
www.msi.org

MSI Reports (ISSN 1545-5041) is published quarterly by the Marketing Science Institute. It is not to be reproduced or published, in any form or by any means, electronic or mechanical, without written permission.

The views expressed here are those of the authors.

Working Paper Series
The articles that appear in *MSI Reports* have not undergone a formal academic review. They are released as part of the MSI Working Paper Series, and are distributed for the benefit of MSI corporate and academic members and the general public.

Subscriptions
Annual subscriptions to *MSI Reports* can be placed online at www.msi.org. Questions regarding subscriptions may be directed to pubs@msi.org.

Single reports
Articles in *MSI Reports* are available in downloadable (PDF) format at www.msi.org.

Past reports
MSI working papers published before 2003 are available as individual hard-copy reports; many are also available in downloadable (PDF) format. To order, go to www.msi.org.

Corporate members
MSI member company personnel receive all MSI reports (PDF and print versions) free of charge.

Academic members
Academics may qualify for free access to PDF (downloadable) versions of MSI reports and for special rates on other MSI print publications. For more information and to apply, go to "Qualify for academic membership" on www.msi.org.

Classroom use
Upon written request, MSI working papers may be copied for one-time classroom use free of charge. Please contact MSI to obtain permission.

Search for publications
See the searchable publications database at www.msi.org.

Submissions
MSI will consider a paper for inclusion in *MSI Reports*, even if the research was not originally supported by MSI, if the paper deals with a priority subject, represents a significant advance over existing literature, and has not been widely disseminated elsewhere. Only submissions from faculty members or doctoral students working with faculty advisors will be considered. "MSI Working Paper Guidelines" and "MSI 2004-2006 Research Priorities" are available in the Research section of www.msi.org.

Publication announcements
To sign up to receive MSI's electronic newsletter, go to www.msi.org.

Change of address
Send old and new address to pubs@msi.org.

Handling Missing Values in Marketing Data: A Comparison of Techniques

James Lemieux and Leigh McAlister

As firms gather and analyze customer data, they must "solve" the problem of missing information. This study compares six techniques and offers guidance on which method will help managers determine the best marketing decisions.

James Lemieux is a Ph.D. student at McCombs School of Business, University of Texas, Austin. **Leigh McAlister** is the 2003-05 MSI Executive Director and the H. E. Hartfelder/The Southland Corporation Regents Chair for Effective Business Leadership, McCombs School of Business, University of Texas, Austin. The order of authorship is alphabetical; both authors contributed equally to the work.

Report Summary

As companies accumulate and analyze data on their customers, they inevitably encounter the problem of missing information. Lemieux and McAlister provide guidelines for managers to use when confronted with this problem. They focus on prescriptive techniques for imputing missing information when there is no special structure to the data.

Using both simulated and real customer data, Lemieux and McAlister compare six imputation techniques—complete case analysis (CCA), hot deck (HD), mean imputation (Mean), expectation maximization (EM), data augmentation (DA), and multiple imputation (MI).

Their results show that CCA—which drops all customers for which there is missing information—should never be used, as it always performs worst. This is an important insight given the fact that CCA is the default treatment for missing information in popular statistical analysis software packages.

The recommended technique depends on the analyst's objective. If the objective is to get the most accurate imputations, or imputations yielding the most accurate estimates of means or covariances, then EM should be used. If the objective is to get the most accurate estimates of variances, then HD should be used. If the objective is to get the most accurate model coefficients or models producing the most accurate model predictions, then Mean should be used.

The results imply that Mean imputation works best when the analyst's objectives are focused on helping managers make good decisions, since accurate model coefficients can be used to assess the impact of marketing actions and accurate model predictions help determine the best marketing decision. This result is a significant contribution to the existing literature since it shows that more sophisticated methods focused on producing accurate intermediate measures (e.g., means, covariances, and variances) do not perform as well as less sophisticated methods for measures most directly applicable to managers. ■

Introduction

As companies accumulate and analyze data on their customers, they inevitably encounter the problem of missing information. Many techniques have been developed to address this problem. In this paper, we review the literature on missing information and provide guidelines for managers to use when confronted with the problem of missing information.

The literature contains two streams of research related to the problem of missing information. The first stream *describes* the way in which consumers behave when confronted with missing information. This descriptive research stream considers the way consumers infer missing product information (Bradlow, Hu, and Ho 2002; Ross and Creyer 1992), the effect of missing information on choice (Kivetz and Simonson 2000), and the way missing information affects consumer inferences and evaluations (Dick, Chakravarti, and Biehal 1990; Johnson and Levin 1985).

The second stream *prescribes* techniques for imputing or handling the missing information. This stream suggests how one should make inferences when confronted with missing information. The literature provides several prescriptive techniques for imputing missing information when the pattern of "missingness" has a particular structure (cf. DeSarbo, Green, and Carroll 1986; Erdem, Keane, and Sun 1999). For example, Kamakura and Wedel (1997) propose techniques for fusing together data from two groups of subjects when there are few variables in common between the groups. Cipra and Trujillo (1995) impute missing information by exploiting structure in time series data. Steenburgh, Ainslie, and Engebretson (2003) and Bronnenberg and Sismeiro (2002) impute missing information by exploiting spatial structure in geographic data.

Prescriptive techniques for dealing with missing information when there is no specific structure to the missingness include techniques designed to estimate a particular type of model and techniques designed to simply impute the missing information. For example, in the presence of missing values, Kamakura and Wedel (2000) provide a technique for estimating a factor model, and DeSarbo, Young, and Rangaswamy (1997) show how to estimate an MDS model.[1]

In this paper, we focus on the final component of the research on missing information: prescriptive techniques for imputing missing information when there is no special structure to the missingness. As have earlier researchers (Gleason and Staelin 1975; Schafer and Graham 2002), we will compare existing imputation techniques to provide guidance in the selection of an imputation technique. Like earlier studies (Gleason and Staelin 1975; Schafer and Graham 2002), we will compare techniques in terms of the accuracy of their imputations, the accuracy of data parameters (i.e., the means, variances, and covariances) implied by the imputations, and the accuracy of the model parameters estimated with imputed data. We go beyond earlier studies to compare techniques in terms of the quality of the decisions implied by the models estimated on imputed data.

Like earlier studies (Gleason and Staelin 1975; Schafer and Graham 2002), we investigate the robustness of the imputation techniques using simulated data. We go beyond existing research to investigate the robustness of the imputation techniques using actual customer data. Like earlier studies, we consider the impact of different factors on the relative performance of imputation techniques. In particular, we vary sample size, average correlation in the data, extent of missing information, and the missing value mechanism (whether the data are missing completely at random [MCAR] or missing at random [MAR]). While existing studies compare techniques by noting differences in accuracy across techniques, we go beyond those studies to provide statistical tests of those differences.

In this paper, we begin by describing the imputation techniques, the criteria on which we compare technique performance, and the dimensions used to test the robustness of performance comparisons. We then discuss our method of analysis and our statistical tests. Finally, we present results from a simulation study and from the analysis of real customer data. We structure the simulated data to be consistent with the real customer data.

We consider the problem of a financial services company who must decide to whom it should offer a credit card. This company has a sample of customers for whom it knows both demographic attributes (predictors) and creditworthiness (dependent variable). Based on this sample, the company must estimate a model that can be applied to readily available demographic information in a new target market in order to assess the (less readily available) creditworthiness of customers in the new market. As is often the case, there are customers in the original sample for whom the financial services company does not have complete demographic information. We compare the imputation techniques on the basis of the accuracy of their imputed values for missing predictors, the accuracy of data parameters (i.e., the predictor means, variances, and covariances) implied by the imputations, the accuracy of the model coefficients estimated using the imputed data, and the quality of the marketing decisions implied by models estimated on the imputed data.

Imputation Techniques

The imputation techniques we consider in this paper include two simple techniques (complete case analysis and Mean), three techniques that exploit the correlation between predictors (expectation maximization, data augmentation, and multiple imputation), and a technique that fills in the missing value for a customer by "borrowing" a non-missing value from a similar customer (hot deck). In this section, we review the techniques and summarize their reported strengths and vulnerabilities.

There are two general ways to handle missing information. The simplest way is to drop all customers with missing information. This approach is known as complete case analysis (CCA) and is the default technique for handling missing information in many popular statistical analysis software packages such as SPSS and SAS. Since this technique throws away information, we should expect it to perform worse than other imputation techniques. Alternatively, the missing information can be "filled in" or "imputed". The oldest and easiest technique for imputing missing information for a particular variable is to use the mean of the non-missing values for that variable (Wilks 1932). Unfortunately, replacing all missing values with the mean will at best put values at the center of a variable's distribution and produce a downwardly biased estimate of a variable's true variance and covariance (Little and Rubin 1987). This problem suggests the application of more sophisticated imputation techniques.

In trying to impute missing information for a predictor, the Mean technique only considers non-missing values for that predictor. If there is correlation between predictors, one might expect that an imputation technique that exploits this correlation would provide a better estimate of the missing information. Gleason and Staelin (1975) use linear regression to exploit predictors' correlations when imputing missing values. They showed that when the average correlation between predictors is less than or equal to .2, the simple technique of mean imputation provides point estimates and estimates of data parameters that are as accurate as those provided by regression.

The Mean technique is known to be vulnerable to the nature of the mechanism causing information to be missing. If the mechanism is such that the probability a value is missing for a customer depends on the values of other predictors for that customer (a pattern known as "missing at random" or MAR), then one would expect that the missing values are more likely to come from one end of a variable's distribution. In this case, the mean of a variable's non-missing values will be a biased estimate of a variable's true mean. To remove this MAR-induced bias, Schafer and Graham (2002) impute missing values using a regression-like technique called expectation maximization (EM). Dempster, Laird, and Rubin (1977) show that the EM algorithm will converge under suitable regularity conditions, but the rate of convergence slows as the amount of missing information increases. When facing a situation with a significant amount of missing information, Schafer (1997) recommends a Bayesian analogue to the EM algorithm known as data augmentation (DA). This technique leverages information contained in a prior probability distribution for the data parameters.

In another extension of EM, Rubin (1987) shows that drawing inferences about parameters estimated on an imputed dataset requires consideration of both the sampling error and the error due to imputation. Multiple imputation (MI) corrects inferences for error due to imputation by estimating *K* imputations for each missing value which results in *K* imputed datasets. A single estimate of a parameter of interest using MI is found by averaging the parameter estimates across the *K* imputed datasets.[2] In principle, MI will work with any imputation technique, but it is most often used in connection with the DA algorithm. Following Schafer and Graham (2002), we implement MI using the DA algorithm.

The final imputation technique we consider is hot deck (HD). This technique uses the values of other predictors when imputing the missing value for a particular predictor variable, but it does not build a regression-like model as do EM, DA, and MI. Rather, if the *i*-th customer is missing a value for the *j*-th predictor, hot deck looks for a "donor" customer most similar to the *i*-th customer (i.e., a customer whose predictor values are closest to those of the *i*-th customer) and "borrows" the donor's non-missing value for the *j*-th predictor. Ford (1983) shows that hot deck provides an excellent estimate of a predictor variance because hot deck's imputations are drawn from a predictor's actual values.

MI was designed to improve upon DA, which in turn was designed to perform better than EM. Hence, we expect MI to perform better than DA and we expect DA to perform better than EM. All of these correlation exploiting techniques (MI, DA, and EM) should perform better than Mean, when the average correlation between predictors is greater than .2, and perform no worse than Mean when the average correlation is less than or equal to .2. Further, we expect CCA to perform worse than all of the other imputation techniques since CCA throws information away. Finally, we expect imputations using HD to provide the most accurate estimates of predictor variances and we expect imputations using Mean to provide the least accurate estimates of predictor variances.

Performance Criteria for Comparing Techniques

We compare imputation techniques using multiple performance criteria. We compare the accuracy of their imputed values, the accuracy of the data parameters (means, variances, and covariances) implied by the imputations, the accuracy of model parameters estimated with the imputed data, and the quality of the marketing decisions implied by models estimated with the imputed data. We begin with a "complete" dataset D (i.e., a dataset for which we know, for every customer, the value of the dependent and predictor variables) and then construct an "incomplete" dataset D_I by deleting values of predictor variables. We will refer to a dataset with deleted values filled in by a particular imputation technique (T) as the technique's "imputed" dataset D_T.

The accuracy of technique T's imputations for a particular predictor variable is estimated by comparing the actual and imputed values for that predictor variable. In particular, for each predictor p_j, we calculate the error between "true" values in D and the imputed values in D_T. We summarize technique T's imputation accuracy for predictor p_j by calculating the mean absolute error across all missing values of p_j, and denote this error by $mae_{\text{values}}(D_T, p_j)$.

The accuracy of estimated predictor means and variances (data parameters) are calculated in a similar manner. For each predictor p_j, we calculate the data parameter in technique T's imputed dataset D_T. This value is compared to the "true" data parameter for p_j calculated in complete dataset D. We summarize technique T's ability to reproduce a data parameter for p_j in D by calculating the absolute error for this data parameter between D and D_T. We denote the error for means and variances on predictor p_j by $ae_{mean}(D_T, p_j)$, and $ae_{variance}(D_T, p_j)$, respectively.

The accuracy of the estimated covariances for predictor p_j is calculated by considering all terms $\text{cov}(p_j, p_k)$, for $j \neq k$ in the imputed dataset D_T. These values are compared to the "true" covariances for p_j in the complete dataset D. We summarize technique T's ability to reproduce covariances for p_j in D by calculating the mean absolute error between D and D_T across all covariances terms for p_j. This error is denoted by $mae_{covariance}(D_T, p_j)$.

Technique T's imputed dataset D_T can be used to estimate a logit model that relates a customer's demographic descriptors to his or her creditworthiness. The accuracy of the coefficient estimate associated to predictor p_j from this logit model is calculated by comparing the estimate derived using the imputed dataset D_T to the "true" coefficient value estimated on the complete dataset D. We summarize technique T's ability to reproduce the logit coefficient for p_j in D by calculating the absolute error for this coefficient between D and D_T, and denote this error by $ae_{coefficient}(D_T, p_j)$.

To estimate the quality of the marketing decision implied by a logit model estimated on technique T's imputed dataset D_T, we need to know the number of "hits" and "misses" implied by the estimated model. The "hits" are the number of creditworthy people to whom a credit card is offered and the number of non-creditworthy people to whom a credit card is not offered. The "misses" are the number of creditworthy people not offered a credit card and the number of non-creditworthy people offered a credit card. To estimate "hits and misses," a score is obtained for each customer in D_T by inserting a customer's demographic descriptors (some of which may be imputed) into the logit model estimated using D_T. This customer-specific score is often compared to a managerially determined cutoff. If a customer's score is above the cutoff, the model suggests that the customer is likely to be creditworthy and should be issued a credit card. If a customer's score is below the cutoff, the model suggests that the customer is likely to *not* be creditworthy and should therefore *not* be issued a credit card.

We evaluate the quality of the decisions implied by an estimated model without referring to a single, managerially determined cutoff. Rather, we use the area under a Receiver Operator Characteristic (ROC) curve that summarizes hits and misses across all possible cutoff values (Faraggi and Reiser 2002). The ROC methodology is derived from signal detection theory where it is used to determine if an electronic receiver is able to able to distinguish between signal and noise (Green 1966). The area under an ROC curve is often used to evaluate the predictive accuracy of discrete-valued model, as in models that determine a managerial decision such as whether to offer someone a new product (Bult and Wansbeek 1995). Higher values for this area are associated with "better" (more "hits" and/or fewer "misses") marketing decisions being implied by the logit model. Let $ROC(D_T)$ be the area under the ROC curve for the logit model estimated using dataset D_T imputed by technique T, and let $ROC\ (D)$ be the area under the ROC curve for the logit model estimated on complete dataset D. We summarize the quality of marketing decisions implied by technique T's imputation for dataset D by calculating the absolute error $|ROC(D) - ROC(D_T)|$, and denote this error by $ae_{ROC}(D_T)$.

Technique Performance Robustness

In order to understand the robustness of the technique performance comparisons, we compare techniques within different "environments." These environments are defined by the way complete datasets are constructed and by the way values are deleted to construct the incomplete datasets. We manipulate four characteristics, or "robustness dimensions," shown to effect imputation technique performance in past studies. Two of these robustness dimensions relate to the way a complete dataset is constructed: sample size (i.e., number of customers in the dataset) and average correlation between predictors. The other two robustness dimensions relate to how values are deleted in constructing the incomplete datasets: proportion of missing information and the missing value mechanism (i.e., whether the probability a predictor value is missing for a particular customer depends on the values of other predictor variables for that customer).

Sample size

Schafer and Graham (2002) suggest that EM and the EM-like imputation techniques (DA and MI) are best for small sample applications. Following their guidelines we consider both a "small" sample (100 customers) and a "large" sample (250 customers) in our simulation. In the real customer data, we also include a "very large" sample (8,940 customers).

Correlation between predictors (simulation only)

Gleason and Staelin (1975) show the importance of considering the correlations between predictors when evaluating different imputation techniques. Their study suggests that Mean may perform as well as the correlation-exploiting techniques (MI, DA, and EM, in this study) when the average level of correlation between predictors is .2 or less. In the simulation analysis, we consider five values of average correlation between predictors (average correlation = .1, .2, .3, .4, .5). For the analysis of real customer data, we note that the average level of correlation between predictors in the small datasets ($N = 100$) is .24, in the medium datasets ($N = 250$) is .16, and in the very large dataset ($N = 8940$) is .12. It is important to note that such low levels of average correlation are common when considering demographic variables.

Proportion missing

Incomplete datasets are constructed that range from very little information missing to a large proportion of information missing. In particular, we consider six different levels for the proportion of information missing: 5%, 10%, 20%, 30%, 40%, and 50%.

Missing value mechanism

Researchers have considered three mechanisms that create patterns of missing information like those observed in practice (Schafer and Graham 2002). The most straightforward of these mechanisms, "missing completely at random" (MCAR), occurs when the probability a particular value is missing for a customer is independent of all other predictor values for that customer. A more general mechanism, "missing at random" (MAR), occurs when the probability a particular value is missing for a customer is not independent of the other predictor values for that customer. For example, this might occur if customers with more education are less likely to report their income level than customers with less education. In this case, the probability that income is missing for any particular customer depends on the education level of the customer. A more complex mechanism, "missing not at random" (MNAR), occurs when a specific, non-random process causes values to be missing. In this case, one must specify a separate model for the MNAR process in order to impute missing values.

We consider only MCAR and MAR missing value mechanisms since we have no reason to posit a specific non-random process causing missing information in the customer dataset under consideration. The MCAR mechanism is implemented by assigning each value in the complete dataset an equal probability of being deleted. The MAR mechanism is implemented by employing the decision rule for deletion found in Little (1992). In particular, for each customer and predictor variable, we take the sum of the other predictors for that customer. If the sum exceeds a predefined cutoff value, then the predictor variable for that customer is considered for deletion.

Method of Analysis

Simulation

We wish to compare all pairs of imputation techniques on all levels of all robustness dimensions. We compare techniques on each of the two sample sizes (N = 100 and N = 250), five levels of correlation between predictors (.1, .2, .3, .4, .5), six levels of proportion missing (5%, 10%, 20%, 30%, 40%, 50%), and two missing value mechanisms (MCAR and MAR). Hence, we simulate 2 x 5 x 6 x 2 = 120 different complete datasets (one dataset for each unique combination of levels of robustness dimensions).

To make the simulated data as consistent as possible with the real customer data, simulated datasets include five predictors (as we have in the real customer data). Values for those five predictors are drawn from a multivariate normal distribution with means equal to zero and variances equal to one (to be consistent with the standardized variables in the real customer data). The covariance structure is set so every pair of predictors has a correlation equal to the average correlation level associated with the dataset. The value for the dependent variable is constructed by taking the sum of an equally weighted predictor and adding a random value drawn from a logistic distribution with location parameter equal to zero and scale parameter equal to one. The resulting sum is transformed to equal one when the sum is greater than zero (indicating the customer is creditworthy) and transformed to equal zero when the sum is less than zero (indicating that this customer is not creditworthy).

For each imputation technique (T) and imputed dataset (D_T), one value of $ae_{ROC}(D_T)$ is calculated and five values (corresponding to each predictor) of the other error measures are calculated: $mae_{values}(D_T, p_j)$, $ae_{mean}(D_T, p_j)$, $ae_{variance}(D_T, p_j)$, $mae_{covariance}(D_T, p_j)$, and $ae_{coefficient}(D_T, p_j)$. Each performance criterion is analyzed in two ways. First, an ANOVA model is estimated with the dependent variable being the performance criterion and the independent variables being imputation technique (MI, DA, EM, MEAN, HD, CCA) and robustness dimensions (sample size, average correlation between predictors, proportion of missing information, and missing value mechanism). Since our primary interest is in the relative performance of the different imputation techniques, the ANOVA model has all main effects and all two-way interactions with the technique factor. The ANOVA indicates: (1) if there are statistically significant differences between techniques on the performance criterion (i.e., if there is a significant main effect for technique), (2) if performance differences change in statistically significant ways across different levels of a robustness dimension (i.e., if there are statistically significant interactions between technique and a robustness dimension).

The ANOVA does not indicate if a difference between any two particular techniques on a performance criterion is statistically significant. To look for statistically significant differences between the six techniques for a particular performance criterion, we consider post hoc contrast for the (6*5)/2 = 15 technique pairs.[3] To consider interactions, we check for a statistically significant difference on a performance criterion between two techniques on a particular level of a robustness dimension using a post hoc contrast. Since there are 15 different levels of robustness dimensions (two sample sizes, five levels of average correlation, six proportions of missing information and two missing value mechanisms), and 15 possible pairs of the six imputation techniques, there are 225 possible contrasts for each of the six performance criterion, yielding a total of 1,265 tests related to interactions.[4] Since these contrast tests are correlated, we use the conservative Scheffé adjustment to assess statistical significance (Lomax 2001).[5]

Real customer data

The real customer data contain five demographic predictors shown to be good indicators of creditworthiness: home ownership, length of residence, number of children, number of adults, and income (Sullivan and Fisher 1988; Black and Morgan 1998). Since it is common in the credit card industry to evaluate someone as creditworthy if he or she has fewer than three delinquent credit card accounts (Lawrence 1992), we set the dependent variable to the value one if the customer has fewer than three delinquent credit card accounts and to zero, otherwise. We begin with a complete dataset containing information on 8,940 customers. Thirty-six incomplete datasets are constructed to reflect a full factorial combination of three sample sizes (100, 250, 8,940), six proportions of missing information (5%, 10%, 20%, 30%, 40%, 50%), and two missing value mechanisms (MCAR and MAR). We do not manipulate average correlation because these data are not generated as in the simulation. The average level of correlation between predictors in these datasets ranges from .12 to .24.

As we did with the simulated data, a separate ANOVA model is estimated for each performance criterion. Each performance criterion is a dependent variable with main effects for imputation technique and robustness dimensions (sample size, proportion of missing information, and missing value mechanism) and with all two-way interactions that include the technique factor. To look for statistically significant overall differences between the six techniques for a particular performance criterion, we consider post hoc contrast for the (6*5)/2 = 15 technique pairs.[6] To consider interactions, we check for a statistically significant difference on a performance criterion between two techniques on a particular level of a robustness dimension using a post hoc contrast. Since there are 11 different levels of robustness dimensions (three sample sizes, six proportions of missing information and two missing value mechanisms), and 15 possible pairs of the six imputation techniques, there are 165 possible contrasts for each of the six performance criterion, yielding a total of 925 tests related to interactions.[7] Since these contrast tests are correlated, we use the conservative Scheffé adjustment to assess statistical significance.

Simulated Data Results

Each column in Table 1 represents an ANOVA model whose dependent variable is the performance criterion listed at the top of the column and whose independent variables are technique (imputation technique = MI, DA, EM, Mean, HD, CCA), sample size (N = 100, N = 250), correlation (average level of correlation between predictors = .1, .2, .3, .4, .5), proportion missing (proportion of information missing = 5%, 10%, 20%, 30%, 40%, 50%), missing value mechanism (MCAR, MAR), and all two-way interactions involving the technique factor. The table values are the level of statistical significance of the factor listed in the row whose dependent variable is listed at the top of the column. Virtually all effects are statistically significant. A significant main effect for the technique factor implies that some techniques perform statistically significantly better than others on a performance criterion. A significant interaction implies statistically significant differences in techniques' relative performance across different levels of a robustness dimension.

In the interest of brevity, we forgo the exploration of significant main effects for each robustness dimension and forgo the explicit exploration of the significant interactions. Since our interest is in the relative performance of the different imputation techniques, we focus instead on the post hoc contrasts comparing techniques' errors on each level of the robustness dimensions.

Table 1

ANOVA Models Using Simulated Datasets

(Each cell reports the level of significance for the row independent variable in the ANOVA with the column dependent variable.)

Independent Variables in ANOVA Models	**Dependent Variable in ANOVA** mae_{values}	mae_{mean}	$mae_{variance}$	$mae_{covariance}$	$mae_{coefficient}$	ae_{ROC}
Technique	.000	.000	.000	.000	.000	.000
Sample size	.000	.000	.000	.000	.000	.000
Correlation	.000	.000	.000	.000	.000	.992
Proportion missing	.000	.000	.000	.000	.000	.000
Mechanism	.000	.000	.000	.000	.040	.000
Technique x Size	.000	.000	.000	.000	.000	.000
Technique x Correlation	.000	.000	.000	.000	.000	.988
Technique x Proportion missing	.000	.000	.000	.000	.000	.000
Technique x Mechanism	.000	.000	.002	.002	.000	.000
# Observations	3,000	3,600	3,600	3,600	3,420	684

* The performance criteria mae_{values} is not defined for CCA. This results in 3,000 observations for the associated ANOVA model. Also, the logit models estimated using CCA datasets do not converge in any dataset with 50% missing. Consequently, there are 3,420 and 684 observations for the ANOVA models associated to $mae_{coefficients}$ and ae_{ROC}, respectively.

To ease reading, we will say that technique A "performs better than" technique B on a particular criterion only when technique A has statistically significantly less error than technique B on that criterion. We will say that technique A "performs best" on a particular performance criterion, only when technique A has statistically significantly less error than all of the other techniques on that criterion. Our use of the phrases "performs worse than" and "performs worst" should be similarly understood to imply statistically significant differences.

Contrast analysis for techniques' overall error

The relative performance implied by the post hoc contrast analyses for each technique's overall error is summarized in Table 2. The six sections of Table 2 correspond to the six performance criteria. Within a section, the first row reports the results of contrasts involving techniques in the simulated datasets. (The remaining rows in each section relate to interpretation of contrasts for techniques in real customer datasets and will be discussed later.)

In the first row of the first section of Table 2, we see that EM is in the column headed "Technique(s) with Least Error." Hence, EM provides the best point estimates across all simulated datasets (i.e., lowest error for mae_{values}). In the same row, HD is in the column headed "Technique(s) with Most Error," indicating that HD provides the worst point estimates. In the column headed "Neither Least Error nor Most Error," MI = DA < Mean indicates that the point estimates from MI did not have statistically significantly more or less error than those from DA. However, both MI and DA provided point estimates that are better than the point estimates from Mean.

The first row of each section in Table 2 reports the relative performance of the different imputation techniques on each of the performance criteria we study. Continuing to focus on the first row of each section and scanning down the final column we see that CCA performs worst on all criteria other than accurately reproducing predictor variances ($ae_{variance}$) and point estimates (mae_{values}). (Recall that CCA does not

Table 2

Summary of Contrast Analyses for Technique's Overall Error

(For each performance criterion, a summary of all statistically significant differences is given for: all simulated datasets, simulated datasets with average correlation = .1, simulated datasets with average correlation = .2, real customer datasets with *N* = 8,940, and all real customer datasets).

Performance Criterion	Technique(s) with Least Error	Technique(s) with Neither Least or Most Error	Technique(s) with Most Error
mae_{values}			
Simulated data: all	EM	MI = DA < Mean	HD
Simulated: corr = .1	EM = Mean	MI = DA	HD
Simulated: corr = .2	EM = Mean*	MI = DA	HD
Real: *N* = 8,940	MI = DA = EM = Mean		HD
Real data: All	MI = DA = EM = Mean		HD
ae_{mean}			
Simulated data: all	EM	MI = DA < Mean < HD	CCA
Simulated: corr = .1	MI = DA = EM = Mean	HD	CCA
Simulated: corr = .2	MI = DA = EM = Mean	HD	CCA
Real: *N* = 8,940	MI = DA = EM = Mean = HD		CCA
Real data: all	MI = DA = EM = Mean	HD	CCA
$ae_{variance}$			
Simulated data: all	HD	MI = DA < EM < CCA	Mean
Simulated: corr = .1	HD	MI = DA = EM = CCA**	Mean
Simulated: corr = .2	HD	MI = DA = EM = CCA	Mean
Real: *N* = 8,940	HD		MI = DA = EM = Mean = CCA
Real data: all	HD	MI = DA = EM = Mean	CCA
$mae_{covariance}$			
Simulated data: all	MI = DA = EM	Mean = HD	CCA
Simulated: corr = .1	MI = DA = EM = Mean	HD***	CCA
Simulated: corr = .2	MI = DA = EM	Mean < HD	CCA
Real: *N* = 8,940	MI = DA = EM = Mean = HD = CCA		
Real data: all	MI = DA = EM = HD		Mean = CCA
$ae_{coefficients}$			
Simulated data: all	HD	MI = DA = EM = Mean****	CCA
Simulated corr = .1	MI = DA = EM = Mean = HD		CCA
Simulated corr = .2	MI = DA = EM = Mean = HD		CCA
Real: *N* = 8,940	MI = DA = EM = MEAN = HD = CCA		
Real data: all	MI = DA = EM = Mean = HD		CCA
ae_{ROC}			
Simulated data: all	MI = DA = EM = Mean	HD	CCA
Simulated: corr = .1	MI = DA = EM = Mean	HD	CCA
Simulated: corr = .2	MI = DA = EM = Mean	HD	CCA
Real: *N* = 8,940	MI = DA = EM = Mean = HD = CCA		
Real data: all	MI = DA = EM = Mean	HD	CCA

* For mae_{values}, simulated data, average correlation = .2, Mean also = MI and DA

** For $ae_{variance}$, simulated data, average correlation = .1, EM < CCA

*** For $mae_{covariance}$, simulated data, average correlation = .1, MI = HD

****For $ae_{coefficient}$, all simulated data, Mean < DA

Figure 1
Coefficient Error and Average Correlation in Complete Dataset for Simulated Data

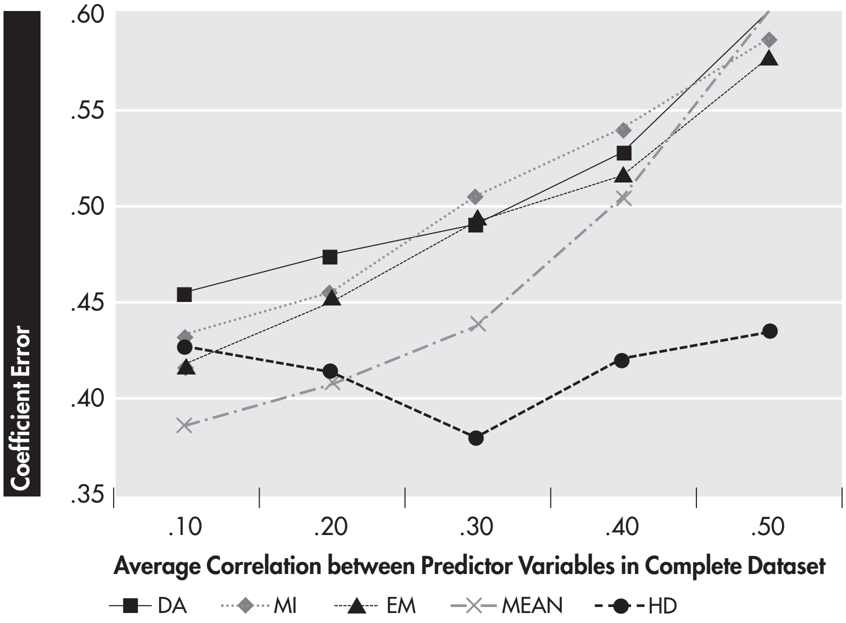
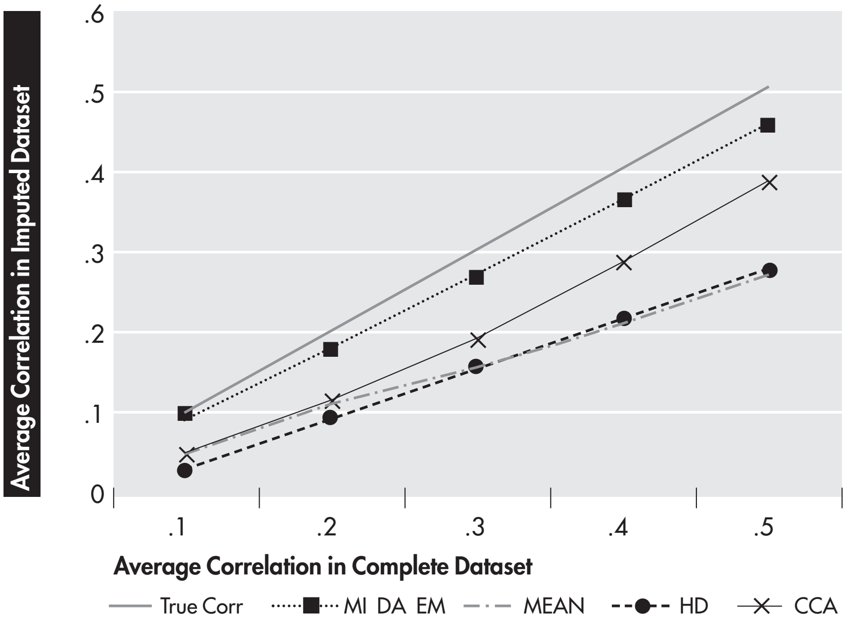


Figure 2
Average Correlation between Complete and Imputed Data



make point estimates and so is undefined for mae_{values}.) Scanning down top row entries in the first column indicates that, except for reproducing predictor variances ($ae_{variance}$) and yielding accurate logit coefficients ($ae_{coefficient}$), EM is never dominated by any other technique. Considering errors in covariance estimates ($mae_{covariance}$), MI and DA are statistically indistinguishable from EM. Considering the quality of marketing decisions (ae_{ROC}), MI, DA, and Mean are statistically indistinguishable from EM. As we speculated in the introduction, HD performs best and Mean performs worst when considering the accuracy of predictor variances ($ae_{variance}$). Surprisingly, HD also performs best when considering the accuracy of logit model coefficients ($ae_{coefficient}$).

Figure 1 shows the relative accuracy of logit coefficients estimated with techniques' imputed datasets ($ae_{coefficient}$) by plotting the error in coefficient estimates ($ae_{coefficient}$) for different levels of the robustness dimension "average correlation between predictors." For values of average correlation ranging from .1 to .5, coefficient error ($ae_{coefficient}$) for CCA is .69, .89, .96, .96, and 1.12, respectively. Rather than illustrate how badly CCA performs, we chose to scale the y-axis in Figure 1 to clearly show the relative performance of MI, DA, EM, Mean, and HD. As seen in the chart, the error in coefficient estimates for models estimated using datasets imputed with MI, DA, and EM grows as the "average level of correlation between predictors" grows. In fact, this growth is exactly what one would expect as multicollinearity (i.e., "average level of correlation between predictors") grows. It is surprising that the accuracy of estimated coefficients for Mean (for low levels of "true" correlation) and HD (for high levels of "true" correlation) are less sensitive to multicollinearity in the underlying, complete datasets.

Figure 2 relates "true" correlation (average level of correlation in the complete dataset) to the average level of correlation in datasets imputed with different techniques. The average level of correlation in datasets imputed with HD and Mean is lower than the "true" level of correlation between predictors. Hence, it seems the datasets imputed with HD and to some extent those imputed with Mean are able to produce “better” coefficient estimates because they poorly reproduce the “true” correlations between predictors. With lower levels of correlation between predictors in the datasets imputed by HD and Mean, models estimated upon these datasets are less vulnerable to the wide swings in coefficient estimates resulting from highly correlated predictors. It is important to note that the quality of marketing decisions (i.e., ae_{ROC}) implied by “less accurate” coefficients estimated using MI, DA, and EM is better than the quality of marketing decisions implied by the “more accurate” coefficients estimated using HD. This is consistent with multicollinearity causing instability in model parameter estimates without causing the quality of predictions to degrade.

In summary, the contrast analysis for techniques’ overall error in the simulated data shows that CCA performs worst for all performance criteria (for which it is defined) except the estimation of predictor variances ($ae_{variance}$). When considering predictor variances ($ae_{variance}$), HD performs best and Mean performs worst. For all performance criteria other than the estimation of predictor variances ($ae_{variance}$), EM is never dominated. MI and DA perform as well as EM when estimating predictor covariances ($mae_{covariance}$) and MI, DA, and Mean perform as well as EM when considering the quality of implied marketing decisions (ae_{ROC}). Results for the accuracy of logit coefficients ($ae_{coefficient}$) are difficult to interpret because of the high multicollinearity inherent in some of the complete datasets.

Contrast analysis for techniques’ error within the robustness dimensions

We study the techniques’ error within a robustness dimension (sample size, predictor correlation, proportion missing, or missing value mechanism) by looking for statistically significant differences between techniques on a criterion variable on each level of a robustness dimension. This generates 225 pair-wise statistical tests for each of the six performance criteria yielding a total of 1,265 tests as discussed earlier. We summarize those tests in six matrices on the left side of Table 3. Each matrix represents the results for a different performance criterion.

For the matrix associated with a particular performance criterion, each cell reports a number reflecting the performance for the technique listed in the row compared to the performance of the technique listed in the column. A value of “100%” means the technique in the row has statistically significantly more error than the technique in the column for every level of each robustness dimension. A value of “0%” means the technique in the row never has statistically significantly more error than the technique in the column for any level of any robustness dimension. A value between 0% and 100% indicates that the technique in the row sometimes has statistically significantly more error than the technique in the column.

The value in any cell in Table 3 is calculated by considering 15 statistical tests: two tests for the two different sample sizes, five tests for the five different levels of correlation, six tests for the six different levels of proportion missing, and two tests for two missing value mechanisms. If we filled in a cell with the number of tests out of 15 for which the technique in a row has statistically significantly more error than the technique in the column, those robustness dimensions represented with more levels (degree of correlation between predictors with five levels and proportion missing with six levels) would be more influential in determining the cell value than would robustness dimensions with fewer levels (sample size and missing value mechanism with two levels each). To balance the influence across robustness dimensions, we first calculate the percentage of a robustness dimension’s levels for which the row technique has statistically significantly more error than the column technique. Next, we calculate the average of these percentages across the robustness dimensions. These average percentage scores are reported in the matrix cells in Table 3. We refer to a cell value as “the percent of time” the row

Table 3

Summary of Contrast Analyses of Techniques' Errors within the Robustness Dimensions

Criterion	Technique	Simulated Data						Real Customer Data					
		MI	DA	EM	Mean	HD	CCA	MI	DA	EM	Mean	HD	CCA
mae_{values}	MI	---	0%	88%	5%	0%	---	---	0%	0%	0%	0%	---
	DA	0%	---	88%	5%	0%	---	0%	---	0%	0%	0%	---
	EM	0%	0%	---	0%	0%	---	0%	0%	---	0%	0%	---
	Mean	86%	86%	90%	---	0%	---	0%	0%	0%	---	0%	---
	HD	90%	100%	100%	100%	---	---	100%	100%	100%	100%	---	---
ae_{mean}	MI	---	0%	0%	0%	0%	0%	---	0%	0%	0%	0%	0%
	DA	0%	---	0%	0%	0%	0%	0%	---	0%	0%	0%	0%
	EM	0%	0%	---	0%	0%	0%	0%	0%	---	0%	0%	0%
	Mean	60%	64%	69%	---	0%	0%	0%	0%	0%	---	0%	0%
	HD	96%	96%	96%	92%	---	0%	50%	50%	50%	50%	---	0%
	CCA	100%	100%	100%	100%	100%	---	100%	100%	100%	100%	94%	---
$ae_{variance}$	MI	---	0%	0%	0%	96%	0%	---	0%	0%	0%	39%	0%
	DA	0%	---	0%	0%	96%	0%	0%	---	0%	0%	44%	0%
	EM	29%	29%	---	0%	96%	18%	0%	0%	---	0%	72%	0%
	Mean	96%	96%	96%	---	100%	74%	0%	0%	0%	---	83%	17%
	HD	0%	0%	0%	0%	---	0%	0%	0%	0%	0%	---	0%
	CCA	65%	65%	60%	4%	100%	---	33%	33%	33%	17%	72%	---
$mae_{covariance}$	MI	---	0%	0%	0%	0%	0%	---	0%	0%	0%	0%	0%
	DA	0%	---	0%	0%	0%	0%	0%	---	0%	0%	0%	0%
	EM	0%	0%	---	0%	0%	0%	0%	0%	---	0%	0%	0%
	Mean	95%	95%	95%	---	13%	5%	61%	67%	72%	---	0%	0%
	HD	100%	100%	100%	31%	---	0%	0%	0%	0%	0%	---	0%
	CCA	100%	100%	100%	43%	48%	---	50%	50%	61%	0%	33%	---
$ae_{coefficients}$	MI	---	0%	0%	4%	53%	0%	---	0%	0%	0%	0%	0%
	DA	0%	---	0%	4%	48%	0%	0%	---	0%	0%	0%	0%
	EM	0%	0%	---	4%	53%	0%	0%	0%	---	0%	0%	0%
	Mean	0%	0%	0%	---	43%	0%	0%	0%	0%	---	0%	0%
	HD	13%	13%	17%	8%	---	0%	0%	0%	0%	0%	---	0%
	CCA	96%	96%	96%	96%	96%	---	33%	33%	33%	17%	6%	---
ae_{ROC}	MI	---	0%	0%	0%	0%	0%	---	0%	0%	0%	0%	0%
	DA	0%	---	0%	0%	0%	0%	0%	---	0%	0%	0%	0%
	EM	0%	0%	---	0%	0%	0%	0%	0%	---	0%	0%	0%
	Mean	0%	0%	0%	---	0%	0%	0%	0%	0%	---	0%	0%
	HD	88%	88%	88%	88%	---	0%	0%	0%	0%	0%	---	0%
	CCA	92%	92%	92%	92%	88%	---	72%	83%	83%	83%	83%	---

*Cell entries are the average, across robustness dimensions, of the proportion of levels within a robustness dimension for which the row technique has statistically significantly more error than the column technique. We refer to this as the "percent of time" that the row technique has statistically significantly more error than the column technique.

technique was statistically significantly inferior to the column technique.

To illustrate how a cell value is calculated, consider the fifth cell down in the first column of Table 3. The cell value of 90% indicates that HD provides worse point estimates (mae_{values}) than MI for most of the levels of the four robustness dimensions. This value is determined by noting that HD had statistically significantly more error than MI for both samples sizes, for three of five levels of correlation between predictors, for all of the levels of proportion missing, and for both MCAR and MAR. An equal-weighted average of 100%, 60%, 100%, and 100% yields the 90% reported in the cell.

The order of the techniques in the columns and rows of each matrix in Table 3 is consistent with the relative performance one might expect based on the literature cited earlier in this paper, i.e., $MI \prec DA \prec EM \prec MEAN \prec HD \prec CCA$ ("$\prec$" indicates less error). For a particular performance criterion, if every statistical test were consistent with this ordering, then all cells above the main diagonal would contain zeroes. We will discuss some of those situations in which this expected pattern does not hold.

Consistent with Schafer and Graham's (2002) concern that techniques focused on accurately recovering missing values may compromise their ability to accurately recover data parameters, our results show that techniques' relative performances differ across performance criteria. For example, MI and DA provide more accurate variance estimates ($ae_{variance}$) than EM 29% of the time while EM provides better point estimates (mae_{values}) than MI and DA 88% of the time.[8] Interestingly, even Mean provides better point estimates than MI and DA 5% of the time.[9]

Other performance criteria with non-zero entries above the main diagonal include $ae_{variance}$ and $ae_{coefficient}$. As in our discussion of main effects and as the literature suggests, HD provides better variance estimates ($ae_{variance}$) than other techniques 96-100% of the time and Mean provides worse variance estimates than other techniques 74-100% of the time. The fact that HD provides better coefficient estimates ($ae_{coefficients}$) than MI, DA, and EM 48-53% of the time is a manifestation of the multicollinearity problem discussed in the section on contrast analysis for techniques' overall error. Other non-zero entries above the main diagonal result from extreme situations.[10]

Combining the insights gleaned from simulated data contrast analysis of both the overall error and the error within robustness dimensions, we provide the following summary of results:

1. CCA performs worst on all criteria except the accurate estimation of predictor variances, where it is second worst.

2. HD provides the best estimates of predictor variances.

3. EM provides the best point estimates and the best estimates of predictor means.[11]

4. MI, DA, and EM provide the best estimates of predictor covariances.

5. MI, DA, EM, and Mean provide the best marketing decisions

Real Customer Data Results

Each column in Table 4 represents an ANOVA model whose dependent variable is the performance criterion at the top of the column and whose independent variables are technique (MI, DA, EM, Mean, HD, CCA), sample size (100, 250, 8,940), proportion missing (5%, 10%, 20%, 30%, 40%, 50%), missing value mechanism (MCAR, MAR), and all two-way interactions involving the technique factor. The table values are the level of statistical significance of the row's independent variable in the ANOVA model with the column's dependent variable. Many fewer effects are statistically significant for the ANOVA models using real customer

Table 4

ANOVA Models Using Real Customer Datasets

(Each cell reports the level of significance for the row independent variable in the ANOVA with the column dependent variable.)

Independent Variables in ANOVA Models	Dependent Variable in ANOVA					
	mae_{values}	ae_{mean}	$ae_{variance}$	$mae_{covariance}$	$ae_{coefficient}$	ae_{ROC}
Technique	.000	.000	.000	.000	.152	.000
Sample size	.940	.000	.264	.000	.000	.000
Proportion missing	.603	.000	.000	.000	.000	.388
Mechanism	.937	.000	.001	.087	.381	.281
Technique x Size	.987	.007	.000	.000	.272	.000
Technique x Proportion missing	.999	.000	.016	.000	.007	.402
Technique x Mechanism	.982	.000	.000	.220	.110	.507
# Observations*	900	1,080	1,080	1,080	1,035	207

* The performance criteria mae_{values} is not defined for CCA. This results in 900 observations for the associated ANOVA model. Also, the logit models estimated using CCA datasets do not converge in any dataset with 50% missing. This results in 1,035 observations and 207 observations for the ANOVA corresponding $mae_{coefficients}$ and ae_{ROC}, respectively.

datasets than were significant for the ANOVA models using simulated datasets. This may result from reduced statistical power (Cohen 1977) since ANOVA models using real customer datasets have fewer observations than ANOVAs using simulated datasets. In addition, two of the robustness dimensions do not take on the same range of values in the simulated and real customer datasets. The level of average correlation between predictors ranges from .1 to .5 in the simulated datasets while it ranges from .12 to .24 in the real customer datasets. Sample size takes on only two values (N = 100 and N = 250) in simulated datasets while it takes on three values in real customer datasets (N = 100, N = 250, and N = 8,940). Given that we found few statistically significant differences between imputation techniques in low correlation simulated datasets and in very large sample (N = 8,940) real customer datasets,[12] it is not surprising that there are fewer statistically significant effects in the ANOVA models estimated using real customer datasets.

As in the discussion of results from simulated datasets, we forgo exploration of significant main effects for each robustness dimension and forgo explicit exploration of significant interactions. In discussing post hoc contrasts, we again use the terms “perform better,” “perform best,” “perform worse,” and “perform worst” to denote statistically significant differences between techniques.

Contrast analysis for techniques' overall error

The relative performance implied by the post hoc contrast analyses for each technique's overall error is summarized in last row of each section of Table 2. Note that there are no statistically significant differences between MI, DA, and EM for any performance criterion. Also, for all performance criteria except predictor variances and covariances ($ae_{variance}$ and $mae_{covariance}$), there are no statistically significant differences between MI, DA, EM, and Mean. MI, DA, and EM perform best for all performance criteria except the accurate estimation of predictor variances ($ae_{variance}$). HD produces the best and CCA produces the worst estimates of predictor variances. Finally, similar to results in the simulated data, HD produces the worst point estimates (ae_{values}) and CCA performs worst on all other criteria.

To see that the contrast analysis for techniques' overall error in real customer datasets is consistent with that from the simulated datasets, consider the implications of the differences in the range of predictor correlation values and the range of sample sizes in the simulated and real customer datasets. Table 2 reports (second and third rows of each section) the results for post hoc contrasts between techniques for the subset of the simulated datasets in which the average correlation between predictors is .1 and for the subset of the simulated datasets in which that average correlation is .2. The fourth row of each section of Table 2 reports the post hoc contrasts between techniques for the subset of the real customer datasets with "very large samples" (*N* = 8,940). Since the average correlation between predictors in the real customer datasets ranges from .12 to .24, the results from rows two and three provide a rough prediction for results in the real customer datasets (reported in row five). Since real customer datasets include "very large samples" (*N* = 8,940), one might expect the rough predictions based on results in rows two and three to be somewhat moderated based on results in row four.

Considering all of the real customer datasets (fifth row of the first section of Table 2), point estimates made by MI and DA do not have statistically significantly more error than those made by EM and Mean as one would hypothesize based on results from simulated datasets with the low correlation between predictors (rows two and three of the first section of Table 2). This may be driven by the fact that in the "very large sample" datasets (fourth row of the first section of Table 2), there is no statistically significant difference in point estimate error between MI, DA, EM, and Mean. Again, considering all of the real customer datasets, note that HD does somewhat better and Mean does somewhat worse at estimating covariance than one would hypothesize based on results using simulated data with low correlation between predictors. Again, this may be driven by the similarity of techniques' performances in the "very large sample," real customer datasets.

Contrast analysis for techniques' error within the robustness dimensions

We study the techniques' error within a robustness dimension (sample size, proportion missing, and missing value mechanism) by looking for statistically significant differences between techniques on a criterion variable on each level of a robustness dimension. As was pointed out earlier in the paper, we create 165 pair-wise statistical tests for each of the six performance criteria yielding a total of 925 tests. We summarize those tests in six matrices on the right side of Table 3, where each matrix represents the results for a different performance criterion. Recall that a cell entry reports the "percent of time" (as defined earlier) that the row technique has statistically significantly more error than the column technique on the criterion variable.

Table 3 indicates that HD provides better estimates of predictor variances than do other techniques 39-83% of the time. MI, DA, and EM provide better estimates of predictor covariances than Mean 61-72% of the time and better estimates than CCA 50-61% of the time. MI, DA, EM, and Mean provide the best point estimates, the best estimates of predictor means, the most accurate logit coefficients, and the best marketing decisions.

Differences between results for post hoc contrasts of error within robustness dimensions in the real datasets compared to results in simulated datasets are consistent with the explanations given for differences in results for post hoc contrasts of overall error presented earlier in this section.

Combining the insights gleaned from real customer data contrast analysis of both the overall error and of error within robustness dimensions, we provide the following summary of results:

1. HD provides the best estimates of predictor variances.

2. CCA performs worst.

3. MI, DA, EM, and Mean provide the best point estimates, estimates of predictor means, logit coefficients, and marketing decisions.

4. MI, DA, EM, and HD provide the best covariance estimates.

Discussion and Conclusions

Taken together, the results for the simulated and real customer datasets provide guidance to marketers concerned with the problem of missing information. We can unambiguously advise that one never resort to complete case analysis (CCA) when confronted with missing information. This is an important insight given the fact that CCA is the default treatment for missing information in popular statistical analysis software packages.

We can further unambiguously advise that one use hot deck (HD) when one's objective is to estimate predictor variances. There was no situation in either our analysis of the simulated data or our analysis of the real customer data in which another imputation technique produced estimates of predictor variances that had statistically significantly less error than those produced by HD.

Considering the three techniques MI, DA, and EM, note that it is only for estimates of predictor variances (in small samples with high levels of missing information) that MI and DA produce estimates that have statistically significantly less error than those produced by EM. If one were to follow the suggestion above and use HD to estimate predictor variances, then note that for all of the other performance criteria in our study EM produces estimates that are either statistically indistinguishable or statistically significantly superior to those of MI and DA. Given that MI and DA each require enormous amounts of computational power (for example, the MI algorithm required nearly 3,000 times more execution time than Mean or EM on a Pentium 4 2.54 GHz computer), it is our recommendation that one use EM for estimating missing values, predictor means, and predictor covariances.

The question of which imputation technique to use if one is focused on coefficient estimates should only be asked in situations in which there is a low level of correlation between predictors. As pointed out in the paper, a high level of correlation (i.e., high multicollinearity) implies that many different sets of coefficient values could each produce good predictions. For those situations in which the average correlation between predictors is .2 or less, our results indicate that datasets imputed with any of the techniques other than CCA will yield model coefficients that are equally good. Given that Mean imputation is simpler that MI, DA, EM, or HD, parsimony suggests that Mean imputation be used. If one is focused on the quality of marketing decisions implied by a model estimated on the imputed dataset, MI, DA, EM, and Mean perform equally well. Again, the law of parsimony suggests that Mean imputation be recommended.

In summary, we recommend that one never use CCA. In determining which imputation technique is most appropriate in a given situation, one needs to consider the analyst's goal. If the goal is to choose an imputation technique that provides the best point estimates for missing values, the best estimates of predictor means or covariances, we suggests using EM. If the goal is to choose an imputation technique that provides the best estimates of predictor variances, we suggest using HD. If the goal is to choose an imputation technique that provides the best estimates of model coefficients in environments in which there is low correlation between predictors, or to choose an imputation technique that implies the best marketing decisions, we suggest using Mean. ■

Acknowledgements

The authors would like to thank Equifax for providing data and the McCombs School Faculty Summer Research Grants for supporting this research.

Notes

[1] Kamakura and Wedel (2000) also note that the estimated factors can be used to impute missing values.

[2] Rubin (1987) shows that creating three to five imputed datasets provides sufficiently accurate parameter estimates. In our implementation of MI, we use five imputations for each missing value.

[3] Since the performance criterion mae_{values} is not defined for CCA, there are only 10 *possible technique pairs for* mae_{values}.

[4] Because there are only 10 technique pairs for mae_{values}, we get only 150 post hoc comparisons for mae_{values}. Logit models estimated using CCA datasets do not converge in any dataset with 50% missing. Hence CCA cannot be compared against any other technique in this condition. This results in 220 possible contrasts for the performance criteria of $ae_{coefficient}$ and ae_{ROC}.

[5] The Scheffé test is used to test simultaneous, multiple contrasts involving treatment effects. It is used for post hoc hypothesis testing and does not assume the tests are orthogonal. In this paper, the imputation techniques are the treatment effects and paired comparisons between two techniques are the contrasts. According to Scheffé (1959), we can reject the null hypothesis that there is no difference between two techniques by first calculating the usual F-statistic for a contrast and then comparing that F to $(K-1)^*F_{\alpha;K-1,N-K}$, where K is the number of treatments, $1-\alpha$ is the significance level, and N is the number of observations. If the F-statistic is greater than $(K-1)^*F_{\alpha;K-1,N-K}$, then we can reject the null.

[6] Since the performance criterion mae_{values} is not defined for CCA, there are only 10 possible technique comparisons for mae_{values}.

[7] Because there are only 10 technique pairs for mae_{values}, we get only 110 post hoc comparisons for mae_{values}. Logit models estimated using CCA datasets do not converge in any dataset with 50% missing. Hence CCA cannot be compared against any other technique in this condition. This results in 160 possible contrasts for the performance criteria of $ae_{coefficient}$ and ae_{ROC}.

[8] For $ae_{variance}$, MI and DA has statistically significantly less error than EM with 50% of information missing, in small sample ($N = 100$), and for MCAR. For mae_{values}, MI and DA have statistically significantly more error than EM for both sample sizes, for all six levels of correlation between predictors, for both MCAR and MAR, and for levels of missing information = 30%, 40%, and 50%.

[9] For mae_{values}, Mean has statistically significantly less error than MI and DA when the average level of correlation between predictors is .1 or .2.

[10] For $ae_{coefficient}$, MI, DA, and EM have statistically significantly more error than HD for average level of correlation between predictors = .3, .4, and .5; for the small sample; for MCAR; and for 5% and 10% of information missing. At 30% of information missing, MI and EM have statistically significantly more error than HD. For $ae_{coefficients}$, Mean outperforms MI, DA, and EM when 50% of the information is missing. On $mae_{covariance}$, HD outperforms Mean when data are MCAR and CCA outperforms Mean when the average correlation between predictors is .5.

[11] In the contrast analysis for techniques' overall error for the simulated data, EM provides better point estimates and better estimates of predictor means relative to MI and DA. In the contrast analysis for techniques' error within a robustness dimension for the simulated data we see that, EM provides better point estimates 88% of the time relative to MI and DA. However, when considering techniques' relative abilities to provide accurate estimates of predictor means using interactions, none of the contrasts between EM and MI or DA were statistically significant. This reflects the lower statistical power in the tests involving interactions effects (Cohen 1977).

[12] The second and third rows of each section of Table 2 show that there are few statistically significant differences in low correlation, simulated datasets. Row four of each section of Table 2 shows that there are very few statistically significant differences in very large sample, real customer datasets.

References

Black, Sandra, and Donald Morgan (1998) "Risk and the Democratization of Credit Cards." New York, N.Y.: Federal Reserve Bank of New York, Paper No. 9815 (June).

Bradlow, Eric, Ye Hu, and Teck-Hua Ho (2002), "A Learning-based Model for Imputing Missing Levels in Partial Conjoint Profiles." Philadelphia, Penn.: University of Pennsylvania, Wharton School,Working Paper.

Bronnenberg, Bart, and Catarina Sismeiro (2002), "Using Multimarket Data to Predict Brand Performance in Markets for Which No or Poor Data Exists." *Journal of Marketing Research* 39 (February), 1–17.

Bult, Jan, and Tom Wansbeek (1995), "Optimal Selection for Direct Mail." *Marketing Science* 14 (4), 378–94.

Cipra, Tomas, and Jose Trujillo (1995), "Holt-Winters Method with Missing Observations." *Journal of Management Science* 41 (1), 174–8.

Cohen, Jacob (1977), *Statistical Power Analysis for the Behavioral Sciences*. New York, N.Y.: Academic Press.

Dempster, Arthur, Nan Laird, and Donald Rubin (1977), "Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society* B39, 1–38.

DeSarbo,Wayne, Paul Green, and D. Carroll (1986), "An Alternative Least-Squares Procedure for Estimating Missing Preference Data in Product-Concept Testing." *Decision Sciences* 17 (2), 163–85.

DeSarbo,Wayne, Martin Young, and Arvind Rangaswamy (1997), "A Parametric Multidimensional Unfolding Procedure for Incomplete Nonmetric Preference/Choice Set Data in Marketing Research." *Journal of Marketing Research* 34 (November), 499–516.

Dick, A., D. Chakravarti, and G. Biehal (1990), "Memory-Based Inference during Consumer Choice." *Journal of Consumer Research* 19, 82–93.

Erdem, Tulin, Michael Keane, and Baohong Sun (1999), "Missing Price and Coupon Availability Data in Scanner Panels: Correcting for the Self-Selection Bias in the Choice Model Parameters." *Journal of Econometrics* 89, 177–96.

Faraggi, David, and Benjamin Reiser (2002), "Estimation of the Area under the ROC Curve." *Statistics in Medicine* 21, 3093–106.

Ford, Barry (1983), "An Overview of Hot Deck Procedures." In *Incomplete Data in Sample Surveys*, vol. 2, eds. William Madow, Ingram Olkin, and Donald Rubin. New York, N.Y.: Academic Press.

Gleason, Terry, and Richard Staelin (1975), "A Proposal for Handling Missing Data." *Psychometrika* 40, 229–52.

Green, David (1966), *Signal Detection Theory and Pychophysics*, New York, N.Y.: Wiley.

Henley, J., and B. McNeil (1982), "The Mean and Use of the Area under a Receiver Operating Characteristic Curve." *Radiology* 143, 29–36.

Johnson, Richard, and Irwin Levin (1985), "More Than Meets The Eye: The Effect of Missing Information on Purchase Evaluations." *Journal of Consumer Research* 12 (September), 169–77.

Kamakura, Wagner, and Michel Wedel (1997), "Statistical Data Fusion for Cross-Tabulation." *Journal of Marketing Research* 34 (November), 485–98.

Kamakura, Wagner, and Michel Wedel (2000), "Factor Analysis and Missing Data." *Journal of Marketing Research* 37 (November), 490–8.

Kivetz, Ran, and Itamar Simonson (2000), "The Effects of Incomplete Information on Consumer Choice." *Journal of Marketing Research* 37 (November), 427–48.

Lawrence, David (1992), *Handbook of Consumer Lending*. Englewood Cliffs, N.J.: Prentice Hall.

Little, Roderick (1992), "Regression with Missing X's: A Review." *Journal of the American Statistics Association* 87 (December), 1227–37.

Little, Roderick, and Donald Rubin (1987), *Statistical Analysis with Missing Data*. New York, N.Y.: John Wiley.

Lomax, Richard (2001), *An Introduction to Statistical Concepts for Education and Behavioral Sciences*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Ross, William, and Elizabeth Creyer (1992), "Making Inferences about Missing Information: The Effects of Existing Information." *Journal of Consumer Research* 19, 14–25.

Rubin, Donald (1987), *Multiple Imputation for Nonresponse in Surveys*. New York, N.Y.: Wiley.

Schafer, Joseph, and John Graham (2002), "Missing Data: Our View of the State of the Art." *Psychological Methods* 7 (2), 147–77.

Schafer, Joseph (1997), *Analysis of Incomplete Multivariate Data*. New York, N.Y.: CRC Press.

Schafer, Joseph, and John Graham (2002), "Missing Data: Our View of the State of the Art." *Psychological Methods* 7 (2), 147–77.

Scheffé, Henry (1959), *The Analysis of Variance*. New York, N.Y.: John Wiley and Sons, Inc.

Sirdeshmukh, Deepak, and H. Rao Unnava (1994) "Reducing Competitive Ad Interference." *Journal of Marketing Research* 31 (3), 403–12.

Steenburgh, Thomas, Andrew Ainslie, and Peder Engebretson (2003), "Massively Categorical Variables: Revealing the Information in Zip Codes." *Marketing Science* 22 (1), 40–57.

Sullivan, Charlene, and Robert Fisher (1988), "Consumer Credit Delinquency Risk: Characteristics of Consumers Who Fall Behind." *Consumer Credit Delinquency Risk* 10.3, 53–64.

Switzer, Fred, Philip Roth, and Deborah Switzer (1998), "Systematic Data Loss in HRM Settings: A Monte Carlo Analysis." *Journal of Management* 24 (6), 763.

Wilks, S. (1932), "Moments and Distributions of Estimates of Population Parameters from Fragmentary Samples." *Annals of Mathematics Statistics* 3 (August), 163–95.

Report No. 05-107
"Handling Missing Values in Marketing Data: A Comparison of Techniques" © 2005 James Lemieux and Leigh McAlister; Report Summary © 2005 Marketing Science Institute