

**A Unified Imputation Approach for the Treatment and Analysis of  
Missing Data in Marketing Research**

**T. E. Raghunathan**

**Venkatram Ramaswamy**

**Steven H. Cohen**

**Kerimcan Ozcan**

**February 1999**

# **A Unified Imputation Approach for the Treatment and Analysis of Missing Data in Marketing Research**

## **ABSTRACT**

We present a unified approach that integrates multiple imputation with simulation, for the treatment and analysis of missing data in marketing research. It is a generalized approach that can accommodate continuous, categorical or censored data, and is suitable for many common situations encountered in practice. The proposed approach preserves associations in the observed multivariate data, focuses on data distributions rather than obtaining point estimates, produces asymptotically unbiased estimates, incorporates sampling variability and uncertainty in standard errors, and can utilize information from auxiliary variables when one has more missing data on core variables of interest. A simulation study with 2000 synthetic data sets demonstrate the robustness and efficiency of MIGS, when compared with alternative approaches. An illustrative commercial application is also provided. We conclude by offering some guidance for dealing with missing data in marketing research.

# **A Unified Imputation Approach for the Treatment and Analysis of Missing Data in Marketing Research**

Missing data plague market research, especially in the analysis of multivariate data from surveys and secondary databases. In this paper, we address the following situations that are quite common in practice:

*Scenario A:* A large company collects customer data on an on-going basis using surveys. The corporate research group makes available databases to different business units within the company. Within each business unit, end-users (analysts) estimate different models using different pieces of information in the core databases. The corporate research group must however contend with missing data on a regular basis. What can the corporate research group do to the core databases, before sharing data with end-users?

Scenario A is also typical of syndicated data suppliers that provide commercial and private data, and in general, any private or public databases that are shared among many ultimate users. Further, these end-users typically have varying degrees of statistical expertise and sophistication, and with different managerial questions and objectives.

*Scenario B:* A company wishes to forecast demand for a new product in a certain geographic region. Income is one of several core predictor variables in a model of buying intentions. The data at hand contain several auxiliary variables (e.g., ownership of other products) that are related to, but not part of, the analyst's model.

While there are many refusals of income, the auxiliary variables have much more complete data. What can the company do to analyze the relationship between the predictors and buying intentions, if one or more of these is often missing?

In general, scenario B is typical of any situation where an analyst can delineate “core variables” (independent and dependent variables that are relevant to the analyst’s model specification) that contain missing data, and “auxiliary variables” containing more complete data that are related to, but not part of, the analyst’s model.

#### *Stylized Approaches for Specific Problems*

Over the past forty years, statisticians and econometricians have responded to the challenge of analyzing incomplete data, particularly using methods entailing classical statistical inference (cf. Afifi and Elashoff 1966; Kim and Curry 1977; Little and Rubin 1987 for reviews). Traditional procedures and estimators have been developed for *specific* types problems such as the multiple factor model (Finkbeiner 1979), multiple regression and discriminant analyses (e.g., Little 1978; 1979), analysis of variance and covariance with experimental data (e.g., Smith 1981), simultaneous equations (e.g., Dagenais 1976) and so on, and in special instances where the missing data mechanism may be known (Little and Rubin 1987). The marketing research literature has also offered approaches for handling missing data for specific models, although primarily when the researcher has complete data on the independent variables, but data are missing on the dependent variable. Malhotra (1987) discusses the application of the EM algorithm (see Dempster, Laird, and Rubin 1977) for

analyzing incomplete data in situations where the researcher has substantial control over the independent variables (e.g., DeSarbo, Green, and Carroll 1986; Punj and Staelin 1978), or the independent variables are known treatments as in experimental research, or in longitudinal panel studies with missing dependent variable observations (e.g., Winer 1983).

Although a majority of the above approaches are suitable for the *idiosyncratic* problems and models they have been designed for, an inherent limitation is their *lack of generalizability* to other situations. The major problem in Scenario A is that the data provider does not know *a priori* what specific models will be estimated. In Scenario B, the user has missing data on the *independent variables* for which there is little guidance in terms of stylized approaches. Moreover, a mainstream analyst may lack the adequate knowledge of the specific problems and proposed solutions in the literature, and may not have the ability to engage in custom programming and/or use specialized routines for different problems and models.

#### *Towards a Unified Approach*

Our goal is to fill the lacuna for a generalized approach to *imputing missing data* in multivariate data sets that can potentially free a mainstream analyst from having to employ stylized approaches designed for specific problems. Currently, there is very little guidance offered about missing data issues in textbooks on data analysis and research design (e.g., Pedhazur and Schmelkin 1991; Greene 1997; Malhotra 1996). It is hardly surprising that in practice, market researchers may choose ad hoc approaches such as *listwise or pairwise deletion* in which the analysis is

conducted only on complete observations, or *mean substitution* where the mean on a given variable is used to replace missing values on that variable. The logic from the mainstream analyst's perspective is often to avail of standard complete data techniques and general purpose software packages such as SPSS, SAS, or BMDP that often require a rectangular file. Unfortunately, the problem with these ad hoc methods is that they typically yield statistically invalid answers for quantities of scientific interest to the analyst, such as data summaries or parameters estimated from a model using the data (Rubin 1996).

We first briefly review the distinct and popular approaches for the imputation of incomplete multivariate data in marketing research. Against this background, we motivate and position our *unified* approach, provide the technical details, and discuss its practical usefulness in the scenarios described above. We then illustrate our unified approach using data from a commercial setting for the analysis of customer satisfaction with a financial services firm. We then evaluate the robustness of the proposed approach by comparing its performance with that of several popular approaches using a multitude of synthetic data sets. We conclude the paper with a discussion of the strengths and limitations of the proposed approach, and offer some guidance for the treatment and analysis of missing data in marketing research.

## **IMPUTATION OF MISSING DATA: A BRIEF REVIEW**

Consider a rectangular dataset (matrix) of complete data,  $Y$ , with  $I$  rows and  $N$  columns. Let  $y_i$  denote the  $i$ -th row of  $Y$ ,  $i = 1, \dots, I$ . Assuming that the rows can be modeled as independent, identically distributed (iid) draws from some multivariate

probability distribution, the probability density of the complete data may be written as:

$$P(Y|\theta) = \prod_{i=1}^I f(y_i|\theta), \quad (1)$$

where  $f$  is the density for a single row, and  $\theta$  is a vector of unknown parameters. The complete data  $Y$  can be factored as:

$$P(Y|\theta) = P(Y_{obs}|\theta)P(Y_{mis}|Y_{obs}, \theta), \quad (2)$$

where  $Y_{obs}$  and  $Y_{mis}$  denote the observed and missing part of  $Y$  respectively.

Let  $R$  denote a binary rectangular matrix with the same dimension as  $Y$  such that  $R_{in} = 1$  if  $Y_{in}$  is observed and  $R_{in} = 0$  if  $Y_{in}$  is missing. Following Rubin (1976), the missing data are denoted *as missing at random* (MAR) if

$$P(R|Y_{obs}, Y_{mis}, \xi) = P(R|Y_{obs}, \xi), \quad (3)$$

and the parameters  $\xi$  of the missing data mechanism are distinct from the parameters  $\theta$  (i.e., is *ignorable*, Little and Rubin, 1987; Rubin, 1987). Despite its name, note that MAR does *not* suggest that the missing data values are a random sample of all data values (this condition is actually denoted as MCAR or *missing completely at random*); rather, MAR implies that the probability that an observation is missing may depend on  $Y_{obs}$ , but not on  $Y_{mis}$ .

In imputation-based procedures, missing values are filled in (imputed) to facilitate standard complete data statistical analyses. Several imputation-based procedures have been proposed that differ in the way in which imputations are generated. The mean substitution method imputes missing data with the

unconditional sample mean of observed values for a particular variable (Wilks 1932; Gleason and Staelin 1975). It can however distort the underlying distribution of the data and is typically not recommended (Little 1992). This method can be somewhat improved with hot decking that fills in the missing item with a value which is drawn from a set of responding units (“donors”) similar to the one with the missing data (Ford 1983; Little 1997). The major limitation of hot-deck procedures is their heuristic nature. They require subjective decisions on which variables to use for adjustment cells, selecting matching metrics for continuous variables, and distinguishing critical variables (for exact matches) from general matching variables (Kamakura and Wedel 1997). Current hot-deck procedures employ a single imputed value and theoretically have poor sampling properties (Li, Raghunathan, and Rubin 1991).

Another method is regression imputation, in which missing values are filled in with means conditioned on the variables recorded for that case. This is achieved by first regressing variables for which imputation is sought on the variables for which complete responses exist, and then calculating the predicted means for missing variables using the estimated regression coefficients. Several variants of this method have been proposed (Buck 1960; Gleason and Staelin 1975; Frane 1976; Kim and Curry 1977; Little and Rubin 1987; Little 1992). This method is generally recommended over ad hoc methods, especially if the predicted values are augmented by a residual so as to reflect uncertainty in the predicted value (called stochastic regression imputation). Little (1997) notes that standard error of estimates from

filled-in data are better, but they are still underestimated due to the omission of imputation error itself.

Relatively newer *model-based* methods can explicitly model the missing data mechanism and estimate parameters based on the likelihood under that model (e.g., DeSarbo et al. 1986; Lee and Chiu 1990). In its most general form, *maximum likelihood* (ML) inference computes those parameter values that maximize the likelihood of having observed both the data and missingness pattern. Since the observed data truly consists of not only  $Y_{\text{obs}}$  but also  $R$ , the probability distribution of the observed data is given by (Schafer 1997):

$$\begin{aligned} P(R, Y_{\text{obs}} | \theta, \xi) &= \int P(R, Y | \theta, \xi) dY_{\text{mis}} \\ &= \int P(R | Y, \xi) P(Y | \theta) dY_{\text{mis}}. \end{aligned} \quad (4)$$

However, under the MAR assumption, (4) becomes:

$$\begin{aligned} P(R, Y_{\text{obs}} | \theta, \xi) &= P(R | Y_{\text{obs}}, \xi) \int P(Y | \theta) dY_{\text{mis}} \\ &= P(R | Y_{\text{obs}}, \xi) P(Y_{\text{obs}} | \theta). \end{aligned} \quad (5)$$

Hence, in common situations where the missing data mechanism can be safely assumed to be ignorable and MAR, maximum-likelihood estimation of  $\theta$  can be performed without regard to the missing data mechanism, i.e. the observed-data likelihood  $L(\theta | Y_{\text{obs}}) \propto P(Y_{\text{obs}} | \theta)$ . Except in special cases, however, these tend to be complicated functions of  $\theta$ , and extracting meaningful summaries such as parameter estimates and standard errors require special computational tools such as the

Expectation-Maximization (EM) Algorithm (Dempster, Laird, and Rubin 1977) discussed next.

*EM Algorithm.* Following expression (2), the complete data log likelihood can be factored as:

$$\log L(\theta|Y) = \log L(\theta|Y_{obs}) + \log P(Y_{mis}|Y_{obs}, \theta) + c, \quad (6)$$

where  $P(Y_{mis}|Y_{obs}, \theta)$  is the conditional predictive distribution of the missing data given  $\theta$ , and  $c$  is an arbitrary constant. The EM method capitalizes on  $P(Y_{mis}|Y_{obs}, \theta)$  which captures the interdependence between  $Y_{mis}$  and  $\theta$ . Since  $Y_{mis}$  is however unknown, (6) can be averaged over the predictive distribution  $P(Y_{mis}|Y_{obs}, \theta^t)$ , where  $\theta^t$  is a preliminary estimate of  $\theta$  yielding (Dempster, Laird, and Rubin 1977; Schafer 1997):

$$\begin{aligned} S(\theta|\theta^t) &= \int l(\theta|Y)P(Y_{mis}|Y_{obs}, \theta^t)dY_{mis} \\ &= l(\theta|Y_{obs}) + \int \log P(Y_{mis}|Y_{obs}, \theta)P(Y_{mis}|Y_{obs}, \theta^t)dY_{mis} + c. \end{aligned} \quad (7)$$

Hence, in the E-step, the function  $S(\theta|\theta^t)$  is calculated using (7) and in the M-step,  $\theta^{t+1}$  is found by maximizing  $S(\theta|\theta^t)$ . The E- and M-steps are alternated with a starting value  $\theta^0$  until convergence (see Malhotra 1987 for a specific application to regression models with incomplete data on a dependent variable). Several extensions of EM have been proposed to speed up convergence (Ruud 1991; Meng and Rubin 1993; Liu and Rubin 1994; Meng and Van Dyk 1997).

Note that in the EM method, the missing values are not being directly filled, but, rather, functions of them are used in the log likelihood. Software packages, such

as the recent SPSS Missing Value Analysis module and MISS (a Gauss program module) contain implementations of the EM method for the computation of means and covariances matrices, based on normal distributions for continuous variables. They output imputed data as well for further analyses. However, only a single imputed value is produced that does not account for sampling and uncertainty in imputation.

### *Multiple Imputation*

All the methods discussed so far impute a single new value for a missing data unit and treat it as if it were known in subsequent analyses. However, single imputation neither reflects the sampling variability for a particular missing model nor compensates for the additional uncertainty of not knowing which missing data model is the true one. The idea of multiple imputation was introduced by Rubin (1987) to overcome these limitations, albeit in the context of nonresponse in sample surveys. The idea, however, is quite general: replace the missing data  $Y_{\text{mis}}$  by simulated values  $Y_{\text{mis}}^{(1)}, Y_{\text{mis}}^{(2)}, \dots, Y_{\text{mis}}^{(M)}$  to create  $M$  completed datasets. Each of the  $m$  datasets can then analyzed by *standard* complete-data methods requiring rectangular files. The variability among the results of the  $M$  analyses provides a measure of the uncertainty due to missing data, which, when combined with typical measures of sample variation, lead to a single inferential statement about the parameters in any model estimated by the analyst.

Rubin (1987, pp 75-76) originally provided a theoretical justification for using a *Bayesian* approach, noting that for multiple imputation to be proper, the simulated

values of  $Y_{\text{mis}}$  must be drawn from the *posterior predictive distribution* of the missing values:

$$P(Y_{\text{mis}}|Y_{\text{obs}}) = \int P(Y_{\text{mis}}|Y_{\text{obs}}, \theta)P(\theta|Y_{\text{obs}})d\theta. \quad (8)$$

Hence, a Bayesianly proper multiple imputation would reflect uncertainty about  $Y_{\text{mis}}$  given the parameters of the complete-data model, as well as uncertainty about unknown model parameters.

## MULTIPLE IMPUTATION WITH GIBBS SAMPLING

In practice, multivariate data sets have missing data on several variables and the expression in (8) consequently entails a multidimensional integral. This poses a major challenge in generating simulated values of  $Y_{\text{mis}}$  from the posterior predictive distribution that invariably does not have a tractable closed form. However, recent advances in Markov Chain Monte Carlo methods of simulation (Gilks, Richardson, and Spiegelhalter, 1996) allow draws to be simulated from non-standard posterior distributions, given the knowledge of the full conditional distributions. In particular, implementation of the Gibbs sampler (e.g., Gelfand and Smith 1990) requires knowledge of only the full conditionals, up to proportionality.

We now discuss a *unified* approach for the imputation of missing data, that we refer to as ‘Multiple Imputation with Gibbs Sampling’ or MIGS. The major highlights of MIGS are as follows:

- It entails a generalized complete data model specification that can handle missing data of mixed scale types. As noted by Schafer (1997), little attention has been

paid to treatment and analysis of multivariate data with variables of different scale types that are practically the norm in commercial market research. Unlike imputation methods such as EM that have been formulated for continuous variables, or adjustment cell hot decks for categorical variables, MIGS accommodates both continuous and categorical variables in a single modeling framework.

- It generates multiply imputed complete data sets thereby accommodating Scenario A. This retains the main advantage of single imputation methods, viz., a rectangular complete data set while correcting its disadvantages by reflecting both sampling variability under missing data and uncertainty of imputation.
- It exploits recent advances in Bayesian computation for simulating missing data using the Gibbs sampler. This provides a computationally efficient way of producing multiply imputed datasets from the posterior predictive distribution of missing data. As demonstrated subsequently, this also enables the incorporation of auxiliary variables as in Scenario B and allows the analyst to capitalize on prior knowledge.

### *The Complete Data Model*

To achieve a generalized framework that can handle mixed scale types of continuous and categorical variables, we utilize an extension of the general location scale model (Olkin and Tate 1961; Little and Rubin 1987; Raghunathan and Grizzle 1995) to specify a multinomial log-linear model for missing categorical variables, a

multivariate normal model for missing continuous variables, and a mixed variable model for censored variables (e.g., a non-negative continuous variable with a spike at zero).

Specifically, suppose that  $Y_{\text{mis}}$  consists of a  $r$ -dimensional continuous variable  $Z$  and a  $q$ -dimensional categorical variable  $X$  (e.g., some of the core variables indicated in Scenario B).<sup>1</sup> Let  $U$  denote a  $p$ -dimensional variable (e.g., containing the auxiliary variables in Scenario B) that is fully observed on all the individuals in a given random sample of size  $I$ . The complete data model involves specifying a joint distribution of  $(Z, X)$  given  $U$ . A convenient representation of this joint distribution is through the specification of the distribution of  $X$  given  $U$  and then the distribution of  $Z$  given  $X$  and  $U$ .

Suppose that  $X_j$  has  $C_j$  levels for  $j=1, \dots, q$ . These categorical variables can form a contingency table with  $C = \prod_j C_j$  cells. Let  $c = (i_1, i_2, \dots, i_q)$  denote a cell in the  $q$ -way contingency table with  $C$  cells. Let  $n_c = n_{i_1 i_2 \dots i_q}$  denote the number of individuals with  $X_1=i_1, X_2=i_2, \dots, X_q=i_q$ , where  $i_j = 1, \dots, C_j$ . Given a random sample of size  $I$ , the  $C$ -dimensional vector of cell counts  $n_c$  is assumed to be multinomially distributed with cell probabilities  $\pi = (\pi_{i_1 i_2 \dots i_q})$ . Next, we specify the conditional distribution of  $Z$  given  $U$  and  $X$ . Given  $X$ , or equivalently a specific cell  $c$ , the continuous responses (or their

---

<sup>1</sup> Although MIGS can handle censored continuous variables (e.g., James 1995), for the sake of notational simplicity, we omit this distinction without loss of generality. The examples that follow demonstrate that MIGS can handle such variables as well.

tarnsforms)  $Y_{ci}$ ,  $i= 1, \dots, n_c$ , are assumed to be identically and independently distributed normal random variables with mean  $\mu_c$  and a covariance matrix  $\Sigma$ .

Even if a few categorical variables, however, the number of cells  $C$  can be very large. Hence, we impose a log-linear model structure on the cell probabilities:

$$\log \pi_{i_1 i_2 \dots i_q} = W_1' \alpha, \tag{9}$$

where  $\alpha$  is a  $s_1 \times 1$  vector of regression coefficients and  $W_1$  is an appropriate design matrix that may involve the known values  $U$  and log-linear parameters representing the main and interaction effects. In practice, we have found that the inclusion of terms up to third-order interactions is sufficient, although this can be easily modified if higher-order interactions are warranted. Also, note that structural zeroes can be accommodated easily as well.

To reduce the dimensionality of the parameters further, we impose additional structure by introducing random effects that allow for borrowing strength across the various cells formed by the categorical variables. Hence, we assume that the cell means  $\mu_c$  are normally distributed with mean  $V_c \tau$  and covariance matrix  $\Omega$ , where  $V_c$  is a  $r \times s_2$  matrix of covariates defined by the categorical variable  $X$  and the fully observed covariates  $U$ , and  $\tau$  is a  $s_2 \times 1$  vector of regression coefficients. The fully observed covariates are assumed to have an arbitrary distribution. The foregoing model includes mixed effects logistics models and linear regression models as particular cases. Finally, we assume a flat prior distribution for the parameters  $\omega = (\alpha, \Sigma, \tau, \Omega)$ , given by

$$P(\omega) \propto |\Sigma|^{-1} |\Omega|^{-(\nu/2+r)} \exp\{tr(B\Omega^{-1})\}, \quad (10)$$

i.e., a flat or non-informative prior for  $(\alpha, \Sigma, \tau)$  and a proper inverted Wishart prior for  $\Omega$ . Technically, a proper prior distribution for  $\Omega$  ensure a proper posterior distribution for  $\Omega$  (Raftery and Banfield 1991; DuMouchel and Waternaux 1992). The hyperparameters  $\nu$  (a scalar) and B (a matrix) may be chosen on the basis of external information. By choosing the degrees of freedom  $\nu$  and the elements of B to be close to zero (subject to the condition that B is positive semidefinite), the prior distribution can be made diffuse relative to the likelihood.

### *Gibbs Sampling*

Given the complete data model specification, the prior distribution for the parameters, and the observed data, imputations are created by drawing values from the posterior predictive distribution of the missing data using data augmentation (Tanner and Wong 1987) embedded in a Gibbs sampling (Gelfand and Smith 1990) framework as discussed below.

Following expression (8), the observed data posterior  $P(\theta|Y_{obs})$  is intractable and cannot easily be computed. However, when  $Y_{obs}$  is 'augmented' by an assumed value of  $Y_{mis}$ , the resulting complete data posterior  $P(\theta|Y_{obs}, Y_{mis})$  becomes much easier to deal with (Tanner and Wong 1987). Given a current guess  $\theta^{(t)}$  of the parameter, we can first draw a value of the missing data from the conditional predictive distribution of  $Y_{mis}$ :

$$Y_{mis}^{t+1} \sim P(Y_{mis} | Y_{obs}, \theta^{(t)}), \quad (11)$$

and then conditioning on  $Y_{mis}^{t+1}$ , we can draw a new value of  $\theta^{(t+1)}$  from its complete data posterior:

$$\theta^{(t+1)} \sim P(\theta | Y_{obs}, Y_{mis}^{(t+1)}). \quad (12)$$

This yields a stochastic sequence whose stationary distribution is  $P(\theta, Y_{mis} | Y_{obs})$ . For a suitably large value of  $t$ , we can regard  $\theta^{(t)}$  as an approximate draw from  $P(\theta | Y_{obs})$  and  $Y_{mis}^t$  as an approximate draw from  $P(Y_{mis} | Y_{obs})$ , the posterior predictive distribution of the missing data.

Gibbs sampling is utilized for generating draws from the posterior distribution in expression (12) which is multidimensional in nature. Briefly, draws from each conditional distribution (or that of a subvector) is obtained in a cyclic fashion, each time replacing the old values by the most recently drawn values of the parameters. After several cycles, we take the drawn values of the parameters as a draw from the joint posterior distribution. In a practical setting, after passing through some prespecified number of cycles (“burn in” period), we store the results of every  $P^{\text{th}}$  draw as an approximately independent draw from the joint posterior distribution (see Gelman and Rubin 1992; Gilks and Wild 1992; Raftery and Lewis 1992; and Smith and Roberts 1993 who discuss several important computational aspects of this approach). Thus to create  $M$  imputed values, we can integrate expression (11) with the Gibbs sampling cycle and draw a total of  $MP$  times from the relevant conditional distributions. Technical details are provided in the Appendix.

### *Using the Proposed Approach*

We return to Scenarios A and B at the beginning of this paper. In Scenario A, what can the corporate research group do to the core databases, before sharing data with end-users? It could use the proposed approach to create multiple imputed data sets and then distribute multiple data sets along with instructions (discussed below) on how to combine estimates across data sets (Rubin 1996). Moving to Scenario B, what can the company do to predict the relationship between the predictors and buying intentions, if one or more of these is often missing? If the analyst already has a specific predictive model that she wishes to use, the analyst could use one of the stylized approaches (if appropriate to her model) for the treatment and analysis of missing data. Alternatively, she could use the proposed approach to fill the “holes” in the data multiple times, estimate parameters for each data set using whatever classical methods that she is familiar with, and then combine the estimates, as discussed below.

In typical applications, good results can be obtained with as few as 3-5 imputations. This is because multiple imputation relies on simulation to solve only the missing data aspect of the problem. If the fraction of missing information about an estimand is  $\lambda$ , the relative efficiency of a point estimate based on  $M$  imputations is  $(1 + \lambda / M)^{-1}$  (Rubin 1987, p. 114). For example, when  $\lambda=0.3$  an estimate based on  $M=3$  imputations will have a standard error of only 1.049 times as large. Another reason why valid inferences can be obtained with small  $M$  is that the instructions for combining the  $M$  complete data analyses explicitly account for Monte Carlo error

(Schafer 1997). Suppose that  $Q$  denotes a  $k$ -dimensional population quantity of interest, such as the population mean of  $k$  characteristics or the regression coefficients in a particular analyst's model, involving some or all of the  $U$ ,  $Z$  and  $X$  variables. Each completed data set results in estimates of  $Q$  and the associated sampling variance. Let  $\hat{Q}_m$  and  $\hat{S}_m$  denote the estimate and the associated variance based on the  $m^{\text{th}}$  completed data set, where  $m = 1, \dots, M$ . The multiply imputed estimate of  $Q$  is given by the average

$$\hat{Q}_{MI} = \sum_m \hat{Q}_m / M, \quad (13)$$

and the associated covariance matrix can be estimated by

$$\hat{S}_{MI} = \mathbf{I} + r_M \sum_m \hat{S}_m / M, \quad (14)$$

where  $r_M = \mathbf{m} M^{-1} \mathbf{r} \sum_m (\hat{Q}_m - \hat{Q}_{MI}) \hat{S}_m^{-1} (\hat{Q}_m - \hat{Q}_{MI})' / \mathbf{k}(M-1)$ .

## COMPARATIVE EVALUATION WITH SYNTHETIC DATA

We conducted a modest simulation study to compare the performance of MIGS with alternative approaches. We first generated 1000 complete data sets, each with a sample size of 300 containing three core variables: a continuous dependent variable  $Y$ , two continuous independent variable,  $X1$  and  $X2$ , as well as four auxiliary variables that were designed to be related to the core variables: a continuous variable  $Z1$  correlated more with  $X1$ , a continuous variable  $Z2$  correlated more with  $X2$ , a categorical variable  $Z3$  correlated with all the core variables, and a censored variable

Z4 (with a spike of zeroes) correlated with all the core variables. We imposed the following “true” regression model (to represent an analyst’s model) in generating these data:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \varepsilon, \quad (15)$$

with  $\beta_1$  (Beta 1) =  $\beta_2$  (Beta 2) = 0.5, and  $\varepsilon$  being a normal random error term. We then imposed varying patterns of missing data across the 1000 data sets as summarized in Table 1. In general, the amount of missing data for X2 was more than X1, and the dependent variable Y had less missing data than X1 or X2. The median amount of missing data (marginal percentage) was 32% for X2, 26% for X1, and 22% for Y. The percent of complete cases varied from 45% to 62% with a median of about 53% (Q1 and Q3 indicate the lower and upper quartiles respectively).

The main objective of the simulation study was to compare the sampling properties of the point and interval estimates of these regression coefficients by using MIGS with those obtained by using traditional imputation approaches. Each of the 1000 data sets was first treated by the following methods discussed earlier: mean substitution, regression imputation, and EM imputation (the SPSS Missing Value Analysis module was used for these analyses); then, the analyst’s regression model in (15) was estimated for each imputed data set, as well as for the complete data set (i.e., prior to imposing missing data). The median  $R^2$  was 45% with a minimum of 30% and a maximum of 58% across the 1000 complete data sets. The estimated regression coefficients using the complete data sets (referred to as “true recovered coefficients”)

were compared to the respective coefficients (referred to as “analyst estimates” using the imputed data sets.

### *Simulation Results*

Figure 1 plots the true recovered coefficient versus the analyst estimate for each of the 1000 data sets (the 45 degree scatterplot represents the true recovered coefficients). The mean substitution method exhibits considerable bias for both Beta 1 and Beta 2. Further, for Beta 1, the regression imputation method performs better, and the EM method performs the best; for Beta 2 however, both these methods exhibit more bias (recall that X2 has more missing data) although they still perform better than the mean substitution method. We then used MIGS to impute the missing data using only the core variables (Y, X1, and X2), and then also using the auxiliary variables Z1, Z2, Z3, and Z4. The resulting plots for MIGS are shown in Figure 2. The performance of MIGS is evidently superior to the other three approaches. Note that MIGS is also able to take advantage of the additional information provided by the auxiliary variables (that would exist in Scenario B discussed earlier).

In practice, there are two additional alternatives to imputation methods depending upon the sophistication of the analyst. A simple, unsophisticated alternative is to engage in listwise deletion and merely use only the complete cases. A more involved approach, assuming the analyst is rather sophisticated with a Bayesian inclination and familiarity with MCMC methods, is for her to engage in a ‘parameter simulation’ to simulate the parameters of her fixed predictive model, treating the missing data as random variables.

We utilized a Bayesian estimation software package for the applied Bayesian statistician called BUGS (Spiegelhalter et al., 1995) to estimate a regression model with each of the 1000 incomplete data sets. In BUGS, the missing data are treated as random variables. The model was specified as follows:

$$y_i \sim N(\mu_i, \sigma^2); \mu_i = \alpha + \beta_1 X_1 + \beta_2 X_2; \quad (16)$$

$$\alpha \sim N(0, 10000); \beta_1 \sim N(0, 10000); \beta_2 \sim N(0, 10000); \quad (17)$$

$$X_1 \sim N(\gamma_1, \tau_1^2); X_2 \sim N(\gamma_2, \tau_2^2); \quad (18)$$

$$\gamma_1, \gamma_2 \sim N(0, 10000); \tau_1^2, \tau_2^2 \sim Ga(1, 100); \sigma^2 \sim Ga(0.001, 1000), \quad (19)$$

where  $N(a,b)$  and  $Ga(a,b)$  denote a Normal and Gamma distribution, respectively, with mean 'a' and variance 'b'. Note that (18) specifies a normal prior distribution (with hyper-parameters) for each of the independent variables  $X_1$  and  $X_2$ , and (19) specifies non-informative priors for the hyper-parameters.

The upper part of Figure 3 depicts the resulting scatterplot from using the listwise deletion method. It appears to be biased somewhat less than the mean substitution method. The lower part of Figure 3 displays the BUGS results. Clearly, the parameter simulation approach is superior and appears almost identical to that of MIGS.

Table 2 summarizes the bias, mean-square error (MSE) and the exact coverage of the nominal 95% confidence interval for the three traditional imputation methods, MIGS, listwise deletion, and parameter simulation with BUGS. The mean-square errors of the estimates are expressed as percentages of the mean-square errors of the estimates based on the complete data sets, i.e., prior to imposition of missing data.

Expressing the mean-square error as a percentage provides a useful measure of the loss of precision due to missing values for the different methods.

It is evident from Table 2 that for mean substitution, the estimates are severely biased, have large mean-square errors, and the exact coverages are far below the nominal level. With this method, estimates are unbiased only if MCAR holds. There is a typical deflation of variances and covariances. Hence, bivariate associations get distorted, and estimates for the correlation and regression coefficients are underestimated (Little 1997).

The estimates for listwise deletion also have large mean-square errors and the exact coverages are below the nominal level. This is due to a loss in sample size and statistical power (e.g., Gilley and Leone 1991; Kim and Curry 1977; Little and Rubin 1987; Little 1992; Roth 1994), although they appear to fare better than mean substitution overall. The appeal of listwise deletion is usually the formation of a rectangular dataset so that standard statistical analyses can be easily conducted. However, the validity of statistical inferences depends crucially on the validity of the MCAR assumption. If that assumption does not hold, then parameter estimates will be biased and inefficient (Little and Rubin, 1987). Even in those rare instances when this assumption holds, unless the ratio of incomplete data to complete data is negligibly small, listwise deletion should not be the method of choice. Both mean substitution and listwise deletion are not recommended (Little 1992), although they are the most widely available options (and we suspect, most often used) in standard statistical software packages.

Both regression imputation and EM imputation appear to provide estimates with less bias and mean-squared errors, especially for Beta1 (recall that X1 has less missing data than X2). The exact coverages are however far below the nominal level due to the larger estimated standard errors of the estimates. In general, the single imputation methods have poor performance since they do not account for sampling variability and imputation uncertainty adequately. The parameter simulation approach (BUGS) entails draws of the missing data while simultaneously estimating the model parameters. As such, it performs well, as does MIGS. The inclusion of auxiliary variables with MIGS further improves its performance. This is because the imputations are more efficient as they borrow strength from the auxiliary variables.

The extent of missing values in the data sets that were simulated thus far were designed to be fairly representative of situations commonly encountered in practice. In the simulated data sets, half the data sets had between 51% and 55% complete cases with missing values on the variables ranging from 16% to 32%. In a commercial application presented in the next section, two of four core variables have about 30% missing values, but the number of complete cases is less than 25%. Although missing data plague market research in general, there are clearly situations where the extent of incompleteness is lower. Hence, we undertook a second round of simulations and generated another 1000 data sets with less missing data as shown in Table 3. Half of these data sets had complete cases around 83%-85% with missing values on the variables ranging between 2% to 12%.

Table 4 presents a summary of the results. In general, all the methods now produce results with less bias, lower mean-square errors, and greater coverage. Note, from Table 4, that mean substitution now performs better than listwise deletion, although they together still perform worse than regression and EM imputation, that exhibit far better performance. The parameter simulation approach and MIGS continue to perform well, although MIGS with auxiliary variables seems to perform the best overall. Our overall conclusions and recommendations are discussed following an illustrative commercial application of MIGS.

## **AN ILLUSTRATIVE COMMERCIAL APPLICATION**

Consider the following commercial scenario. A leading financial services firm conducts customer satisfaction surveys on a regular basis. A group has the responsibility of generating customer satisfaction reports for managers for tracking purposes. These reports summarize informational quantities of interest such as top-two box scores on satisfaction, etc. In addition, the staff of the advanced market research group is responsible for more detailed modeling and analyses of the data, such as investigating the drivers of overall satisfaction, and providing some conclusions from their analyses. This commercial application is representative of both Scenario A (central data collection, cleaning, generation of summary reports for tracking purposes) and Scenario B (customer satisfaction modeling for diagnostic purposes) discussed at the beginning of this paper. For the sake of brevity, we focus

more on estimation of the customer satisfaction model, rather than the summary reports of statistics tabulated from the data.

The firm offered a number of investment products/services. In certain prime regions, the firm had begun to implement an alternative distribution channel for its offerings. A corporate manager responsible for investments in the new distribution channel was interested in investigating the satisfaction of customers who had transacted with this alternative distribution channel. The firm conducted a customer satisfaction survey of a sample of these customers drawn from its internal database. In the survey, customers were asked to provide an overall satisfaction measure of their experience with the firm on a five-point scale ranging from very satisfied (7) to very dissatisfied (1). A mid-level score of four indicated neutral evaluation by the customers. Other questions were asked on satisfaction with the firm's product offerings, customer services, and satisfaction with the service delivery through the traditional and alternative distribution channels. Besides questions pertaining to customer satisfaction, the firm also sought to obtain information on customers' usage patterns and behaviors, as well as some descriptive background characteristics of its customers.

The customer satisfaction data had less than 25% of the respondents providing complete responses to the five core variables (overall satisfaction with firm, and satisfaction with product offerings, customer services, traditional, and alternative distribution channels). While almost all of the respondents gave their overall satisfaction rating, a majority of the respondents had some form of missing data

pattern in their responses to the remaining four core variables. For instance, on the variables pertaining to satisfaction with service delivery through the two distribution channels, as much as a third of the data were missing. Hence, using traditional approaches that rely on a rectangular data file would result in severe loss of information.

The distribution of the overall satisfaction ratings were skewed towards the higher end which is not surprising for a high-performance firm in the financial services sector (Peterson and Wilson 1992). In fact, the firm had stated goals of total satisfaction, and was focusing its quality initiatives on shifting customers to the very satisfied category to ensure long-term loyalty. Consequently for analysis purposes, the analysts decided to combine the bottom four categories. The customer satisfaction model was specified to be an ordered probit model, given the ordinal nature of response on the dependent variable (the difference in satisfaction between levels five and six on the scale is not necessarily the same as that between levels six and seven on the same scale). The LIMDEP software package (Greene, 1998) was to be employed for estimating the ordered probit model. In this model (see Greene, 1995), we define a latent (unobserved) continuous variable that measures the continuous (cardinal) value of overall satisfaction. For the  $i$ -th customer, the latent variable  $Z_i$  is specified as a linear function of the explanatory variables:

$$Z_i = \tilde{X}_i' \tilde{\beta} + \varepsilon_i \quad (20)$$

where  $\tilde{X}_i$  is a vector of the four core explanatory variables for the  $i$ -th customer,  $\tilde{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)$  is the respective vector of parameters associated with these drivers of overall satisfaction, and  $\varepsilon_i$  is a randomly distributed error term that follows a standard normal distribution. For each customer, this unobserved cardinal value  $Z_i$  is mapped to the observed overall satisfaction rating  $Y_i$  as follows:

$$\Pr(Y_i = j) = \Pr(\delta_{j-1} < Z_i \leq \delta_j) \quad (21)$$

where  $\Pr(Y_i = j)$  is the (ordered) probability that the observed rating for the  $i$ -th customer is the  $j$ -th level of the overall satisfaction scale ( $j = 1, \dots, J$ ), and  $\delta_j$  is the threshold value ( $j = 0, \dots, J+1$ ) such that  $\delta_0 < \delta_1 < \dots < \delta_J < \delta_{J+1}$ .

Table 5 compares the results from using only the complete cases versus MIGS (with auxiliary variables capturing customer characteristics and usage of products and services). The second and third columns give the parameter estimates  $\hat{\beta}$  (for brevity, the estimated threshold values have been suppressed). Note that the complete case analysis produces estimates that are lower in magnitude (and with somewhat larger standard errors) than MIGS. Since the effects of the explanatory variables are nonlinear, the implications of the estimated coefficients are better reflected in the last two columns, which show the marginal effect of a unit increase in satisfaction on the gain in very satisfied customers. Note that the complete case analysis systematically under-estimates the impact of each of explanatory variables. In particular, the two distribution channel variables (which had a preponderance of missing data) exhibit the largest differences in the results. Clearly, any conclusions

drawn from revenue/profit calculations of increased satisfaction through investments made in the new alternative channel would be misguided with the complete case analysis.

## CONCLUSION

We now offer some observations and guidance in dealing with missing data in marketing research, based on our collective experience. Foremost, analysts who wish to avail of rectangular data files for utilizing standard statistical software packages like SAS and SPSS must recognize the limitations of popular ad hoc approaches like mean substitution and listwise deletion. Unless the preponderance of missing data is small (say, about 10%) and are missing completely at random, there is considerable danger in obtaining statistically invalid inferences with these approaches. In other words, they must be avoided if possible. In general, regression and EM imputation are to be preferred. If not incorporated into the main software domain, these approaches are often available as additional modules. However, the disadvantage of the single imputation methods is that they do not reflect sampling variability and uncertainty of imputation.

Multiple imputation, as advocated by Rubin (1996) over the years, has recently come into its own, particularly with advances in computation (e.g., Gelman, King, and Liu 1998). As discussed, it requires the analyst to estimate a model multiple times ( $M = 3-5$  will usually suffice), and then appropriately combine and test the significance of the estimates. In previous years, this was indeed more laborious and

computationally intensive. However, as recently noted by Rubin (1996, p. 486), “multiple imputation is substantially easier for the ultimate user than any other current method that can satisfy the dual objectives of reliance only on complete-data methods and general validity of inference... it is becoming relatively easy for the data collector to create multiply-imputed files using modern computing hardware and accompanying algorithmic developments for Bayesian models.” When combined with the increasingly popular Gibbs sampling approach, an integrated method is possible (referred to as MIGS), as discussed and demonstrated in this paper.

MIGS integrates Multiple Imputation with Gibbs Sampling for the treatment and analysis of missing data on several variables. It is a generalized approach that can accommodate continuous, categorical or censored data, and is suitable for many common situations encountered in market research. MIGS preserves associations in the observed multivariate data, focuses on data distributions rather than obtaining point estimates, produces asymptotically unbiased estimates, incorporates sampling variability and uncertainty in standard errors, and can use information from auxiliary variables when one has more missing data on core variables. Our simulations with 2000 data sets demonstrate the robustness and efficiency of MIGS, when compared with alternative approaches.

The main advantage of MIGS stems from its general nature, potentially freeing a mainstream analyst from having to employ stylized approaches designed for specific problems. It can handle missing data in multivariate data sets arising from several missing data situations involving mixed scale types, and both

dependent and independent variables. Unlike parameter simulation (with BUGS), the analyst does not need to have a Bayesian inclination, possess specialized knowledge of Bayesian methods, and have the ability to engage in custom programming and/or use specialized routines for different problems and models. In Scenario A discussed earlier, the data provider does not know *a priori* what specific models will be estimated, but has the need for a robust method for “filling the holes” in the data. In Scenario B, the user has access to auxiliary variables that can help in imputing missing data on core variables of interest to the analyst. This cannot be accommodated easily in standard parameter simulation approaches. Further, although our focus has been on descriptive market research situations, MIGS can also be utilized in experimental research contexts (e.g., behavioral research studies) where certain covariates may confound the main analysis, but may provide additional information for imputation of key response variables and/or provide better estimates of treatment effects.

Unfortunately, a major limitation currently is that there is a lack of current software for multiple imputation, although this is changing rapidly (e.g., Schafer 1997; van Buuren, van Mulligen, and Brand, 1995). For mainstream use however, a user-friendly software package and/or interfaces for popular software packages is badly needed. To this end, one of the authors has developed a SAS interface for MIGS (that relies on SAS macros), which allows users to simply pick the variables to be imputed, delineate which variables are continuous, categorical, and censored, place restrictions and bounds on their imputation, and specify the number of

multiple imputations (e.g., our illustrative commercial application was estimated with this SAS interface). We hope that this paper serves as an impetus for further methodological and software development in this area.

## Appendix

Suppose that we have an initial draw of the missing values and the cell means  $\mu_c, c = 1, \dots, C$ . The initial draw of the set of missing values can be obtained by using simple techniques such as hot decking or random mean imputation (Rubin 1987). After filling in the set of missing values with the initial draw, the initial draw of the cell means can be obtained by using the mean of a bootstrap sample of observations in each cell.

For notational simplicity, we use the generic notation “ $|\cdot$ ” to denote the values that are being conditioned on, other than the argument of the posterior density. The computational process cycles through the following steps:

$$\text{Step 1. } \Omega^{-1}|\cdot \sim \text{Wishart}\left[\mathbf{I}B + \sum_c (\mu_c - V_c \tau)(\mu_c - V_c \tau)' \mathbf{Q}, C + \nu\right].$$

$$\text{Step 2. } \tau|\cdot \sim \text{Normal}(\hat{\tau}, \hat{P}) \text{ where } \hat{P} = (\sum_c V_c' \Omega^{-1} V_c)^{-1} \text{ and } \hat{\tau} = \hat{P} \sum_c V_c' \Omega^{-1} \mu_c.$$

$$\text{Step 3. } \Sigma^{-1}|\cdot \sim \text{Wishart}\left[\mathbf{I} \sum_{ci} (Z_{ci} - \mu_c)(Z_{ci} - \mu_c)' \mathbf{Q}, N - 2r + 2\right].$$

$$\text{Step 4. } \mu_c|\cdot \sim \text{Normal}(\hat{\mu}_c, S_c) \text{ where } S_c = (n_c \Sigma^{-1} + \Omega^{-1})^{-1} \text{ and } \hat{\mu}_c = S_c (n_c \Sigma^{-1} \hat{Z}_c + \Omega^{-1} V_c \tau)$$

and  $Z_c$  is the mean of the  $n_c$   $Z$ -values in cell  $c$ .

*Step 5.* Draw the values of the parameters in the log linear model. Given the initial draw of the missing categorical variables and  $U$ , we can fit a log-linear model using maximum likelihood. Any structural zero probabilities can be fit by using the constrained maximum likelihood (Bishop, Feinberg, and Holland 1975). Let  $\hat{\alpha}$  denote the maximum likelihood estimate and  $\hat{T}$  denote the observed Fisher information

matrix. We approximate the posterior distribution of  $\alpha|.$  by a multivariate normal distribution with mean  $\hat{\alpha}$  and covariance matrix  $-\hat{T}^{-1}$ .

*Step 6.* The final step is to draw the missing values, given the parameters. First, we draw the missing categorical outcome variable  $X_j$  for an individual. Given  $\alpha$  and the observed values of  $X_k, k \neq j$ , the previously drawn values of the missing  $X_k, k \neq j$ , and  $U$ , we can compute the conditional probability that  $X_j$  takes on a value  $i_j$  given  $X_k, k \neq j$  and  $U$  for an individual who is missing  $X_j$  where  $i_j = 1, \dots, C_j$ . Specifically, this conditional probability is  $\phi_{i_1, \dots, i_{j-1}, i_{j+1}, \dots, i_q}^{(i_j)} = \pi_{i_1, \dots, i_j, \dots, i_q} / \pi_{i_1, \dots, *, \dots, i_q}$  where "\*" in the subscript denotes summation over the particular margin. The drawn value is then a result of a multinomial experiment with these conditional probabilities. Next, to draw the missing continuous variables, note that given all the categorical variables we then know precisely the cell in which the individual  $i$  belongs. For notational brevity, we shall use  $obs \subset \{1, \dots, p\}$  to denote the indices of the observed continuous variables and  $Y_{i,obs}$  and  $Y_{i,mis}$  to denote the observed and missing continuous variables respectively on individual  $i$ . The predictive distribution of  $Y_{i,mis}$  is a multivariate normal with mean  $\mu_{m,mis} + \Sigma_{mis,obs} \Sigma_{obs,obs}^{-1} (Y_{i,obs} - \mu_{c,obs})$  and covariance  $\Sigma_{mis,mis} - \Sigma_{mis,obs} \Sigma_{obs,obs}^{-1} \Sigma_{obs,mis}$ , where  $\Sigma_{A,B}$  denotes a submatrix of  $\Sigma$  formed by the row indices in A and column indices in B. This completes the Gibbs cycle.

Table 1  
 Characteristics of Simulated Missing Data Sets  
 (High Incompleteness Condition)

<i>Missing Data Characteristics:</i>	<i>Min.</i>	<i>Q1</i>	<i>Median</i>	<i>Q3</i>	<i>Max</i>
Complete Case %	45	51	53	55	62
X2 Missing (Marginal %)	13	26	28	32	46
X1 Missing (Marginal %)	10	19	22	26	39
Y Missing (Marginal %)	5	16	18	22	34
Only X2 Missing %	8	13	14	15	20
Only X1 Missing %	5	8	9	10	14
X1 and X2 Missing %	3	5	6	7	11
Only Y Missing %	2	6	7	8	12
Y and X2 Missing %	1	4	4	5	8
Y and X1 Missing %	1	2	3	4	7
Y1, X1, and X2 Missing	1	4	4	5	7

Table 2  
 Comparison of Characteristics of Estimates from Analyst Model  
 (High Incompleteness Condition)

<i>Method</i>	<i>Beta1</i>			<i>Beta2</i>		
	<i>Bias (x1000)</i>	<i>MSE</i>	<i>95% Coverage</i>	<i>Bias (x1000)</i>	<i>MSE</i>	<i>95% Coverage</i>
Mean Substitution	104	400	68	112	531	64
Regression Imputation	55	210	78	72	380	72
EM Imputation	22	222	77	71	398	66
MIGS	16	182	93	12	209	93
MIGS with Auxiliary Variables	7	165	95	9	179	95
Listwise Deletion	95	455	79	87	461	81
Parameter Simulation (BUGS)	17	182	93	2	197	94

Table 3  
 Characteristics of Simulated Missing Data Sets  
 (Low Incompleteness Condition)

<i>Missing Data Characteristics:</i>	<i>Min.</i>	<i>Q1</i>	<i>Median</i>	<i>Q3</i>	<i>Max</i>
Complete Case %	76	83	84	85	90
X2 Missing (Marginal %)	3	8	10	12	21
X1 Missing (Marginal %)	1	4	4	7	14
Y Missing (Marginal %)	0	2	3	5	13
Only X2 Missing %	3	7	8	9	13
Only X1 Missing %	1	3	3	4	7
X1 and X2 Missing %	0	1	1	2	3
Only Y Missing %	0	2	2	3	6
Y and X2 Missing %	0	0	1	1	2
Y and X1 Missing %	0	0	0	1	3
Y1, X1, and X2 Missing	0	0	0	0	3

Table 4  
 Comparison of Characteristics of Estimates from Analyst Model  
 (Low Incompleteness Condition)

<i>Method</i>	<i>Beta1</i>			<i>Beta2</i>		
	<i>Bias (x1000)</i>	<i>MSE</i>	<i>95% Coverage</i>	<i>Bias (x1000)</i>	<i>MSE</i>	<i>95% Coverage</i>
Mean Substitution	26	132	93	52	201	89
Regression Imputation	7	106	95	13	108	95
EM Imputation	1	95	96	33	137	91
MIGS	6	118	96	2	117	96
MIGS with Auxiliary Variables	2	113	96	1	114	95
Listwise Deletion	55	214	87	40	175	90
Parameter Simulation (BUGS)	12	125	95	3	125	95

Table 5  
Empirical Results for Commercial Application

<i>Variables</i>	<i>Parameter Estimates<sup>a</sup></i>		<i>Marginal Effects</i>	
	<i>Complete Case (CC) Analysis</i>	<i>MIGS (M=5)</i>	<i>Complete Case (CC) Analysis</i>	<i>MIGS (M=5)</i>
<i>Product Offerings</i>	0.46 (0.08)	0.52 (0.04)	8.9%	10.2%
<i>Customer Services</i>	0.18 (0.06)	0.25 (0.03)	2.9%	4.0%
<i>Traditional Channel.</i>	0.35 (0.07)	0.46 (0.05)	5.4%	8.3%
<i>Alternative Chennel</i>	0.13 (0.06)	0.18 (0.05)	1.9%	4.0%

<sup>a</sup> Standard errors in parentheses.

FIGURE 1  
COMPARISON OF IMPUTATION METHODS

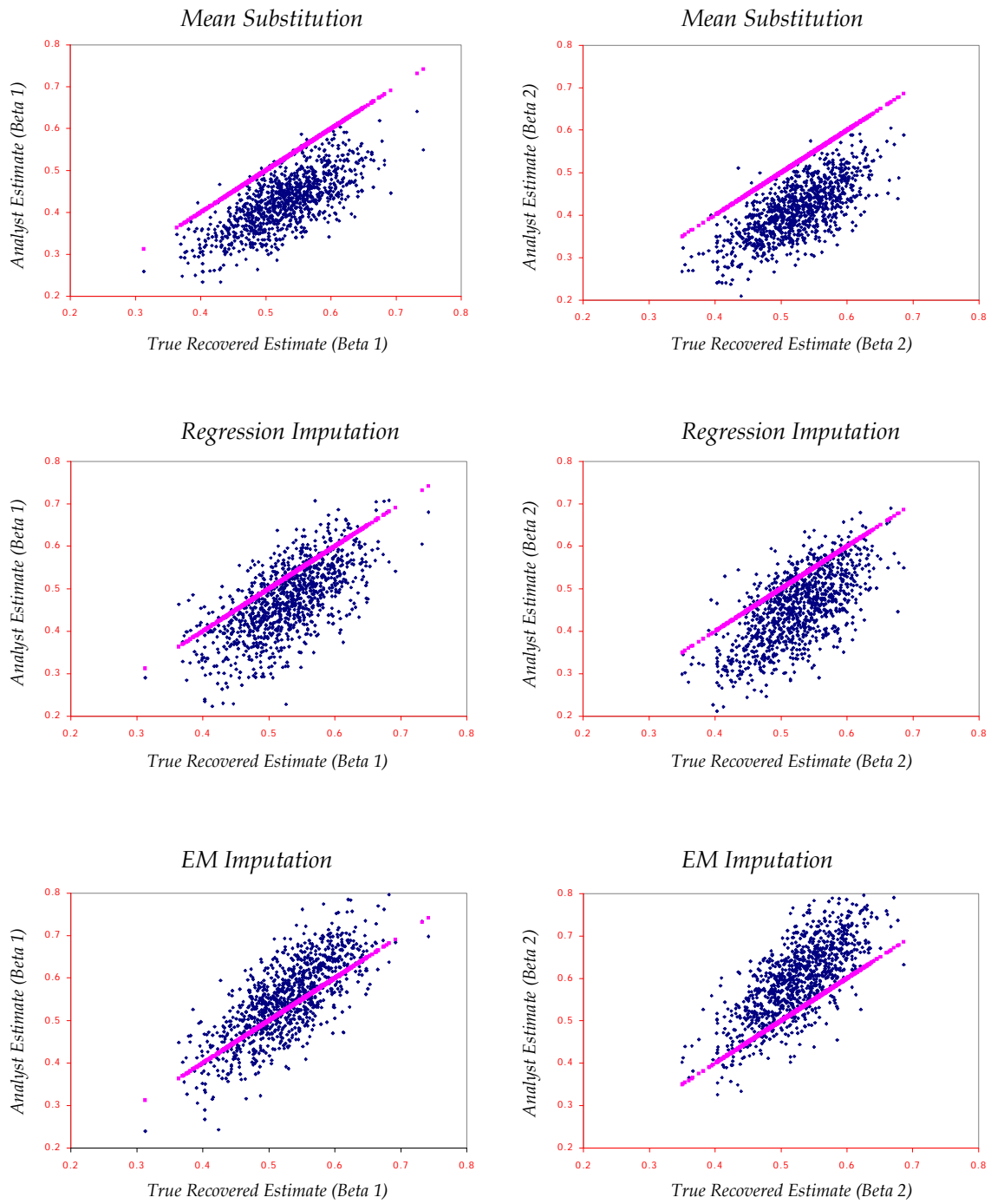


FIGURE 2  
PERFORMANCE OF MIGS

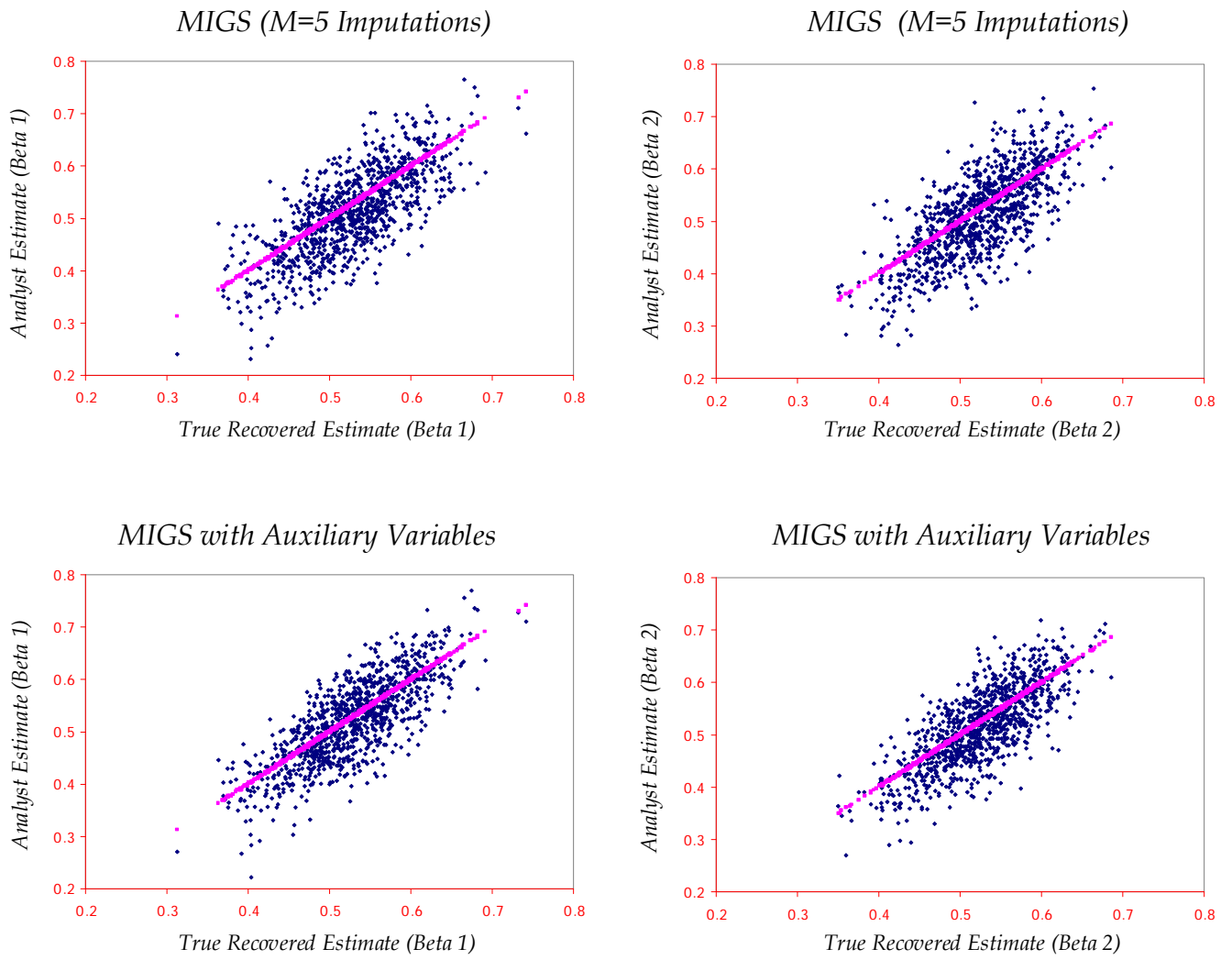
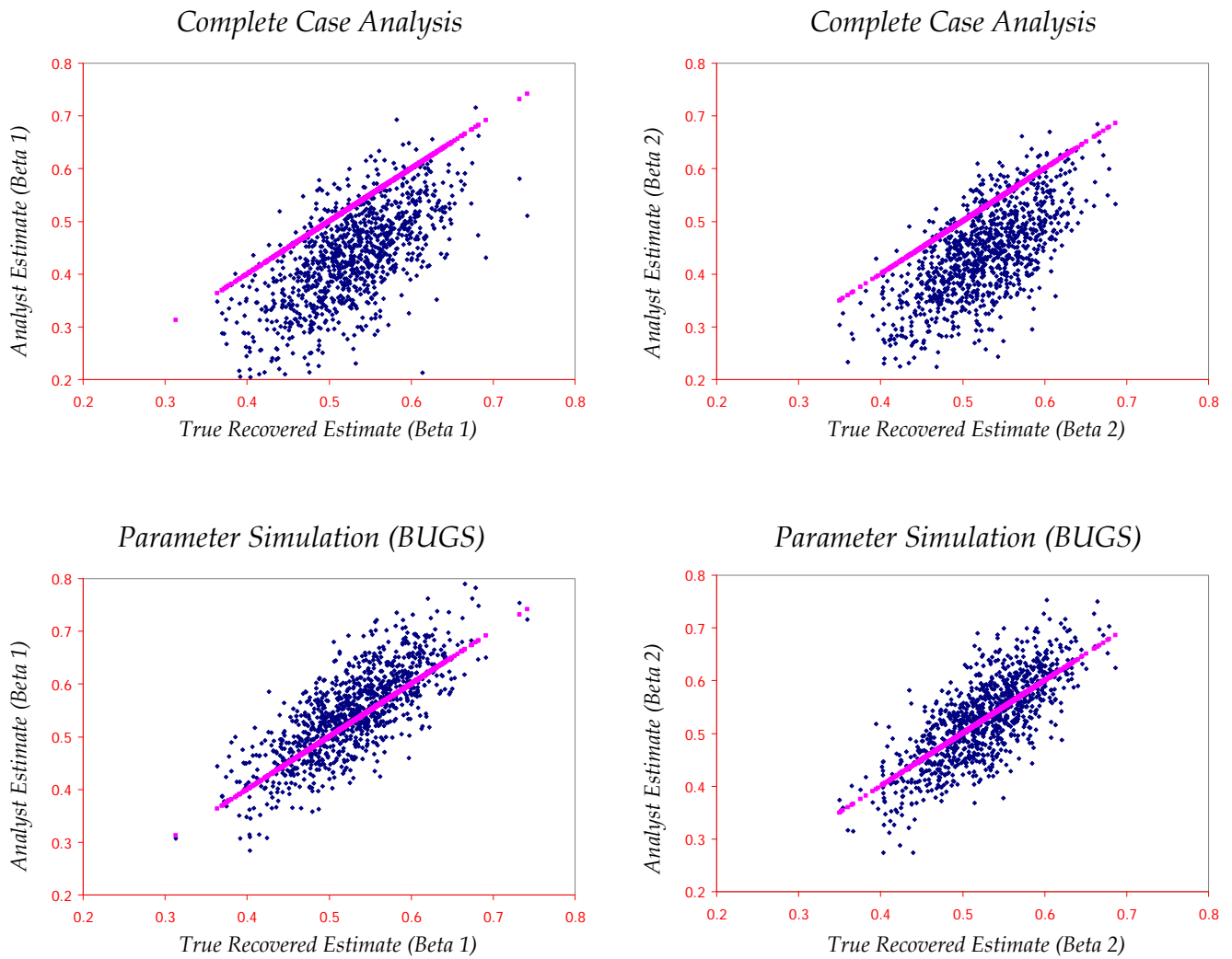


FIGURE 3  
COMPARISON OF ALTERNATIVE METHODS



## References

- Kamakura, Wagner A. and Michel Wedel (1997), "Statistical Data Fusion for Cross-Tabulation," *Journal of Marketing Research*, 34 (4), 485-498.
- Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall.