Survey Methodology, June 2001
Vol. 27, No. 1, pp. 85-95
Statistics Canada, Catalogue No. 12-001

A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models

Trivellore E. Raghunathan, James M. Lepkowski, John Van Hoewyk and Peter Solenberger [1]

Abstract

This article describes and evaluates a procedure for imputing missing values for a relatively complex data structure when the data are missing at random. The imputations are obtained by fitting a sequence of regression models and drawing values from the corresponding predictive distributions. The types of regression models used are linear, logistic, Poisson, generalized logit or a mixture of these depending on the type of variable being imputed. Two additional common features in the imputation process are incorporated: restriction to a relevant subpopulation for some variables and logical bounds or constraints for the imputed values. The restrictions involve subsetting the sample individuals that satisfy certain criteria while fitting the regression models. The bounds involve drawing values from a truncated predictive distribution. The development of this method was partly motivated by the analysis of two data sets which are used as illustrations. The sequential regression procedure is applied to perform multiple imputation analysis for the two applied problems. The sampling properties of inferences from multiply imputed data sets created using the sequential regression method are evaluated through simulated data sets.

Key Words: Item nonresponse; Missing at random; Multiple imputation; Nonignorable missing mechanism; Regression; Sampling properties and simulations.

1. Introduction

Incomplete data is a pervasive problem faced by most applied researchers. Several methods have been, and continue to be, developed to draw inferences from data sets with missing values (Little and Rubin 1987). The multiple imputation framework suggested by Rubin (1978, 1987a, 1996) is an attractive option if a data set is to be used by multiple researchers with differing levels of statistical expertise. This approach involves imputing several plausible sets of missing values in the incomplete data set resulting in several completed data sets. Each completed data set is analyzed separately, say by fitting a particular regression model. The resulting inferences – point estimates and the covariance matrices – are then combined using the formula given in Rubin (1987a, Chapter 3) and refinements thereof (Li, Raghunathan and Rubin 1991; Li, Meng, Raghunathan and Rubin 1991; Meng and Rubin 1992; and Barnard 1995).

Imputation based approaches for handling missing data, in general, are quite useful in practice because once the missing values have been imputed, existing complete-data software can be used to analyze the data. Since software development for complete data analysis is keeping pace with the introduction of new statistical methods, applied researchers without knowledge of particular missing data techniques or resources to generate their own code for implementing new missing data procedures will be able to fit finely tuned substantive models for a specific problem at hand. An added advantage of the multiple imputation approach is that by repeatedly applying the complete data software, one can obtain valid point and interval estimates under a fairly general set of conditions (Rubin 1987a). Several researchers (see, for example, the list of references in Rubin 1996) have applied this technique under a variety of settings and have demonstrated, through analysis of simulated and actual data sets, the appropriateness of this approach. Alternatives such as single imputation with an appropriate variance estimation procedure, for example, modified Jackknife Repeated Replication Technique (Rao and Shao 1992) also have this advantage. The imputation approach described in this paper can also be used to create single imputation with an alternative variance estimation procedure.

The development of imputation methods from varying perspectives has a long history (Madow, Nisselson, Olkin and Rubin 1983). A theoretically appealing framework for developing imputation methods is the Bayesian approach. This approach specifies an explicit model for variables with missing values, conditional on the fully observed variables and some unknown parameters, a prior distribution for the unknown parameters, and a model for the missing data mechanism, which does not need to be specified under an ignorable missing data mechanism (Rubin 1976). This explicit model then generates a posterior predictive distribution of the missing values conditional on the observed values. The imputations are draws from this posterior predictive distribution. Several computer programs and

1. Trivellore E. Raghunathan, James M. Lepkowski, John van Hoewyk and Peter Solenberger, University of Michigan, Institute for Social Research, Survey Methodology Program, P.O. Box 1248, Ann Arbor, MI 48106-1248, U.S.A.

algorithms are available for imputing missing values under multivariate normality (Rubin and Schafer 1990), the multivariate *t* distribution (Liu 1995), and several variations of the general location model (Schafer 1997; Raghunathan and Grizzle 1995; and Raghunathan and Siscovick 1996). The latter model can handle the joint distribution of categorical and continuous variables and was first proposed by Olkin and Tate (1961), and used by Little and Schluchter (1985) explicitly for missing data problems. An important property of these approaches is that they are fully conditional on all the observed information. Several simulation studies (for example, Raghunathan and Grizzle 1995) indicate that the inferences drawn from such imputed data have desirable sampling properties.

Survey data sets often consist of large numbers of variables which have a variety of distributional forms. Typically, such data sets have hundreds of variables, some continuous, others counts, many dichotomous or polytomous, and even some semi-continuous or limited dependent variables. Moreover, the distributions of the continuous variables alone may involve normal, lognormal, and other distributions. Postulating a full Bayesian model can be very difficult in this situation. Furthermore, survey data commonly have two additional features that make the modeling process even more complex. First, certain restrictions are imperative. For example, the variable "Number of Years Since Quit Smoking" is defined only for former smokers; hence, the imputation process for this variable should be restricted only to former smokers. Restrictions also arise due to skip patterns in the questionnaire. For example, certain questions about income from a second job are asked only when the respondent indicates that he/she has a second job. The imputation of such variables has to be handled in a hierarchical manner.

Second, there are certain logical or consistency bounds for the missing values that must be incorporated in the imputation process. Such interrelationships among the variables make the model specification difficult. For instance, "Years of Smoking" is restricted to current or past smokers and the imputed values must be less than Age – x years, where x may be chosen based on certain other characteristics, such as evidence of smoking as a teen-ager. For a former smoker, x also includes years since smoking ceased. Another example of bounds is discussed in Heeringa, Little and Raghunathan (1997). They address imputation of bracketed response questions in which a respondent is unable or unwilling to provide an exact response (*e.g.*, income and assets), but does define the bounds within which the imputed values must lie.

The goal of this paper is to propose and evaluate a general purpose multivariate imputation procedure that can handle a relatively complex data structure where explicit full multivariate models cannot be easily formulated but the imputed values for each individual are fully conditional on all the values observed for that individual. The approach is to consider imputation on a variable by variable basis but to condition on all observed variables. The basic strategy creates imputations through a sequence of multiple regressions, varying the type of regression model by the type of variable being imputed. Covariates include all other variables observed or imputed for that individual. The imputations are defined as draws from the posterior predictive distribution specified by the regression model with a flat or non-informative prior distribution for the parameters in the regression model. The sequence of imputing missing values can be continued in a cyclical manner, each time overwriting previously drawn values, building interdependence among imputed values and exploiting the correlational structure among covariates. To generate multiple imputations, the same procedure can be applied with different random starting seeds or taking every P^{th} imputed set of values in the cycles mentioned above.

The variables in the data set are assumed to be of the following five types: (1) continuous, (2) binary, (3) categorical (polytomous with more than two categories), (4) counts and (5) mixed (a continuous variable with a non-zero probability mass at zero). Computationally, binary and categorical variables can be treated identically, but distinguishing them helps in conceptual understanding and in the description of the basic algorithm. We also assume that the population is essentially infinite, the sample is a simple random sample and the missing data mechanism is ignorable (Rubin 1976). The use of multiple imputation in a complex design setting has, as yet, not been fully investigated and is beyond the scope of the current paper.

In this paper we describe the sequential regression multivariate imputation (SRMI) approach in section 2 and evaluate two applications of the approach in sections 3 and 4. In the first application, it is difficult to postulate a joint multivariate distribution because of the complex systematic relationship between the variables and restrictions. In the second application, a general location model can be used to create multiple imputations (Olkin and Tate 1961; and Little and Schluchter 1985). Hence, we compare multiple imputation inferences resulting from the SRMI approach to those resulting from a joint multivariate model. The results of a simulation study investigating the sampling properties of imputed data inferences are presented in section 5, and a concluding discussion with directions for future research are given in section 6.

2. Imputation Method

For a sample of size n, let X denote a $n \times p$ design or predictor matrix containing all the variables with no missing values. X consists of continuous, binary, count or mixed variables, and appropriate dummy variables representing categorical variables. In addition, X may also consist of a column of ones to model an intercept parameter, offset variables, and certain design variables. Let $Y_1, Y_2, ..., Y_k$ denote k variables with missing values, ordered, without

loss of generality, by the amount of missing values, from least to most. The pattern need not be monotone. (In a monotone pattern of missing data, Y_2 is observed only for a subset of subjects on whom Y_1 is observed, Y_3 is observed only for a subset of those on whom Y_2 is observed and so on.)

For model based imputations, the joint conditional density of $Y_1, Y_2, ..., Y_k$ given X can be factored as

$$f(Y_1, Y_2, ..., Y_k \mid X, \theta_1, \theta_2, ..., \theta_k) = f_1(Y_1 \mid X, \theta_1) f_2(Y_2 \mid X, Y_1, \theta_2) ... f_k(Y_k \mid X, Y_1, Y_2, ..., Y_{k-1}, \theta_k) \quad (1)$$

where $f_j, j = 1, 2, ..., k$ are the conditional density functions and θ_j is a vector of parameters in the conditional distribution (*e.g.*, regression coefficients and dispersion parameters). In the sample survey context this can be viewed as a superpopulation model. We model each conditional density through an appropriate regression model with unknown parameters, θ_j, and draw from the corresponding predictive distribution of the missing values given the observed values. We assume that the prior distribution for the parameters $\theta = (\theta_1, \theta_2, ..., \theta_k)$ is $\pi(\theta) \propto 1$ (diffuse relative to the likelihood). However, the method can easily be modified for specified proper prior distributions.

Each conditional regression is based on one of the following models:

1. A normal linear regression model on a suitable scale (for example, a Box-Cox power transformation may be used to achieve normality) if Y_j is continuous;
2. A logistic regression model if Y_j is binary;
3. A polytomous or generalized logit regression model if Y_j categorical;
4. A Poisson loglinear model if Y_j is a count variable; and
5. A two-stage model where zero-non zero status is imputed using logistic regression, and conditional on non-zero status, a normal linear regression model is used to impute non-zero values, if Y_j is mixed.

Each imputation consists of c "rounds". Start round 1 by regressing the variable with the fewest number of missing values, Y_1 on X, imputing the missing values under the appropriate regression model. Assuming a flat prior for the regression coefficients, the imputations, for the missing values in Y_1 are the draws from the corresponding posterior predictive distribution (See Appendix A for a detailed discussion about drawing values for various regression models.) Then update X by appending Y_1 appropriately (for example, dummy variables, if it is categorical) and move on to the next variable, Y_2, with the next fewest missing values. Repeat the imputation process using updated X as predictors until all the variables have been imputed. That is, Y_1 is regressed on $U = X$; Y_2 is regressed on $U = (X, Y_1)$ where Y_1 has imputed values; Y_3 is regressed on $U = (X, Y_1, Y_2)$ where Y_1 and Y_2 have imputed values; and so on.

The imputation process is then repeated in rounds 2 through c, modifying the predictor set to include all Y variables except the one used as the dependent variable. Thus, regress Y_1 on X and $Y_2, Y_3, ..., Y_k$; regress Y_2 on X and $Y_1, Y_3, ..., Y_k$; and so on. Repeated cycles continue for a prespecified number of rounds, or until stable imputed values occur.

The procedure outlined above needs modification to incorporate restrictions and bounds. The restrictions are handled by fitting the models to an appropriate subset of individuals. For example, a Poisson regression model could be applied to impute any missing values for the variable "Number of Pregnancies." The imputation will be restricted to women in the sample. As a covariate, though, this variable may be treated differently when imputing subsequent variables. For instance, certain dummy variables may be created based on this variable, which hare then appended to the matrix U before proceeding with the imputation of the next variable.

Consider another example, "Years Smoking Cigarettes," where the sample would be restricted to current or past smokers. If there is no evidence of smoking as a teenager, "Years Smoking Cigarettes" for a current smoker should satisfy the bound (0, Age - 18). If there is some indication of smoking as a teenager then the range may be restricted to, say (0, Age - 12). For a past smoker these ranges will be (0, Age - 18 – YRSQUIT) and (0, Age - 12 – YRSQUIT) respectively, where YRSQUIT is the years since the individual quit smoking. The appropriate regression model for this variable is a truncated version of the normal linear regression model (possibly on a transformed scale). The parameters, the regression coefficients and the residual variance need to be drawn from the corresponding posterior distributions. The imputations are then drawn from the corresponding truncated normal distribution conditional on the drawn value of the parameters.

It is difficult to draw values of parameters directly from their posterior distribution with truncated normal likelihoods. However, it can be easily computed for a given parameter value. The Sampling-Importance-Resampling (SIR) algorithm (Rubin 1987b, Raghunathan and Rubin 1988) can be used to draw from the actual posterior distribution. First, draw several trial parameter values from the posterior distribution without applying the bounds (untruncated normal linear regression model). Second, attach an importance ratio to each trial value, defined as the ratio of the actual posterior density with bounds to the trial density (the posterior density without bounds), both evaluated at the drawn value. Finally, resample a single parameter value with probability proportional to the importance ratios. This method requires careful monitoring of the distribution of importance ratios (Gelman, Carlin, Stern and Rubin 1995).

The bounds can also be applied to polytomous variables. For instance, suppose that a variable Y can take one of k values, but the observed data suggests that the missing value for a particular subject can either be j or l. The contribution to the likelihood from this subject corresponds to the conditional binomial distribution. The draws in the multinomial step (see Appendix A) are made from the conditional distribution for these two categories. That is, the imputed value is j with probabilities $s_{j^*} = P_{j^*}/(P_{j^*} + P_{l^*})$ and l with probability $1 - s_{j^*}$.

At the completion of the initial round of imputations, the first complete data set with no missing values is available. The factorization in Equation (1) defines a joint conditional distribution of $Y_1, Y_2, ..., Y_k$, given X. If the pattern of missing data is monotone, the imputations in the first round are approximate draws from the joint posterior predictive density of the missing values given the observed values. Note that the draws from the logistic, polytomous, and count variables are from large sample approximations of the posterior density of the regression coefficients. It is possible to improve upon these approximations by using, for example, the SIR algorithm or another rejection algorithm in each subsequent round.

When the pattern of missing data is not monotone, one can develop a Gibbs sampling algorithm (Geman and Geman 1984; Gelfand and Smith 1990) corresponding to Model (1). For example, conditional on the drawn values of the parameters $\theta_2, \theta_3, ..., \theta_k$ and the missing values drawn in the first round, the second round would draw values of θ_1 from the appropriate conditional posterior density which is proportional to the first term in Equation (1). Next draw the missing values in Y_1 conditional on this drawn value of the parameter θ_1, all other observed or imputed values for that subject and other parameters, $\theta_2, \theta_3, ..., \theta_k$ in the model. That is, the missing values in Y_j at round $(t+1)$ need to be drawn from the conditional density,

$$f_j^*(Y_j \mid \theta_1^{(t+1)}, Y_1^{(t+1)}, ..., \theta_j^{(t+1)}, \theta_{j+1}^{(t)}, Y_{j+1}^{(t)}, ..., \theta_k^{(t)}, Y_k^{(t)}, X), \quad (2)$$

computed based on the joint distribution in (1), where $Y_l^{(t)}$ is the imputed or observed values for variable Y_l at round t. Though this is conceptually possible, it is difficult even to compute this density in most practical settings with restrictions, bounds, and the types of variables being considered.

Our proposal is to draw missing values in Y_j at round $(t+1)$ from a predictive distribution corresponding to conditional density,

$$g_j(Y_j \mid Y_1^{(t+1)}, Y_2^{(t+1)}, ..., Y_{j-1}^{(t+1)}, Y_{j+1}^{(t)}, ..., Y_k^{(t)}, X, \phi_j), \quad (3)$$

where the conditional density g_j is specified by one of the regression models described earlier that depends upon the variable type for Y_j, and φ_j is the unknown regression parameters with diffuse prior. That is, the new imputed values for a variable are conditional on the previously imputed values of other variables, and the newly imputed values of variables that preceded the currently imputed variable. This proposal may be viewed as an approximation to an actual Gibbs sampling where the conditional density (2) is approximated by the conditional density (3). Furthermore, this approximation can be improved by considering the SIR or some other rejection type algorithm if the conditional density in (2) can be computed up to a constant.

There are some other particular cases where this approximation is equivalent to drawing values from a posterior predictive distribution under a fully parametric model. For example, if all the variables are continuous and each conditional regression model is a normal linear regression model with constant variance, then the algorithm converges to a joint predictive distribution under a multivariate normal distribution with an improper prior for the mean and the covariance matrix.

It is theoretically possible that a sequence of draws based on densities in (3) may not converge to a stationary distribution, because these conditional densities may not be compatible with any multivariate joint conditional distribution of $Y_1, Y_2, ..., Y_k$ given X (Gelman and Speed 1993). Our empirical investigations using several practical data sets have not identified, so far, any such anomalies. In several large data sets, we find the conditional densities (2) and (3) to be quite similar. As discussed in sections 4 and 5, the draws from this approach are comparable to those based on an explicit Bayesian model.

3. Effect of Smoking on Primary Cardiac Arrest

In our first illustration, the SRMI approach is applied to a case-control study examining the relationship between cigarette smoking and the incidence of primary cardiac arrest (Siscovick, Raghunathan, King, Weinmann, Wicklund, Albright, Bovbjerg, Arbogast, Kushi, Cobb, Copass, Psaty, Retzlaff, Childs and Knopp 1995). In this study it is difficult to formulate an explicit model which captures the full complexity of the data. The case subjects were all King County, Washington residents who had out-of-hospital primary cardiac arrests between 1988 and 1994. The case subjects were identified through a review of paramedic incident reports. Control subjects were selected by random digit dialing from King County and matched to case subjects on gender and age (within seven years). To be eligible, subjects (case and control) were required to be between 25 and 74 years of age, married, and free of clinically-diagnosed heart disease or some other life-threatening conditions such as cancer, liver disease, lung disease, or end-stage renal disease.

Because primary cardiac arrest has a case-fatality rate greater than 80%, the eligibility criterion of marriage was included so that information regarding risk factor exposure (*i.e.*, smoker status, years smoked) could be ascertained from surrogate respondents (*i.e.*, spouses). Among control and surviving cases subjects, both subject and surrogate were interviewed to gather exposure data. The control and

the surviving cases subjects were interviewed mainly to study the reliability of measurements from their surrogates. Among the variables considered in this paper, there were practically no differences in the measurements obtained from the subjects and their surrogates for control or case subjects.

Table 1 gives the means, standard deviations, and percent missing values for key variables by case-control status. The exposure variables are indicator variables for Former Smoker (X_1), Current Smoker (X_2) and Years Smoked (X_3). The confounding variables considered are Age, Body Mass Index (BMI) (BMI = Weight [in Kg]/Height^2[in Meters]), and the binary variables Female and Education (High School Graduate). The substantive model of interest is the logistic regression model,

$$\log[\Pr(C=1)/\Pr(C=0)] = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_1 X_3 + \alpha_4 X_2 X_3 + \alpha_5 \text{Age} + \alpha_6 \text{BMI} + \alpha_7 \text{Female} + \alpha_8 \text{Education},$$

where C is an indicator of cardiac arrest. Preliminary investigations indicated that linear terms for Age and BMI, are appropriate.

Table 1
Means and Proportions (in %) for Key Variables and Percent Missing

Variable	Control (n = 551)		Cases (n = 347)	
	% Missing	Mean (SD)	% Missing	Mean (SD)
Age	0.0	58.4 (10.4)	0.0	59.4 (9.9)
BMI	8.2	25.8 (4.1)	2.6	26.4 (4.6)
Years Smoked	16.8	24.8 (14.7)	5.4	31.7 (13.8)
		Proportion		Proportion
Female	0.0	23.2	0.0	19.9
≥ High School	0.0	76.8	0.0	61.9
Smoking Status				
Never Smoked	0.0	47.2	0.0	27.3
Former Smoker	0.0	42.1	0.0	38.2
Current Smoker	0.0	10.7	0.0	34.5

There are no missing values for the variables Age, Female, Education, Smoking Status (X_1, X_2), and C. Thus, for purposes of imputation, define $X = (1$, Age, Female Education, $X_1, X_2, C)$. Log (BMI), having the fewest missing values, was regressed first on X through a normal linear regression model. Residual diagnostics indicated a log-transform improved the normality of residuals.

Next, Years Smoked was regressed on $U = (X$, log (BMI)). For this variable the sample was restricted to current and former smokers. Moreover, imputed values for Years Smoked were bounded by AGE-18, unless a respondent reported that they smoked in school (SCHSMK), and then they were bounded by AGE-12. For former smokers, imputed values were also bounded by how long ago the respondent had quit smoking (YRSQUIT). Thus, imputed values for former smokers who did not smoke in school were bounded by AGE-18-YRSQUIT, while imputed values for former smokers that did smoke in school were bounded by AGE-12–YRSQUIT. Some subjects (5%) had missing values on the two auxiliary items (SCHSMK, YRSQUIT) which were imputed prior to defining the upper bounds of Years Smoked. The inherent structure of this data set makes it difficult to develop explicitly a joint distribution of the variables with missing values conditional on the completed observed variables. SRMI is thus an appealing approach to handle for this type of data.

In imputing the missing values, we performed 1,000 rounds for each of 25 different starting random seeds resulting in $M = 25$ imputations. The logistic regression model was fit to each imputed data set to obtain maximum likelihood estimates of the regression coefficients and asymptotic covariance matrices.

We used the standard multiple imputation variance formula (Rubin 1987a, Chapter 3) to compute the multiply imputed estimate of the regression coefficients and the covariance matrix. Briefly, suppose that $\hat{\alpha}^{(l)}$ is the estimate of the vector of regression coefficients α in the logistic model, and $V^{(l)}$ its covariance matrix, based on imputed data set l. The multiply imputed estimate of α is

$$\hat{\alpha}_{\text{MI}} = \sum_{l=1}^{M} \hat{\alpha}^{(l)} / M$$

and its covariance matrix is

$$V_{\text{MI}} = \sum_{l=1}^{M} V^{(l)} / M + \frac{M+1}{M} B_M$$

where

$$B_M = \sum_{l=1}^{M} (\hat{\alpha}^{(l)} - \hat{\alpha}_{\text{MI}})(\hat{\alpha}^{(l)} - \hat{\alpha}_{\text{MI}})^t / (M-1).$$

The number of imputations is larger than what is usually recommended. We performed 25 imputations with different random seeds to assess whether the Gibbs style rounds lead us to a region of the imputed values that is very different from the observed data. Graphical displays of the imputed and observed values indicated that none of the imputations in the 25,000 rounds were incompatible with the observed data distribution.

Table 2, the complete-case analysis, gives the point estimates and their standard errors based on subjects with all variables observed. A total of 103 subjects (11.5%) had missing values in one or more predictors. A complete-case analysis, which is generally valid only when the data are missing completely at random was performed after deleting these 103 subjects (See Column 2, Table 2). Logistic regression analyses with a missing data indicator as the dependent variable and a number of completely observed variables as predictors indicated that the data are not missing completely at random. One may expect, therefore, that the complete case estimates and standard errors are biased.

Table 2
Point Estimates (Standard Errors) of Logistic Regression Coefficients for Model of Primary Cardiac Arrest for Complete Cases, SRMI Methods 1* and 2**

Predictor Variables	Complete Case		SRMI			
	(n = 795)		Method 1 (n = 898)		Method 2 (n = 898)	
	Estimate (SE)		Estimate (SE)		Estimate (SE)	
Intercept	-2.922	(0.791)	-2.610	(0.757)	-2.348	(0.627)
Age	0.015	(0.009)	0.015	(0.009)	0.014	(0.008)
Female	-0.007	(0.203)	-0.115	(0.189)	-0.119	(0.177)
Education	-0.448	(0.173)	-0.467	(0.166)	-0.444	(0.133)
BMI	0.056	(0.018)	0.049	(0.013)	0.055	(0.009)
Current Smoker	1.693	(0.569)	2.001	(0.543)	1.998	(0.448)
Former Smoker	0.003	(0.284)	-0.029	(0.262)	-0.011	(0.223)
Current Smoker × Yrs Smoked	-0.003	(0.015)	-0.008	(0.013)	-0.005	(0.011)
Former Smoker × Yrs Smoked	0.019	(0.009)	0.014	(0.009)	0.014	(0.009)

* Method 1 – Imputation restricted to model variables
** Method 2 – Imputation includes model and auxiliary variables

Table 2, SRMI Method 1, gives estimates and their standard errors for SRMI using only the variables in the substantive model. These estimates are quite similar to the complete-case analysis estimates. The multiple imputation standard errors are smaller due to additional subjects with imputed data. There are modest changes in the relationship between smoking and primary cardiac arrest. The complete-case analysis indicates a statistically significant relationship between years smoked and primary cardiac arrest for former smokers, while no such association is indicated in the analysis of multiply imputed data.

One of the advantages of the multiple imputation approach is that the imputation process can use additional variables not in the substantive analysis. Such situations arise when a common research database with many variables is used by different researchers, each using a subset of the variables. The imputation may be carried out for the entire database, where prediction for missing values in each variable borrows strength from all other variables in the data set. Such imputations have been shown to improve efficiency compared to those based only on variables in the particular substantive model (Raghunathan and Siscovick 1996).

Table 2, SRMI Method 2, provides multiple imputation estimates and their standard errors obtained when the entire data set was imputed using 50 additional variables. These included dietary indicators, physiological measures, socio-economic status, and behavioural variables. The point estimates are modestly different for all the variables. The standard errors, though, are considerably smaller when compared to the multiple imputation approach using only variables in the substantive model (SRMI, Method 1). This is not surprising because many of the additional variables such as blood pressure, cholesterol counts, alcohol consumption, and physical activity were highly predictive of BMI and smoking related variables.

4. Parental Psychological Disorders and Child Development

A second illustration examines the effects of parental psychological disorders on several measures of childhood development. Little and Schuchter (1985) analyzed the data using a general location model to obtain maximum likelihood estimates of the parameters of the joint distribution. This general location model was employed to create multiple imputations using Markov Chain Monte Carlo methods (Schafer 1997), producing fully Bayesian model-based multiply imputed data sets. We also created multiple imputations using the SRMI procedure.

The study data consists of 69 families with two children each. Each family was classified into one of the three risk categories: (1) Normal Risk – no parental psychiatric disorders; (2) Moderate Risk – one parent diagnosed with a psychiatric illness or a chronic physical illness; and (3) High Risk – one parent diagnosed with schizophrenia or an affective mental disorder. There are three primary dependent variables of interest: Y_{1c}, number of psychiatric symptoms (dichotomized as high/low) for child c; Y_{2c}, the standardized reading scores for child c; and Y_{3c}, the standardized verbal comprehension score for child c.

We consider three models in investigating the impact of parental psychological disorders on childhood development. The first is a mixed effects logistic regression model:

$$\text{logit}[\Pr(Y_{1ic} = 1)] = \beta_0 + \beta_1 U_{1i} + \beta_2 U_{2i} + \gamma_i,$$

where $Y_{1ic} = 1$ if child c in family i is classified as having a high number of symptoms and 0 otherwise; $U_{1i} = 1$ if family i is classified as a moderate risk group and 0 otherwise; $U_{2i} = 1$ if family i is classified as a high risk group and 0 otherwise; and γ_i are random effects assumed to be identically and independently distributed normal random variables with mean 0 and variance φ_γ^2. This

random effect accounts for intraclass correlation between the two children within the same family. With complete data, this model may be fit by maximizing the numerically integrated likelihood function of $(\beta_0, \beta_1, \beta_2, \varphi_\gamma^2)$ using the Newton-Raphson algorithm and the Gaussian quadrature method for the numerical integration of the likelihood function. These types of models can be easily fit with complete data, but are difficult to fit with missing data.

The second and third regression models relate the child's reading and verbal scores, respectively, to risk group after adjusting for the number of symptoms (Y_1). An investigation of the residuals after a few preliminary rounds or reading and verbal score imputations indicated a log scale was appropriate. Thus, denoting Y_{2ic} and Y_{3ic} as the logarithm of the reading and verbal scores, respectively, for child c in family i, we posited the following mixed effects regression model,

$$Y_{2ic} = \alpha_0 + \alpha_1 U_{1i} + \alpha_2 U_{2i} + \alpha_3 Y_{1ic} + \delta_i + \epsilon_{ic}.$$

where δ_i and ϵ_{ic} are mutually independent normal random variables with mean 0 and variances σ_δ^2 and σ_ε^2 respectively. Again, with no missing data in the covariates, the maximum likelihood estimates of the unknown parameters can be readily obtained using, for example, the PROC MIXED procedure in SAS.

There were no missing values in the classification of the risk groups, and thus we defined $X = (1, U_1, U_2)$. The variables with missing values, Y_{21}, Y_{22}, Y_{31} and Y_{32} were imputed using normal linear regression, and the missing values in Y_{11} and Y_{12} were imputed using logistic regression. We created $M = 25$ SRMIs, repeating the process through 1,000 rounds and 25 different seeds. The SRMI multiply imputed data sets were analyzed and combined using the methods described earlier. To compare these results with the multiply imputed inferences when the imputations are draws from the posterior predictive distribution under the general location model we created 25 imputations under a fully Bayesian model using software developed by Schafer (1997). The point estimates and standard errors for the three models using SRMI and Bayes multiple imputation approaches are presented in Table 3. There are no real meaningful differences between the SRMI estimates and standard errors and those resulting from the Bayesian imputation. Children of parents in the high risk group are approximately 7.8 [exp (2.048)] times more likely to have a high number of symptoms than children with parents in the normal group under the SRMI. The 95% confidence interval for this relative risk is (3.8, 16.0). For the moderate risk, group, the corresponding point and interval estimates are 3.7 and (1.8, 7.8). These estimates may be contrasted with those obtained based on the complete-case analysis (not shown): 7.4 (2.3, 24.2) for the high risk group, and 3.5 (1.0, 11.9) for the moderate risk group (data not shown). Though the point estimates of the relative risks are similar, the complete-case confidence intervals are wider because they are based only on 60% of the observations.

Based on the estimated regression coefficients in Table 3, one can infer, after adjusting, for the number of symptoms, that children in the moderate and high risk groups have lower reading scores, by about 11 points [exp(4.654)-exp(4.654-0.110)], when compared to the normal group. On the other hand, the complete-case analysis estimates a score of 16 points lower for children in the moderate risk group than their counterparts in the normal group, and children in the high risk group score about 19 points lower when compared to the normal group.

The SRMI analysis of verbal scores suggests that the children in the moderate and high risk groups score about 20 and 24 points lower, respectively, than their counterparts in the normal group. However, the complete-case analysis shows the moderate risk group scores lower by 36 points and the high risk group scores lower by about 39 points when compared to the normal group. Thus, the complete-case estimates of the effects of parental psychological disorders on the child's reading and verbal scores are quite different than those obtained by the analysis of the multiply imputed data. This is not surprising because the data on reading and verbal scores are not missing completely at

Table 3
Point Estimates (Standard Errors) of Regression Coefficients for Three Models of Child Development Under SRMI and Bayesian Imputation

Predictor Variables	Imp. Method	Dependent Variable					
		Symptoms		Reading Score		Verbal Score	
Intercept	SRMI	-0.678	(0.256)	4.654	(0.013)	4.873	(0.020)
	Bayes	-0.688	(0.257)	4.556	(0.013)	4.991	(0.021)
High Risk Group	SRMI	2.048	(0.356)	-0.109	(0.022)	-0.191	(0.032)
	Bayes	2.033	(0.350)	-0.108	(0.021)	-0.180	(0.033)
Moderate Risk Group	SRMI	1.289	(0.366)	-0.110	(0.022)	-0.162	(0.033)
	Bayes	1.300	(0.360)	-0.109	(0.023)	-0.167	(0.035)
Symptoms	SRMI	-		0.032	(0.022)	-0.083	(0.032)
	Bayes	-		0.031	(0.019)	-0.080	(0.030)

random and are related to the risk group as well as the number of symptoms of the child.

5. Simulation Study

The analyses described in sections 3 and 4 indicate that sensible results can be obtained by applying the SRMI approach to handling missing values. Nevertheless, it is difficult to conclude based on such case studies whether or not the approach will result in valid inferences in routine applications. A simulation study was designed to investigate the repeated sampling properties of inferences from imputed data sets created with the SRMI approach. Complete data sets were generated from hypothetical populations, and elements deleted under an ignorable missing data mechanism. The deleted values were imputed and differences in summary statistics based on the imputed data sets and the before deletion or full data sets were assessed.

More formally, the strategy:

(1) generated a complete data set which did not agree perfectly with our multiple imputation strategy,
(2) estimated selected regression parameters,
(3) deleted certain values using an ignorable missing data mechanism,
(4) used SRMI to multiply impute the missing values, and
(5) obtained multiply imputed estimates for the regression parameters estimated in step 2.

The differences in the parameter are examined across several independent replications of this strategy.

A total of 2,500 complete data sets with three variables (U, Y_1, Y_2) and sample size 100 were generated using the following models:

1. $U \sim \text{Normal}\,(0, 1)$;
2. $Y_1 \sim \text{Gamma}$ with mean $\mu_1 = \exp(U-1)$ and variance $\mu_1^2/5$; and
3. $Y_2 \sim \text{Gamma}$ with mean $\mu_2 = \exp(-1+0.5U+0.5Y_1)$ and variance $\mu_2^2/2$.

The model for Y_2 in step 3 is the primary regression model of interest with true regression coefficients $\beta_0 = -1, \beta_1 = \beta_2 = 0.5,$ and dispersion parameter $\varphi^2 = 0.5$. For the complete data this model can be fixed using statistical software packages such as GLIM or Splus.

The deletion or missing data mechanisms were as follows:

(1) No missing values in U;
(2) the missing values in Y_1 depend on U through a logistic function logit[Pr(Y_1 is missing)] = 1.5 + U; and
(3) the missing values in Y_2 depend on U and Y_1 through a logistic function logit[Pr(Y_2 is missing)] = 1.5 – 0.5 Y_1 – 0.5U.

These missing data mechanisms generated 22% missing data in Y_1 and 29% missing data in Y_2. The complete-case analysis would have only used 48% of the data.

Since SRMI allows us only to fit a normal linear regression model, the imputations were carried out as follows. Suppose that Y_1 has fewer missing values, and let $Z_1 = (Y_1^{\lambda_1} - 1)/\lambda_1$ be the Box-Cox transformation of the continuous variable. In the first round of imputations, assume that Z_1 has a normal distribution with mean $a_0 + a_1 U$ and variance σ_1^2, where λ_1 was estimated using the maximum likelihood approach, and that $Z_2 = (Y_2^{\lambda_2} - 1)/\lambda_2$ has a normal distribution with mean $b_0 + b_1 U + b_2 Z_1$ and variance σ_2^2, where λ_2 was estimated using maximum likelihood. In the subsequent rounds, U and Z_2 are predictors for Z_1, and U and Z_1 are predictors for Z_2. The estimation of a power transformation using maximum likelihood was automated while fitting each regression model.

For each of the 2,500 simulated data sets with missing values, a total 250 rounds with $M = 5$ different random starts were created using SRMI. For each replicate, the resulting $M = 5$ imputed data sets and the full data set (before deletion) were analyzed by fitting the Gamma model for Y_2 using maximum likelihood. The multiple imputation estimate was constructed as the average of the five imputed data estimates. To assess the differences in the point estimates we computed the standardized difference between the SRMI and full data estimates,

$$\Delta(\beta) = \frac{100 \times \text{abs}\,(\text{SRMI estimate} - \text{Full Data Estimate})}{\text{SE}\,(\text{SRMI Estimate}).}$$

Table 4 gives the mean and standard deviation of $\Delta(\beta)$ for three regression coefficients $\beta_0, \beta_1,$ and β_2 in the model. The SRMI estimates are typically within 8% of the full standard units. The actual coverage and the average length of the 95% SRMI confidence intervals were computed for the regression coefficients using the t reference distribution described in Rubin (1987b). For each simulated data set and parameter, it was determined whether or not the true value $(e.g., \beta_1 = 0.5)$ is contained within the corresponding interval. The proportion of intervals containing the true values were computed across the 2,500 replications and are provided in Table 4. For the full data sets, the actual coverage for β_1, for example, was 94.9% and for SRMI it was 95.4. In addition the average length of the confidence intervals were also computed. The average width of the full data confidence interval for β_1 was 0.91 and for SRMI the average length was 1.22. That is, the SRMI data resulted in well calibrated intervals estimates.

The same simulation study was also used to compare the distributional properties of imputations from SRMI and a fully Bayesian method. For the model assumptions used to generate complete data, we developed a Markov Chain Monte-Carlo algorithm for drawing values from the actual posterior predictive distribution of the missing values given

the observed values. Each step of the draw used Metropolis-Hastings algorithm and required considerably more computational time than the SRMI method. Therefore, only the first 500 simulated data sets were used in this comparison. We computed two Kolmogrove-Smimoff (KS) statistics from each simulated data set: One comparing the imputations from the SRMI method and the actual hidden values and the other comparing the Bayesian imputations and the actual hidden values. There were no discernible differences in these two statistics across the 500 simulated data sets. A scatter plot of those 500 pairs of KS statistics showed a narrow scatter of points around a 45 degree line.

Table 4
Means and Standard Deviations for Standardized Differences Between SRMI Estimates and Full Data Estimates and Actual Coverage of Nominal 95% Confidence Intervals

Regression Coefficient	Std. Difference		Confidence Coverage	
	Mean	SD	SRMI	Full Data
β_0	8.2	2.0	96.1	95.4
β_1	8.8	1.7	95.4	94.9
β_2	8.0	2.2	95.3	94.7

6. Discussion

We have described and evaluated a sequential regression multivariate imputation procedure that can be used to impute missing values in a variety of complex data structures involving many types of variables, restrictions, and bounds. This procedure should be useful when the specification of a joint distribution of all the variables with missing values is difficult. A real advantage of the approach is its flexibility in handling each variable on a case by case basis. For instance, to preserve all the bivariate correlations, all the main effect terms must be included as regressors, and to preserve, say, three factor interactions all two factor interactions must be included as regressors in the imputation model. Implementation of this procedure only requires a good random number generator and fitting routines for a variety of multiple regression routines. A SAS based application implementing this approach can be downloaded from a web site (www.isr.umich.edu/ src/smp/ive).

In certain instances, one can modify the algorithm to reduce it to Gibbs sampling from the joint predictive distribution of the missing values given the observed values. However, the SRMI procedure will be more useful where an explicit model is difficult to formulate. In both the illustrations and the simulation, different random starts were used to monitor imputed values, an important aspect in many practical applications. This is a good practice when Gibbs sampling is used under an explicit Bayesian model (Gelman and Rubin 1992) and should be used when the sequential regression method discussed in this paper is used.

The simulation study described in section 5, though limited, is favorable as far as inferences based on the SRMI are concerned. The imputations from SRMI and Bayes model were comparable. The goal here, however, was to develop an imputation approach that is finely tuned on a variable by variable basis fully conditional on all the observed information, rather than an explicit joint multivariate distribution of all the variables. Furthermore, model sensitivity may be reduced by using a semiparametric regression model for each conditional regression. The Bayesian interpretation of the spline smoothing models (Silverman 1985) can be used to draw imputed values from the predictive distribution. Such modifications also deserve further investigation.

For some large data sets with many variables, the SRMI can be computationally intense. The algorithm can be modified to apply a variable selection method for each regression in each round. We compared the inferences with and without the variable selection on several large data sets such as the National Health Interview Survey and the National Medical Expenditure Survey using several hundred variables. The descriptive inferences as well as inferences based on linear and logistic regression models were very similar, still further detailed investigation is needed.

It is also possible to use the imputation approach discussed in this paper in conjunction with, for example, the Jackknife Repeated Replication (JRR) technique for variance estimation. Specifically, (1) re-impute, singly, the missing values in each jackknife replicate SRMI; (2) analyze the imputed replicate data set; and, finally, (3) combine the replicate estimates to obtain the point estimate and its covariance matrix. This approach is more computationally intensive than the multiple imputation approach. This integrated JRR imputation approach and several of its variations are currently under investigation.

Finally, it has been assumed that the data set arises from a simple random sample design. However, most surveys employ complex sample designs involving stratification, clustering, and weighting. Further work is needed to modify the sequential regression method to incorporate complex design features not reflected in the X variables in expression (1). However, even if the imputation process ignores the complex design features, the analysis of completed data should be design based. Though this does not provide valid design-based inferences, it maintains the robustness underlying the design-based analysis to a certain degree. The integrated JRR imputation approach discussed above may have more appealing design-based properties in a complex design setting.

Acknowledgements

The authors would like to thank the three referees for their careful reading of this article and their helpful suggestions. The research was partially supported by a NSF grant DMS–0803720.

Appendix

Regression Models and Imputations

Dropping the subscript indexing of the variables for brevity, the necessary steps for imputing each type of variable are as follows:

Continuous variable: For Y (possibly transformed from the original scale for normality), a continuous variable, build a normal linear regression model, $Y = U\beta + e$, where U is the most recently updated predictor matrix, e has a multivariate normal distribution with mean zero and variance $\sigma^2 I$, and I is an identity matrix. Suppose that $\theta = (\beta, \log\sigma)$ has a uniform prior distribution over the appropriate dimensional real space. Fit this model based on the individuals for whom Y is observed.

Let $B = (U^t U)^{-1} U^t Y$ be the estimated regression coefficient, $\text{SSE} = (Y - UB)^t (Y - UB)$ be the residual sum of squares and $\text{df} = \text{rows}(Y) - \text{cols}(U)$ be the residual degrees of freedom, and T be the Cholesky decomposition such that $TT^t = (U^t U)^{-1}$. The relevant posterior distributions can be derived easily (see, for example, Gelman, Carlin, Stern and Rubin 1995, Chapter 7), and the following steps then provide draws from the posterior predictive distribution of missing Y values:

1. Generate a chi-square random deviate u with df degrees of freedom and define $\sigma_*^2 = \text{SSE}/u$.
2. Generate a vector $z = (z_1, z_2, ..., z_p)$ of dimension $p = \text{rows}(B)$ of random normal deviates and define $\beta_* = B + \sigma_* Tz$.
3. Let U_{miss} denote the U-matrix for those with missing Y values. The impute values are $Y_* = U_{\text{miss}}\beta_* + \sigma_* v$, where v is an independent vector of dimension rows (U_{miss}) of random normal deviates.

Binary Variable: When Y is a binary variable, fit a logistic regression model relating Y to U (most recently updated), $\text{logit}[\Pr(Y = 1 | U)] = U\beta$, using individuals with observed Y. The imputed values for Y are created through the following steps:

1. Let B denote the maximum likelihood estimates of β and V its asymptotic covariance matrix (negative inverse of the observed Fisher information matrix). Let T be the Cholesky decomposition of V (that is, $TT^t = V$). Generate a vector z of random normal deviates of dimension rows (B). Define $\beta_* = B + Tz$.
2. Let U_{miss} denote the portion of U for which Y is missing. Define $P_* = [1 + \exp(-U_{\text{miss}}\beta_*)]^{-1}$. Generate a vector u, of dimension rows (U_{miss}) of uniform random numbers between 0 and 1. Impute 1 if a particular component of u is less than or equal to the corresponding component of P_* and impute 0 otherwise.

This approach results only in approximate draws from the posterior predictive distribution of the missing values as the draws of the parameter β are from the asymptotic approximation of its actual posterior distribution. It is possible to draw from the actual distribution by modifying Step 1 using, for example, Sampling-Importance-Resampling (Rubin 1987b).

Mixed Variable: For Y, a mixed variable (that is, Y either takes the value zero or a continuous value), model the zero values by a 0-1 indicator to distinguish between 0 and non-zero values, and then model a normally distributed variable for the continuous portion of the distribution conditional on the indicator variable being equal to 1. That is, use a two stage approach: impute a one or zero using the logistic approach described above; and then restricting the sample to those with non-zero values, use the continuous variable approach described above to impute a continuous value to replace the just imputed value of 1.

Count Variable: For Y, a count variable, fit a Poisson regression model $Y \sim \text{Poisson}(\lambda)$ where $\log\lambda = U\beta$. The imputations for missing values in Y are created using the following steps:

1. Let B denote the maximum likelihood estimate of β, V its covariance matrix and T the Cholesky decomposition of V. Generate a vector z of random normal deviates of dimension rows (B) and define $\beta_* = B + Tz$.
2. Let U_{miss} denote the portion of U for which Y is missing. Define $\lambda_* = \exp(U_{\text{miss}}\beta_*)$. Generate independent Poisson random variables with means as the elements of λ_*.

Polytomous Variable: For Y that can take k values, $j = 1, 2, ..., k$, let $\pi_j = \Pr(Y = j | U)$. Fit a polytomous regression model relating Y to U where $\log = (\pi_j / \pi_k) = U\beta_j$ for $j = 1, 2, ..., k-1$. Under the restriction $\sum_j^k \pi_j = 1$, it follows that $\pi_k = (1 + \sum_j^{k-1} \exp(U\beta_j))^{-1}$.

Let B denote the maximum likelihood estimate of the regression coefficients $(\beta_1^t, \beta_2^t, ..., \beta_{k-1}^t)$, V be the asymptotic covariance matrix and T its Cholesky decomposition.
The following steps create imputations:

1. Define $\beta_* = B + Tz$ where z is a vector of random normal deviates of dimension rows (B).
2. Let U_{miss} denote the rows of U with missing Y and let $P_i^* = \exp\{U_{\text{miss}}\beta_{i^*}\} / \{1 + \sum_i \exp(U_{\text{miss}}\beta_{i^*})\}$ where β_{i^*} is the appropriate elements of β_* where $i = 1, 2, ..., k-1$ and $P_k^* = 1 - \sum_i P_i^*$.
3. Let $R_0 = 0, R_j = \sum_i^j P_i^*$ and $R_k = 1$ be the cumulative sums of the probabilities. To impute values generate random uniform number u and take j as the imputed category if $R_{j-1} \le u \le R_j$.

Again, the imputation of mixed, count and categorical variables are from approximate posterior predictive distributions because the corresponding parameters are drawn from their asymptotic normal approximate posterior distributions.

References

Barnard, J. (1995). Cross-Match Procedures for Multiple Imputation Inference: Bayesian Theory and Frequentist Evaluation. Unpublished Doctoral Thesis, University of Chicago, Department of Statistics.

Gelfand, A.E., and Smith, A.M.F. (1990). Sampling based approaches to calculating marginal densities. *Journal of American Statistical Association*, 85, 398-409.

Gelman, A., Carlin, J., Stern, H. and Rubin, D.B. (1995). *Bayesian Data Analysis*. London. Chapman and Hall.

Gelman, A., and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, 7, 457-472.

Gelman, A., and Speed T.P. (1993). Characterizing a joint probability distribution by conditionals. *Journal of Royal Statistical Society*, B, 55, 185-188.

Geman, S., and Geman, D. (1984). Stochatic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.

Heeringa, S.G., Little, R.J.A. and Raghunathan, T.E. (1997). Imputation of Multivariate Data on Household Net Worth. University of Michigan, Ann Arbor, Michigan.

Li, K.H., Meng, X.L., Raghunathan, T.E. and Rubin, D.B. (1991). Significance levels from repeated *p* values from multiply-imputed data. *Statistica Sinica*, 1, 65-92.

Li, K.H., Raghunathan, T.E. and Rubin, D.B. (1991). Large sample significance levels from multiply imputed data using moment-based statistics and an F reference distribution. *Journal of American Statistical Association*, 86, 1065-1073.

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

Little, R.J.A., and Schluchter, M.D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika*, 72, 497-512.

Liu, C. (1995). Missing data imputation using the multivariate *t* distribution. *Journal of Multivariate Analysis*, 53, 139-158.

Madow, W.G., Nisselson, H., Olkin, I. and Rubin, D.B. (1983). *Incomplete Data in Sample Surveys*. 1, 2, and 3, New York: Academic Press.

Meng, X.L., and Rubin, D.B. (1992). Performing likelihood ratio tests with multiply imputed data sets. *Biometrika*, 79, 103-111.

Olkin, I., and Tate, R.F. (1961). Multivariate correlation models with mixed discrete and continuous variables. *Annals of Mathematical Statistics*, 32, 448-465.

Raghunathan, T.E., and Grizzle, J.E. (1995). A split questionnaire survey design. *Journal of American Statistical Association*, 90, 54-63.

Raghunathan, T.E., and Rubin, D.B. (1988). An application of Bayesian statistics using sampling/importance resampling to a deceptively simple problem in quality control. *Data Quality Control*: *Theory and Pragmatics*, (Eds., G.E. Liepins and V.R.R. Uppuluri). New York: Marcel Dekker.

Raghunathan, T.E., and Siscovick, D.S. (1996). A multiple imputation analysis of a case-control study of the risk of primary cardiac arrest among pharmacologically treated hypertensives. *Applied Statistics*, 45, 335-352.

Rao, J.N.K., and Shao, J. (1992). Jackknife variance estimation with survey data under hot-deck imputation. *Biometrika*, 79, 811-822.

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

Rubin, D.B. (1978). Multiple imputation in sample surveys – A phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 20-34.

Rubin, D.B. (1987a). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.

Rubin, D.B. (1987b). The SIR-algorithm – A discussion of Tanner and Wong's. The calculation of posterior distributions by data augmentation. *Journal of American Statistical Association*.

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

Rubin, D.B., and Schafer, J.L. (1990). Efficiently creating multiple imputations for incomplete multivariate normal data. *Proceeding of the Statistical Computing Section of the American Statistical Association*, 83-88.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data by Simulation*. New York: Chapman and Hall.

Silverman, B.W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting. *Journal of Royal Statistical Society*, B, 47, 1-52.

Siscovick, D.S., Raghunathan, T.E., King, I., Weinmann, S., Wicklund, K.G., Albright, J., Bovbjerg, V., Arbogast, P., Kushi, L., Cobb, L., Copass, M.K., Psaty, B.M., Retzlaff, B., Childs, M. and Knopp, R.H. (1995). Dietary intake and cell-membrane levels of long-chain n-3 polyunsaturated fatty acids and the risk of primary cardiac arrest. *Journal of American Medical Association*, 274, 1363-1367.