Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Multiple Edit/Multiple Imputation for Multivariate Continuous Data

Bonnie Ghosh-Dastidar & Joseph L Schafer

To cite this article: Bonnie Ghosh-Dastidar & Joseph L Schafer (2003) Multiple Edit/Multiple Imputation for Multivariate Continuous Data, Journal of the American Statistical Association, 98:464, 807-817, DOI: 10.1198/016214503000000738

To link to this article: https://doi.org/10.1198/016214503000000738

Multiple Edit/Multiple Imputation for Multivariate Continuous Data

Bonnie GHOSH-DASTIDAR and Joseph L. SCHAFER

Multiple imputation replaces an incomplete dataset with $m > 1$ simulated complete versions that are analyzed separately by standard methods. We present a natural extension of multiple imputation for handling the dual problems of nonresponse and response error. This extension, which we call multiple edit/multiple imputation (MEMI), replaces an observed dataset containing missing values and errors with $m > 1$ simulated versions of the *ideal* dataset that is complete and error-free. These ideal data sets are analyzed separately, and the results are combined using the same rules as for multiple imputation. The resulting inferences simultaneously reflect uncertainty due to nonresponse and response error. MEMI may be an attractive alternative to deterministic or quasi-statistical edit and imputation procedures used by many data-collecting agencies. Producing MEMI's requires assumptions about the distribution of the ideal data, the nature of nonresponse, and a model for the response error mechanism. However, fitting such a model does not necessarily require data from a follow-up study. In this article we develop and implement MEMI for preliminary data from the Third National Health and Nutrition Examination Survey, Phase I (1988–1991). Raw body measurements for 1,345 children age 2–3 years are imputed under a Bayesian model for intermittent or semicontinuous errors. The resulting population estimates are found to be quite insensitive to prior assumptions about the rates and magnitude of errors.

KEY WORDS: Gibbs sampling; Markov chain Monte Carlo; Missing data; NHANES III; Response error; Semicontinuous.

1. INTRODUCTION

1.1 Background

A measurement error describes a discrepancy in the observed value and the true value that it attempts to measure. In survey sampling, measurement error in data collected from human respondents is usually called *response error* (Biemer, Groves, Lyberg, Mathiowetz, and Sudman 1991; Lessler and Kalsbeek 1992). Surveys often attempt to measure demographic, economic, or other characteristics for which objective, well-defined true values may exist. Recalling and recording the true values may pose considerable challenges to data collection, however.

An example of this can be found in the Third National Health and Nutrition Examination Survey (NHANES III), designed to collect information about the health and diet of people in the United States. In NHANES III, medical staff obtained various physical measurements (e.g., height, weight, waist, and hip circumference, skinfolds). When the standard protocol was followed, these characteristics were measured accurately with negligible amounts of error. Occasionally, however, mistakes or deviations from the intended protocol introduced intermittent errors of large magnitude in one or a few of the variables. Similar patterns of intermittent response error may be detected in other continuous, multivariate data with strong intervariable correlations.

Response errors may complicate the analysis of survey data. However, if ignored, they may lead to biases in estimates produced, as well as underestimation of their variances. Most analysts approach the problem of response error in one of three ways. The first approach, which is generally undesirable, is to do nothing and proceed as if the data were error-free. The second approach is to describe the data with a statistical model that includes the effect of measurement error by an explicit component of variability (e.g., Lord and Novick 1968). A third approach is to use data editing to "clean up" the data by removing gross errors and then analyze the edited data as if it were the truth (e.g., Granquist and Kovar 1997). The second and third approaches are discussed here.

The basis of most measurement-error models in survey sampling is an additive relationship that treats the observed response as the true value plus an error (Biemer and Stokes 1991). The mean of the error distribution is sometimes assumed to be 0, but it may be shifted to accommodate response bias. It is also assumed that the errors are uncorrelated with each other and with the true values. The bias may be estimated by acquiring more accurate measurements by performing a record check or a reinterview study, whereas measurement variability may be estimated by simple repetition of the original survey (Fuller 1987; Bollen 1989; Groves 1989).

Measurement error modeling from reinterview data is attractive in principle, but often not feasible in practice. More commonly, survey organizations use manual or automated edit procedures developed by subject matter specialists to determine whether an observation is reliable (Barcaroli and Venturi 1993; Thompson and Sigman 1999). An influential article by Fellegi and Holt (1976) led to the development of such major systems as the Census Bureau's Structured Program for Economic Editing and Referrals (Winkler and Draper 1997), and Statistics Canada's Generalized Edit and Imputation System (Kovar and Whitridge 1990). These systems are somewhat general and yet must be reconfigured for each specific survey. As a result, they have met with, only limited success.

1.2 Motivating Example

The motivating example for this research is NHANES III, conducted by the National Center for Health Statistics (1994).

Bonnie Ghosh-Dastidar is Statistician, RAND Corporation, Santa Monica, CA 90407-2138 (E-mail: *bonnieg@rand.org*). Joseph L. Schafer is Associate Professor, Department of Statistics, Pennsylvania State University, University Park, PA 16802 (E-mail: *jls@stat.psu.edu*). This research was supported by National Science Foundation grant SBR 93-10101; additional support for manuscript preparation was provided by RAND. The authors thank Meena Khare of the National Center for Health Statistics for providing the dataset analyzed in the article, through a contract between the second author and the National Center for Health Statistics. The authors thank Dan McCaffrey, Brian Williams, the editor the associate editor, and three referees for comments that helped improve the article. Finally, the authors acknowledge the help provided by Stef van Buuren, Catherine Cruz, J. R. Lockwood, Sally Morton, and William Winkler.

Journal of the American Statistical Association
December 2003, Vol. 98, No. 464, Applications and Case Studies
DOI 10.1198/016214503000000738

Unlike many surveys, NHANES includes a large number of biological and health measurements for sampled adults and children. As a result, NHANES data are heavily used to influence policy aimed at improving the health of the U.S. population. In this application, we use raw body measurements available from physical examinations of 1,345 children age 2–3 years. These measurements are preliminary Phase I (1988–1991) data made available by the NCHS. The variables discussed here are weight (WT), standing height (HT), sitting height (SITHT), recumbent length (RECUM), and head circumference (HEADC), all of which are continuous and moderately to strongly correlated. The only variable that requires an explanation is RECUM, which refers to a full body length measurement taken whith the child reclining. The units of measurement are kilograms for WT and centimeters for the rest. This dataset also has a large number of design variables. We use AGE, SEX, RACE, and the geographic indicator SMSA in this application. AGE has two levels in this dataset of 2 to 3 year-olds, RACE has three levels (white/other, black, and Mexican-American), and SMSA has two levels: urban and rural.

While preparing the NHANES IIII Phase I data for imputation, Ezzati-Rice, Khare, and Schafer (1993) found that although most observations appeared to be essentially accurate, a handful of observations were clearly unusual. Data editing using univariate and bivariate rules was found to be inadequate because it missed some of the gross outliers, which had values that seemed plausible but violated multivariate relationships. Furthermore, data editing ignores any uncertainty in the edit process. Therefore, measurement error modeling seems preferable to data editing. However, measurement error models for continuous data have typically assumed continuous errors. In contrast, the NHANES III errors are often (essentially) zero and continuously distributed otherwise, and thus are inherently semicontinuous (Olsen and Schafer 2001).

For this reason, we assume a semicontinuous model for the intermittent response errors and a multivariate normal model for the underlying true body measurements. We adopt a Bayesian approach to generate simulated draws of the true values conditional on the observed values. The computations needed to analyze data under a semicontinuous error model are far more complicated than those needed with a continuous model. As a result, we turn to computationally intensive Markov chain Monte Carlo (MCMC) simulation techniques. We summarize the simulation results with a new method called multiple edit/multiple imputation (MEMI), a natural extension of multiple imputation. Multiple imputation replaces each missing value in a dataset with two or more plausible values. MEMI extends the idea of multiple imputation by performing stochastic editing two or more times. The end result of MEMI is a multiply imputed data product that reflects uncertainty due to response error as well as nonresponse.

1.3 Extending Multiple Imputation

Multiple imputation, originally proposed by Rubin (1976), was designed to handle the problem of missing data in public-use databases where the database constructor and the ultimate user may be distinct entities. It is a Monte Carlo approach that produces multiple plausible values for the missing data under an explicit probability model for the observed data and assumptions about the nonresponse mechanism. Each set of imputed values results in a complete dataset that can be analyzed by standard methods. The results of the complete-data analyses are then combined to produce inferences that reflect missing-data uncertainty (Rubin 1987, pp. 76–79). In many cases, good results can be obtained with only three to five to five imputations (Schafer 1997, pp. 106–107). Once multiple imputations have been generated, there is no further need for specialized missing-data methods or software. Multiple imputation is especially attractive in situations where the imputer has access to more information than the analyst (Rubin 1996).

The same principles that make imputation attractive for missing data make it also attractive for measurement error. Given a raw dataset subject to gross errors and nonresponse, MEMI will simulate multiple versions of an *ideal* sample, without response error or missing values, to be analyzed using standard methods and the results combined with the same procedures as in multiple imputation. Hypothesis testing can be conducted using estimates of the tail areas available from the empirical distribution function of the simulated values (Schafer 1997, pp. 96–97). Measurement-error models are no longer needed, because information on the error process is encapsulated in the distribution of the simulated ideal datasets.

In the remainder of this article, we describe and implement a specific example of MEMI without data from a follow-up study. Section 2 provides the framework for MEMI and a discussion of ignorability and validity of MEMI inference. Section 3 gives the details of a MCMC algorithm for continuous data with intermittent errors. Section 4 gives results for preliminary body measurements from NHANES III, Phase I, and Section 5 discusses assumptions and potential extensions of MEMI.

2. MULTIPLE EDIT/MULTIPLE IMPUTATION

2.1 General Setup

Suppose that the survey data, $\mathbf{Y}$, are a distorted version of the unobserved ideal sample, $\mathbf{X}$. The goal of MEMI is to produce plausible versions of $\mathbf{X}$ given $\mathbf{Y}$, under explicit probability models for $\mathbf{X}$ and the error mechanism. Let θ and Ψ be unknown parameters of the ideal data model and the error distribution. The form of the ideal-data model, $P(\mathbf{X}|\theta)$, and the relationship between the observed and true values, $P(\mathbf{Y}|\mathbf{X}, \Psi)$, are assumed known. When fully observed covariate measurements $\mathbf{U}$ are present (from, e.g., the sample frame), $\mathbf{X}$ may be modeled as a multivariate regression on $\mathbf{U}$. We implicitly condition on $\mathbf{U}$ throughout, so that the ideal data model is written as $P(\mathbf{X}|\theta)$ rather than as $P(\mathbf{X}|\mathbf{U}, \theta)$.

When the survey data have missing values in addition to response errors, they may be partitioned as $\mathbf{Y} = (Y_{obs}, Y_{mis})$, where Y_{obs} and Y_{mis} denote the observed and missing parts. Let $\mathbf{R}$ be a fully observed matrix of response indicators whose elements are 0 or 1, depending on whether the corresponding elements of $\mathbf{Y}$ are missing or observed. In general, the distribution of $\mathbf{R}$ may be expected to depend on $\mathbf{X}$ and $\mathbf{Y}$, so that we write the nonresponse mechanism as $P(\mathbf{R}|\mathbf{X}, \mathbf{Y}, \phi)$ where ϕ represents unknown parameters. The observed data consist of both Y_{obs} and $\mathbf{R}$, so the observed-data probability density

function (pdf) is given by

$$P(\mathbf{R}, Y_{obs}|\theta, \Psi, \phi) = \int\int P(\mathbf{X}|\theta)\, P(\mathbf{Y}|\mathbf{X}, \Psi)\, P(\mathbf{R}|\mathbf{X}, \mathbf{Y}, \phi)\, d\,\mathbf{X}\, dY_{mis}. \quad (1)$$

We assume that the missing-data mechanism depends only on the observed data, so that $P(\mathbf{R}|\mathbf{X}, \mathbf{Y}, \phi) = P(\mathbf{R}|Y_{obs}, \phi)$ and (1) simplifies to

$$P(\mathbf{R}, Y_{obs}|\theta, \Psi, \phi) = P(\mathbf{R}|Y_{obs}, \phi)\, P(Y_{obs}|\theta, \Psi). \quad (2)$$

If θ and Ψ are of interest and ϕ is a nuisance parameter, then $P(\mathbf{R}|Y_{obs}, \phi)$ becomes a proportionality constant, and likelihood inferences for θ and Ψ may be based solely on $P(Y_{obs}|\theta, \Psi)$, referred to as the *observed-data likelihood*. Bayesian inferences about θ and Ψ are based on an observed-data posterior density for θ and Ψ, which is proportional to the product of the observed-data likelihood $P(Y_{obs}|\theta, \Psi)$ and a prior density $P(\theta, \Psi)$,

$$P(\theta, \Psi|Y_{obs}) \propto P(Y_{obs}|\theta, \Psi)\; P(\theta, \Psi). \quad (3)$$

MEMI is motivated by Bayesian arguments and samples from $P(\mathbf{X}|Y_{obs})$, the posterior predictive distribution of $\mathbf{X}$ given Y_{obs}. Appealing to Bayes' theorem yields

$$P(\mathbf{X}|Y_{obs}) = \int\int P(\mathbf{X}|Y_{obs}, \theta, \Psi)\, P(\theta, \Psi|Y_{obs})\, d\theta\, d\Psi. \quad (4)$$

MEMI's, which we denote by $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(m)}$, are independent draws from (4). When the survey data have no missing values, $\mathbf{Y} \equiv Y_{obs}$ and Y_{mis} is empty, so that the complete-data situation is a special case of the more general setup presented here. The resulting m versions of the ideal data are analyzed separately by complete-data methods, and the results are combined using simple rules to obtain inferences that effectively incorporate uncertainty due to missing data and response error (Rubin 1987, pp. 76–79).

2.2 Ignorability

In missing-data problems, assumptions about the nonresponse mechanism are necessary. We assume that the missing-data mechanism is ignorable in the sense defined by Rubin (1976, p. 582). First, this implies that $\mathbf{R}$ depends only on the observed data,

$$P(\mathbf{R}|\mathbf{X}, \mathbf{Y}, \phi) = P(\mathbf{R}|Y_{obs}, \phi), \quad (5)$$

and not on the unobserved quantities $\mathbf{X}$ and Y_{mis}. This assumption is known as *missing at random* (MAR). Note that the MAR assumption implies that the nonresponse is independent of the response errors. The second part of ignorability assumes that the parameter sets (θ, Ψ) and ϕ are distinct. From a Bayesian perspective, this means that any joint prior distribution applied to (θ, Ψ, ϕ) must factor into independent priors for (θ, Ψ) and ϕ.

2.3 Validity of Multiple Edit / Multiple Imputation Inferences

In MEMI, the imputer and the analyst may be distinct entities. As a resut, while the imputer may assume a fully parametric model, the ensuing analysis may be nonparametric or design based. A natural question to ask is whether MEMI inferences are valid when the ideal-data model and the analyst's model are uncongenial, that is, do not correspond (Meng 1994). One possible inconsistency occurs when the ideal-data model and the analyst's model differ because the imputer assumes more than the analyst. When the additional assumptions made by the imputer are true, the MEMI point estimates and standard errors should be more efficient. Although the models are uncongenial, MEMI inferences are still valid. When the additional assumptions made by the imputer are false, MEMI inferences may be adversely affected.

A second type of inconsistency arises when the analyst assumes more than the imputer. In this case the inferences will be valid with a slight loss of power if the assumptions made by the analyst are true. When the analyst's additional assumptions are incorrect, the results of the inferences may be invalid. Thus MEMI inferences will be valid if the assumptions made by the analyst and the imputer are true. Further discussion of congeniality and choice of imputation models has been given by Rubin (1996) and Schafer (1997, chap. 4).

3. MULTIPLE EDIT/MULTIPLE IMPUTATION FOR CONTINUOUS DATA WITH INTERMITTENT ERRORS

3.1 Model Specification

Let $\{\mathbf{y}_i : i = 1, \ldots, n\}$ be a sample of survey data where $\mathbf{y}_i$ represents a vector of p measurements on unit i. Under a true-score model, each $\mathbf{y}_i$ is modeled as the sum of an underlying true value $\mathbf{x}_i$ and an independent error $\boldsymbol{\epsilon}_i^*$, both of length p,

$$\mathbf{y}_i = \mathbf{x}_i + \boldsymbol{\epsilon}_i^*. \quad (6)$$

As in standard measurement error models, we assume a priori independence between $\mathbf{x}_i$ and $\boldsymbol{\epsilon}_i^*$. We model $\mathbf{x}_i$ as multivariate normal; future work will explore departures from normality. We assume that $\mathbf{x}_i \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}_i$ is a linear function of known covariates $\mathbf{u}_i$, $\boldsymbol{\mu}_i = \boldsymbol{\beta}^T \mathbf{u}_i$. With q covariates, $\mathbf{u}_i$ is a vector of length q and $\boldsymbol{\beta}$ is a $q \times p$ matrix of coefficients. Under a multivariate normal assumption, the likelihood function for n independent observations of the ideal data $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$, if they were seen, would be

$$\mathrm{L}(\theta|\mathbf{X}) = |2\Pi\boldsymbol{\Sigma}|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_i)\right\}, \quad (7)$$

where Π is the known constant of proportionality and $\theta = (\boldsymbol{\beta}, \boldsymbol{\Sigma})$ is unknown.

The intermittent response errors, $\boldsymbol{\epsilon}_i^{*\,T} = (\epsilon_{i1}^*, \ldots, \epsilon_{ip}^*)$, are generated by a semicontinuous distribution such that ϵ_{ij}^* is 0 with probability $(1 - \pi_j)$ and $\epsilon_{ij}^* \sim N(0, \tau_j)$ with probability π_j. The mixing probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_p)$ describe the proportion of error in variables $1, \ldots, p$, and $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_p)$ are the variances of those errors. Let $\mathrm{I}(\epsilon_{ij}^*)$ be an indicator

random variable such that $I(\epsilon_{ij}^*) = 1$ when $\epsilon_{ij}^* = 0$. The pdf of ϵ_{ij}^* given $\Psi = (\boldsymbol{\pi}, \boldsymbol{\tau})$ is

$$p(\epsilon_{ij}^*|\Psi) = (1-\pi_j)I(\epsilon_{ij}^*) + \pi_j(2\Pi\tau_j)^{-\frac{1}{2}} \exp\{-\epsilon_{ij}^{*2}/2\tau_j\}. \quad (8)$$

We assume that the errors $\epsilon_{i1}^*, \ldots, \epsilon_{ip}^*$ associated with sample unit i are mutually independent and also independent of $\mathbf{X}$. Under these assumptions, the marginal distribution of $\mathbf{y}_i$ is a high-dimensional mixture of normals with 2^p components (Titterington, Smith, and Makov 1985; Peel and McLachlan 2000).

3.2 Previous Work on Multivariate Outlier Models

Outliers are observations that deviate from the specified data model. The presence of outliers may indicate that the data model does not have sufficiently heavy tails, or that there are real errors in the data. In MEMI, we assume that our model is reasonably correct; points identified as not from the specified model are assumed to be erroneous. A review of parametric modeling of outliers in continuous data revealed extensive literature on univariate outlier models, but little pertaining to multivariate situations. The mutivariate applications generally involve robust parameter estimation. Little and Smith (1987) proposed an expectation-robust algorithm to compute robust parameter estimates under a multivariate normal model. Lange, Little, and Taylor (1989) considered a heavy-tailed model, applying a multivariate-t distribution with low degrees of freedom to datasets with potential outliers. Parameter estimation under this model downweights extreme observations. However, the multivariate-t model generates a continuum of unusual values rather than the intermittent errors encountered in NHANES III.

Little (1988) examined another model, a contaminated normal, in which observations are drawn from a mixture of two multivariate normal distributions with the same mean but different covariance matrices, one proportionately larger than the other. The outliers are fit by the normal component with larger covariances. The mixing probability determines the proportion of the observed data expected to be outliers. For data with intermittent errors, the contaminated normal seems more appropriate than the multivariate-t. However, the contaminated normal assumes a single contamination rate and constant inflation of variance for all variables. In contrast, measurement error modeling allows us to model the errors on a variable-by-variable basis, with different rates of contamination $\pi_1, \ldots, \pi_p$ and different error variances $\tau_1, \ldots, \tau_p$ across the variables.

3.3 Choosing Prior Distributions

As described in Section 2.1, producing MEMIs requires prior distributions for $\theta = (\boldsymbol{\beta}, \boldsymbol{\Sigma})$, and $\Psi = (\boldsymbol{\pi}, \boldsymbol{\tau})$. The priors for θ and Ψ are constructed in a hierarchical fashion. First, independent priors for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are specified, and then a prior for $\boldsymbol{\tau}$ conditional on $\boldsymbol{\Sigma}$ is assumed. Finally, a prior for π is specified independently of the other parameters,

$$p(\theta, \Psi) = p(\boldsymbol{\beta}, \boldsymbol{\Sigma})\, p(\boldsymbol{\tau}|\boldsymbol{\Sigma})\, p(\boldsymbol{\pi}). \quad (9)$$

We assume an improper uniform prior for $\boldsymbol{\beta}$ and the standard noninformative improper prior for $\boldsymbol{\Sigma}$, so that $P(\boldsymbol{\beta}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(p+1)/2}$ (Box and Tiao 1992, p. 426).

Given θ, the conditional priors of $\tau_1, \ldots, \tau_p$ are assumed to be independent, scaled inverse chi-squared distributions with known degrees of freedom $\nu_1, \ldots, \nu_p$. When the scale of the variables is quite different, it may be reflected in the size of the response errors. We model this feature by making the error variance, τ_j, proportional to the variance of the ideal data, σ_{jj}. Thus the scale parameter of the prior is set equal to σ_{jj} times a positive scalar c_j. The prior $\tau_j \sim c_j\sigma_{jj}\chi_{\nu_j}^{-2}$ is like proposing ν_j hypothetical erroneous measurements for variable j, with sample variance $c_j\sigma_{jj}/\nu_j$. The pdf of $\tau_j|\sigma_{jj}$ is proportional to

$$p(\tau_j|\sigma_{jj}) \propto \tau_j^{-(\nu_j/2+1)} \exp\{-c_j\sigma_{jj}/2\tau_j\}, \qquad \tau_j > 0, \quad (10)$$

where $\nu_j, c_j, \sigma_{jj} > 0$.

A standard beta prior is specified for π_j, so that $\pi_j \sim \text{beta}(a_j, b_j)$. The hyperparameters a_j and b_j are chosen so that the prior mean, $a_j/(a_j + b_j)$, is small, indicating a low a priori probability of error. The sum $(a_j + b_j)$ plays the role of a imaginary number of prior observations, and a_j plays the role of a imaginary number of errors in a sample of size $(a_j + b_j)$. The pdf of π_j is proportional to

$$p(\pi_j) \propto \pi_j^{a_j-1}(1-\pi_j)^{b_j-1}, \qquad 0 < \pi_j < 1, \quad (11)$$

where $a_j,\ b_j > 0$.

3.4 A Hybrid Markov Chain Monte Carlo Algorithm

Our goal is to produce MEMI's by drawing from $P(\mathbf{X}|Y_{obs})$ multiple times, under the probability models discussed in Sections 3.1 and 3.3. The complicated form of $P(\mathbf{X}|Y_{obs})$ makes this distribution difficult to sample from directly. Our strategy for carrying out MEMI is based on computationally intensive simulation methods of MCMC, namely Gibbs sampling and the Metropolis–Hastings algorithm (Gilks, Richardson, and Spiegelhalter 1996). MCMC methods are attractive because they can be implemented in complicated problems.

Let the survey data be denoted by $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^T$ and the corresponding errors by $\boldsymbol{\epsilon}^* = (\boldsymbol{\epsilon}_1^*, \ldots, \boldsymbol{\epsilon}_n^*)^T$. For the sake of convenience, $\boldsymbol{\epsilon}^*$ and $\mathbf{X}$ may be partitioned as $\boldsymbol{\epsilon}^* = (\epsilon_{obs}^*, \epsilon_{mis}^*)$ and $\mathbf{X} = (X_{obs}, X_{mis})$, where ϵ_{obs}^* are response errors and X_{obs} are true values corresponding to Y_{obs}. Similarly, ϵ_{mis}^* and X_{mis} are response errors and true values, corresponding to Y_{mis}. Our method of generating samples from $P(\mathbf{X}|Y_{obs})$ exploits the relationship between $\mathbf{X}$, $\boldsymbol{\epsilon}^*$, and $\mathbf{Y}$ given by (6); thus simulating $P(\mathbf{X}|Y_{obs})$ is equivalent to simulating $P(\epsilon_{obs}^*, X_{mis}|Y_{obs})$.

Our algorithm simulates $P(\epsilon_{obs}^*, X_{mis}|Y_{obs})$ by a Gibbs sampler. Drawing repeatedly from the full conditional distributions of ϵ_{obs}^*, X_{mis}, θ, $\boldsymbol{\pi}$, and $\boldsymbol{\tau}$ eventually yields draws from the joint distribution $P(\epsilon_{obs}^*, X_{mis}, \theta, \boldsymbol{\pi}, \boldsymbol{\tau}|Y_{obs})$. The first step of this Gibbs sampler is computationally expensive; therefore, we replace it by a local chain of Metropolis–Hastings (Carlin and Louis 2000). Under this proposed scheme, ϵ_{mis}^* is not simulated. We also exclude the unit nonrespondents from the computational part of the algorithm, because they provide no additional information, and simply impute them at the end.

3.5 Posterior Distributions

In this section we derive the full conditionals in the Gibbs sampler described earlier. The full conditional of θ depends

only on $\mathbf{X}$. Under a multivariate regression model for the ideal data $\mathbf{X}$ and a semicontinuous error model (8), the posterior distribution of θ factors into two distinct pieces corresponding to the posterior distributions of $\boldsymbol{\Sigma}$ and $\boldsymbol{\beta}|\boldsymbol{\Sigma}$. The posterior density of $\boldsymbol{\beta}|\mathbf{X}, \boldsymbol{\Sigma}$ is a multivariate normal,

$$\text{vec}(\boldsymbol{\beta})|\mathbf{X}, \boldsymbol{\Sigma} \sim \text{N}\big(\text{vec}(\hat{\boldsymbol{\beta}}), (\mathbf{U}^T\mathbf{U})^{-1} \otimes \boldsymbol{\Sigma}\big). \quad (12)$$

The posterior of Σ is given by

$$\boldsymbol{\Sigma}|\mathbf{X} \sim \text{W}^{-1}\big(n-q, (\hat{\boldsymbol{\gamma}}^T\hat{\boldsymbol{\gamma}})^{-1}\big), \quad (13)$$

an inverse Wishart density with degrees of freedom $(n-q)$ and a $p \times p$ symmetric, positive definite scale matrix $\hat{\boldsymbol{\gamma}}^T\hat{\boldsymbol{\gamma}}$. In this notation, $\mathbf{A} \sim \text{W}(\nu, \mathbf{B})$ denotes an ordinary Wishart distribution with ν degrees of freedom and scale matrix $\mathbf{B}$, that is, $\mathbf{A} = \mathbf{X}^T\mathbf{X}$, where $\mathbf{X}$ consists of ν rows drawn from $\text{N}(\mathbf{0}, \mathbf{B})$; then $\mathbf{A}^{-1} \sim \text{W}^{-1}(\nu, \mathbf{B})$.

The conditional distributions of π_j and τ_j, the parameters of the error model, depend only on $\boldsymbol{\epsilon}^*$. These calculations are facilitated by writing the semicontinuous errors as $\epsilon^*_{ij} = z_{ij}\epsilon_{ij}$, the product of an unobserved Bernoulli random variable z_{ij} with marginal distribution $p(z_{ij}=1) = \pi_j$, and a continuous error $\epsilon_{ij} \sim \text{N}(0, \tau_j)$. Now consider a vector of p Bernoulli variables, $\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{ip})^T$, associated with $\mathbf{y}_i$. Let $\mathbf{Z}$ be a $n \times p$ matrix, $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n)^T$. Under an assumption of independent errors, the pdf of $\mathbf{Z}$ is

$$P(\mathbf{Z}|\boldsymbol{\pi}) = \prod_{j=1}^{p}\big[{\pi_j}^{d_j}(1-\pi_j)^{n-d_j}\big], \quad (14)$$

where $d_j = \sum_{i=1}^{n} z_{ij}$ is the number of errors in variable j.

Note that if $\boldsymbol{\epsilon}^*$ is known, then $\mathbf{Z}$ is known; further, conditioning on $\mathbf{Z}$ makes $\boldsymbol{\pi}$ independent of $\boldsymbol{\epsilon}^*$. Ignoring the events $(z_{ij}=1,\ \epsilon_{ij}=0)$, which occur with probability 0, we can simply combine the prior density of π_j (11) with the likelihood of $\mathbf{Z}$ (14) to yield the posterior of π_j,

$$\pi_j|\mathbf{Z} \sim \text{beta}(a_j - 1 + d_j, b_j - 1 + n - d_j), \quad j = 1, \ldots, p. \quad (15)$$

Now $\boldsymbol{\tau}$ can be directly estimated from the nonzero elements of $\boldsymbol{\epsilon}^*$. Let $\mathcal{I}_j$ denote the subset of indices $\{i = 1, \ldots, n\}$ for which $z_{ij} = 1$. Combining the likelihood of the nonzero errors (8) with the prior of τ_j (10) yields a scaled inverse chi-squared posterior density with scale parameter $(c_j\sigma_{jj} + \sum_{i\in\mathcal{I}_j}\epsilon^{*2}_{ij})$ and degrees of freedom $\nu_j + d_j$,

$$\tau_j|\boldsymbol{\epsilon}^*, \mathbf{Z}, \boldsymbol{\Sigma} \sim \left(c_j\sigma_{jj} + \sum_{i\in\mathcal{I}_j}\epsilon^{*2}_{ij}\right)\chi^{-2}_{\nu_j+d_j}, \quad (16)$$

independently for $j = 1, 2, \ldots, p$.

Under an assumption of independence, the full conditional of X_{mis} may be simulated in n independent steps, drawing from the conditionals $P(\mathbf{x}_{i(mis)}|\mathbf{x}_{i(obs)}, \theta, \boldsymbol{\pi}, \boldsymbol{\tau})$, $i = 1, \ldots, n$, in turn. The posterior predictive distribution of $\mathbf{x}_{i(mis)}$ is multivariate normal with parameters that are functions of $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}$. Similarly, the full conditional of ϵ^*_{obs} can be simulated in n independent steps. However, the conditional distribution of $\epsilon^*_{i(obs)}$ under the semicontinuous error model is complicated; therefore, we simulate $\boldsymbol{\epsilon}^*_{i(obs)}$ by first drawing Bernoulli indicators $\mathbf{z}_{i(obs)}$ and then drawing continuous errors ϵ_{ij} whenever $z_{ij} = 1$. The posterior predictive distribution of the continuous errors is multivariate normal, whereas the distribution of $\mathbf{z}_{i(obs)}$ is discrete with 2^{p_i} support points, where p_i is the length of $\mathbf{z}_{i(obs)}$. Direct sampling of the conditional of $\mathbf{z}_{i(obs)}$ is computationally inefficient due to the large number of probability masses that must be calculated for all of the support points. Therefore, this step of the Gibbs sampler is replaced with a single step of Metropolis–Hastings (Ghosh-Dastidar 1998).

4. APPLICATION TO THE NATIONAL HEALTH AND NUTRITIONAL EXAMINATION SURVEY III

4.1 Data Setup

We apply the MEMI algorithm to preliminary data from NHANES III, Phase I, described in Section 1.2. The design variables in this dataset are fully observed; however, the body measurement variables have nontrivial rates of unit (9.2%) and item nonresponse (1.7%–3.1%), and thus listwise deletion would omit 16% of the cases. The unit nonresponse is due primarily to failure to show up for the medical examination. The body measurement variables were assessed for normality, and data transformations were considered for skewed variables. We applied a logarithmic transformation to WT, referred to as LGWT. Another variable, SITHT, has slight negative skewness, but we left it untransformed, implicitly saying that the unusually small measurements of SITHT are probably erroneous. The multivariate regression model specified here included the full interaction of AGE, SEX, and RACE and only the main effects for SMSA.

Preliminary examination of this dataset revealed a handful of gross outliers in the body measurements (Fig. 1). Because of the moderate sample size of this dataset and a relatively small number of errors, we used prior distributions $\pi_j \sim \text{beta}(1, 99)$, yielding an a priori mean contamination rate of 1% for each variable. The parameters for the prior distribution of τ_j were chosen to reflect an a priori error variance of σ_{jj} and a small prior sample size for the errors; thus $\tau_j \sim 5\,\sigma_{jj}\chi^{-2}_5$. We have assumed the same values of a_j, b_j, ν_j, and c_j for each variable; however, different values could have been assumed if prior information had led us to believe otherwise. Although the choice of priors may seem somewhat arbitrary, sensitivity analysis suggests that the population estimates produced by MEMI are robust to prior assumptions.

We proceeded to run the MCMC algorithm with the model and prior specifications discussed earlier. Exploratory time series and sample autocorrelation function plots of the simulated values suggested that the chains of π_j and τ_j are sticky, with autocorrelations that do not die out until lag 200. But there were no discernible autocorrelations for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ beyond lag 10. We decided on a conservative burn-in period of 500 iterations, and spaced the MEMI's 250 iterations apart to ensure approximate independence between subsequent draws. Therefore, we required a MCMC chain of 3,000 iterations to generate 10 MEMI's for this application. Parameter estimates for π_j and τ_j were calculated from a chain of 100,000 iterations, minus the burn-in iterations.

4.2 Population Estimates

The NHANES III is designed to collect information about the health and diet of the general population, as well as age-

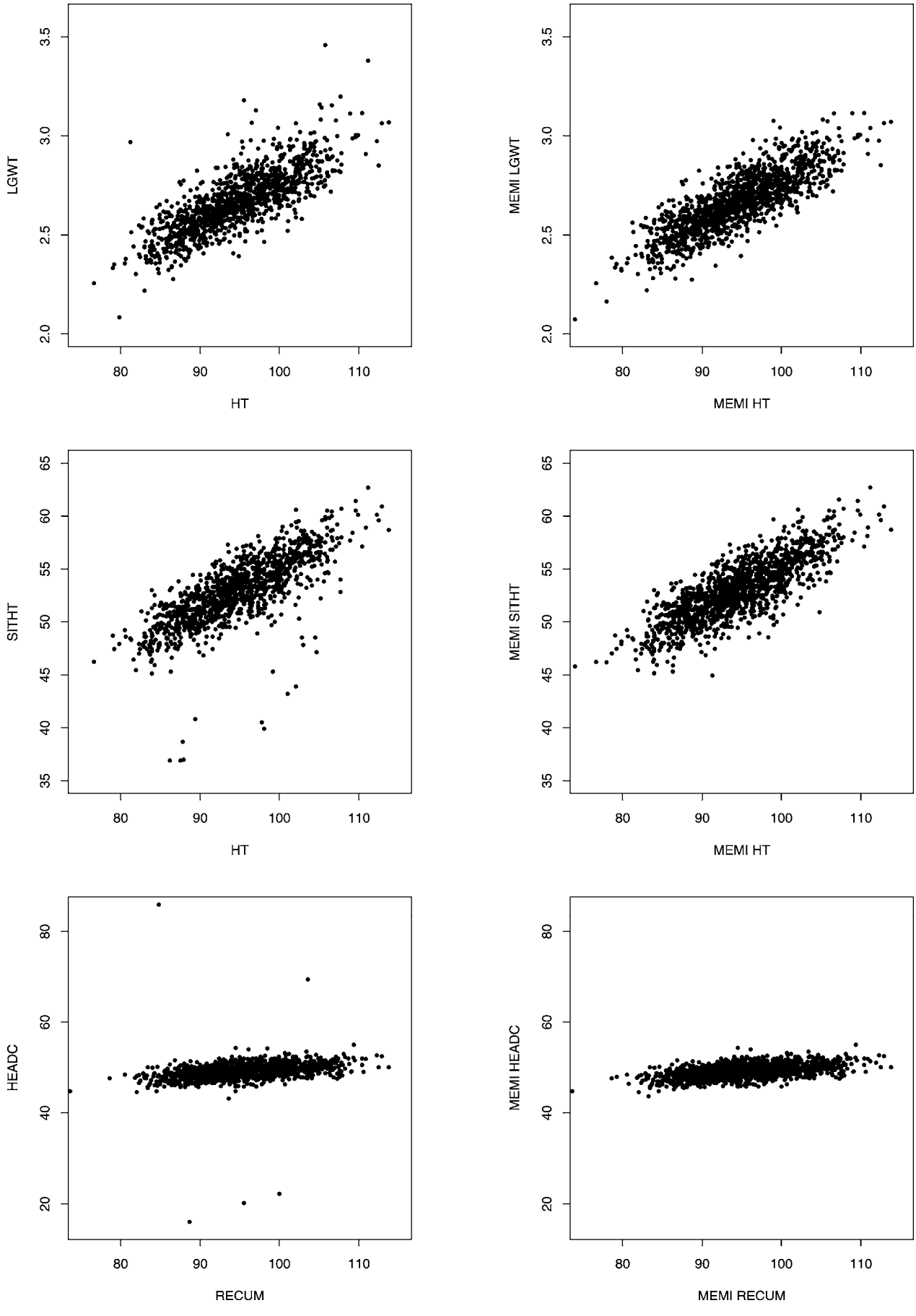

Figure 1. Scatterplots of Body Measurement Variables Before and After Editing. Plots of the unedited data are on the left; plots of the first MEMI dataset are on the right side.

Table 1. MEMI Results From $m = 10$ Imputations

Estimand	*Estimates*	*SE*	*df*	*95% interval*	*100r*	*100$\hat{\lambda}$*
Overall mean BMI	16.21	.043	159	(16.13, 16.30)	31.2	24.7
Male White	16.19	.090	720	(16.01, 16.37)	12.6	11.4
Black	16.16	.093	1,913	(15.98, 16.34)	7.4	7.0
Mexican-American	16.68	.137	40	(16.41, 16.96)	89.3	49.6
Female White	16.08	.094	126	(15.89, 16.27)	36.5	27.9
Black	15.70	.105	2,049	(15.49, 15.91)	7.1	6.7
Mexican-American	16.40	.110	84	(16.19, 16.62)	48.5	34.2
Overall mean HEADC	49.27	.045	346	(49.18, 49.36)	19.2	16.6
Male White	49.93	.099	370	(49.73, 50.12)	18.5	16.1
Black	49.73	.106	689	(49.52, 49.94)	12.9	11.7
Mexican-American	49.61	.108	10,397	(49.40, 49.82)	3.0	3.0
Female White	48.97	.112	54	(48.75, 49.20)	68.7	42.8
Black	49.06	.120	1,117	(48.83, 49.30)	9.9	9.1
Mexican-American	48.49	.092	496	(48.31, 48.67)	15.6	13.8

specific and ethnic groups, in the United States. It is used to produce population estimates of such quantities as body mass index (BMI) and head circumference, which are leading indicators of obesity and normal health functioning in U.S. children. Note that BMI is the ratio of body weight in kilograms to squared height in meters. We generated ideal-data sample means and standard errors for BMI and HEADC for our entire sample of children, and within cells of SEX $\times$ RACE, for each MEMI. We combined the results of the individual inferences using Rubin's rules of multiple-imputation inference to get the overall results, which are given in Table 1.

Table 1 lists strata definitions, MEMI point estimates, and standard errors (SE's) in the first three columns. The multiple-imputation degrees of freedom, df, required for calculation of the 95% t intervals, is one column over. The next column, $100r$, gives the percentage relative increase in variance due to nonresponse and response error. If the observed data were ideal, without missing values and response error, then this quantity would be 0%. The last column, $100\hat{\lambda}$, gives the estimated percent of missing information. Again, when the data are ideal, the percent of missing information is 0%.

It appears that, on average, BMI is lower for girls than for boys. Of the three races, Mexican-Americans have the highest mean BMI for both males and females, followed by whites and then blacks. The fraction of missing information associated with mean BMI differs for the three levels of RACE, with Mexican-Americans having the highest values and blacks having the lowest. In the case of HEADC, SEX appears to have a strong effect, with the average HEADC of males being substantially higher than that of females for all three levels of RACE. RACE, on the other hand, seems to have little effect on HEADC. The fraction of missing information associated with mean HEADC is particularly high for white females.

In the context of MEMI's, the fraction of missing information reflects uncertainty due to both nonresponse error and response error. Thus a large value of $\hat{\lambda}$ associated with mean HEADC of white females is partially explained by the fact that the nonresponse rates for white females is the highest among all six levels of SEX $\times$ RACE. In addition, there are very large errors in the HEADC measurements of white females, which increase the amount of missing information (Fig. 2).

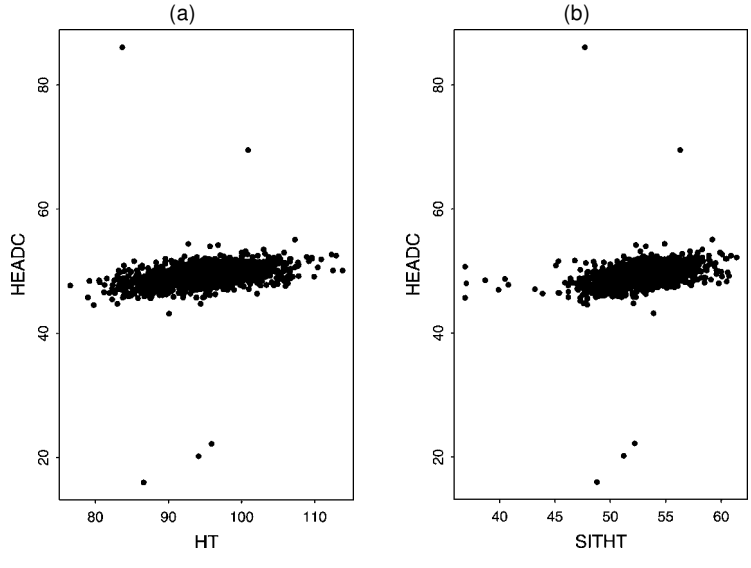

Figure 2. Gross Response Error in HEADC Values of White Females. Plotted against: (a) HT; (b) SITHT.

Table 2. Posterior Means and 95% Intervals of π_j and $\sqrt{\tau_j/\sigma_{jj}}$

	WT	*HT*	*SITHT*	*RECUM*	*HEADC*
$100\pi_j$	$3.2_{(1.4,5.9)}$	$2.1_{(1.0,3.5)}$	$2.5_{(1.4,4.1)}$	$.7_{(.1,1.8)}$	$.8_{(.3,1.5)}$
$\sqrt{\tau_j/\sigma_{jj}}$	$1.64_{(1.2,2.3)}$	$1.09_{(.8,1.5)}$	$2.63_{(2.0,3.6)}$	$1.12_{(.7,1.9)}$	$13.55_{(9.2,20.9)}$

4.3 Other Products

Although the key product of MEMI inference is population estimates for estimands of interest, it also can produce parameter estimates of the error mechanism. By averaging the simulated values of the contamination rates π_j and error variances τ_j across iterations of the MCMC algorithm, we obtain simulated posterior means for these parameters (Table 2). The estimated contamination rate is largest for WT, whereas the estimated error variance is largest for HEADC. Because the scale of measurement for the variables is quite different, we also looked at estimates of $\sqrt{\tau_j/\sigma_{jj}}$. HEADC had the largest ratio of $\sqrt{\tau_j/\sigma_{jj}}$, implying that there are large response errors in the HEADC measurements. Simulated posterior credible sets for π_j and τ_j were calculated using quantiles of the chains of MCMC output. The wide intervals reflect large amounts of uncertainty in the estimation of these quantities.

MEMI can also estimate posterior probabilities of error for individual cases and suggest likely error patterns. This provides information about potentially contaminated cases and their joint probabilities of contamination, which can be used to identify suspect values within cases. Consider a binary representation of the error pattern with 1 indicating an error and 0 indicating no error, with the indicators listed in the order WT, HT, SITHT, RECUM, and HEADC. Results for a few cases are provided for illustration. Case 1 has an estimated posterior probability of .92 associated with the pattern *00000*, which indicates that it has no errors, whereas case 109 appears to have an erroneous value of HEADC, with the most likely pattern being *00001*. Finally, HT and SITHT appear to be jointly erroneous for case 708, because the most likely pattern is *01100*.

4.4 Sensitivity of Results to Priors

Here we verify whether MEMI inferences are sensitive to assumptions about priors. We first tried six different scaled inverse chi-squared prior distributions for the τ_j's by varying c_j and ν_j, holding those for π_j constant. We let ν_j be 5 or 15 and chose values of c_j so that the a priori estimates of the ratio τ_j/σ_{jj} ranged between $1/5$ and 5. We drew 10 MEMI's for each of the priors. For comparison, we computed the complete-data sample mean and standard error of HEADC from the MEMI's. The results of the combined MEMI inference for mean HEADC are shown in Table 1. It is important to remember that when sampling from stationary distributions, there will naturally be fluctuation in results due to random variability. Taking pure randomness into account, it appears that the results of MEMI inference are not sensitive to prior assumptions about τ_j (Table 3).

We repeated our study with six different priors for π_j, keeping ν_j and c_j fixed. We let the a priori sample size of $a_j + b_j$ be either 10 or 100 and set a_j to .2, 1, or 5 when $a_j + b_j$ equaled 100, or .02, .1, or .5 when $a_j + b_j$ equaled 10. This yielded a priori means of .002, .01, and .05, respectively. The MEMI population estimates again were very little affected by changes made to the prior of π_j (Ghosh-Dastidar 1998). Because the choice of priors for the τ_j's and π_j's is somewhat subjective, we investigated how the posterior distributions of these parameters are affected by prior assumptions. We considered the two groups of priors mentioned earlier. We produced boxplots of the simulated draws of π_j and $\sqrt{\tau_j/\sigma_{jj}}$ from their respective posterior distributions using 100,000 iterations, with the first 500 discarded. We concluded that the posterior distributions of τ_j and π_j are, relatively speaking, more sensitive to assumptions about priors than the results of MEMI inference.

4.5 Univariate and Bivariate Editing

Survey agencies typically use data editing to clean survey data. Therefore, we compared MEMI to results obtained from univariate and bivariate editing. First, we screened the raw body measurements one variable at a time, using a robust 1.5-interquartile range (IQR) rule. Note that the IQR represents the absolute distance between the first and third quartiles. The lower and upper outlier detection bounds here are $Q_1 - 1.5 \times$ IQR and $Q_3 + 1.5 \times$ IQR, beyond which values are flagged as outlying. We then generated scatterplots for multiple pairs of variables; points that violated the correlation pattern were flagged using the *Identify* function in S-PLUS. We inspected suspect values on a case-by-case basis. The IQR rule identified 51 cases (3.8% overall), and 2% of the WT, .2% of the HT, 1.3% of the SITHT, .2% of the RECUM, and 1.4% of the HEADC values as potential outliers. The bivariate plots flagged 44 cases, of which 24 were gross outliers; 16 of the 44 had passed univariate editing. When suspect values were removed, the tails of the marginal distributions (especially HEADC) were substantially pulled in and the standard deviations decreased, but the means and medians did not change. The correlations strengthened, with the correlation of HEADC increasing from .28 to .34 with HT and from .10 to .17 with BMI.

Table 3. Posterior Means and Standard Errors of HEADC for Several Priors of τ_j Holding Prior for π_j Constant

	$\nu_j = 5$			$\nu_j = 15$		
Mean HEADC	$c_j = 1$	$c_j = 5$	$c_j = 25$	$c_j = 3$	$c_j = 15$	$c_j = 45$
Male White	$49.93_{.101}$	$49.96_{.096}$	$49.94_{.097}$	$49.94_{.104}$	$49.93_{.096}$	$49.93_{.094}$
Black	$49.74_{.106}$	$49.72_{.104}$	$49.70_{.102}$	$49.71_{.114}$	$49.71_{.110}$	$49.72_{.103}$
Mexican-American	$49.61_{.113}$	$49.61_{.110}$	$49.60_{.113}$	$49.61_{.124}$	$49.60_{.116}$	$49.61_{.114}$
Female White	$48.97_{.090}$	$48.95_{.095}$	$48.96_{.087}$	$48.95_{.095}$	$48.97_{.093}$	$48.96_{.094}$
Black	$49.05_{.120}$	$49.06_{.119}$	$49.06_{.116}$	$49.05_{.119}$	$49.05_{.119}$	$49.04_{.117}$
Mexican-American	$48.53_{.091}$	$48.51_{.093}$	$48.54_{.098}$	$48.51_{.090}$	$48.53_{.092}$	$48.51_{.087}$

Similarly, we derived a MEMI-based edit rule that flagged cases with an estimated posterior probability of error of .5 or greater as errors. We then looked for agreement among the three edit rules. We identified 22 cases as erroneous by all three; generally, these were the gross outliers in the plots (Fig. 1). Of the 51 cases flagged by univariate editing, we identified 55% as potentially erroneous by MEMI, with 23 likely to have only 1 error (often HEADC) and 5 with 2 or more errors (often HT and SITHT). Of the 44 cases flagged in bivariate editing, a higher percentage (68%) was likely to have 1 or more erroneous values under MEMI. Overall, MEMI and bivariate editing had better agreement. Although the IQR rule flagged a larger number of cases, many of these looked fine on inspection. Moreover, univariate edits missed values that appeared plausible but violated intervariable relationships. Tracking a case across multiple scatterplots was also difficult, and cases missing a value were omitted from the plots. MEMI was the only one to reflect uncertainty in data editing. Also, editing this relatively small dataset was sufficiently labor intensive to suggest the need for automated edit and imputation systems such as MEMI.

4.6 Parameter Simulation

MEMI is not the only valid approach that can be used to generate the population estimates in Table 1; likelihood inference is equally appealing. We chose to use the model and prior assumptions of MEMI for the likelihood approach because it seemed to yield the most parallel comparison between the two methods. Assuming that the data model, response-error model, and ignorability assumptions are correct, all relevant statistical information about the parameters is contained in the observed-data posterior, (3), within a Bayesian framework. Due to the complicated form of this likelihood function, we used Gibbs sampling to simulate draws from the posterior distributions of the parameters. These can be used to provide summaries of the posterior likelihood function. We performed 10,000 draws of the regression parameters $\boldsymbol{\beta}$ from their observed-data posterior using the Gibbs sampler described in Sections 3.4 and 3.5; we discarded the initial 500 burn-in values. We computed posterior means of HEADC within strata defined by SEX $\times$ RACE from these simulated values; we used quantiles of the simulated distribution to estimate an equal-tailed 95% Bayesian posterior

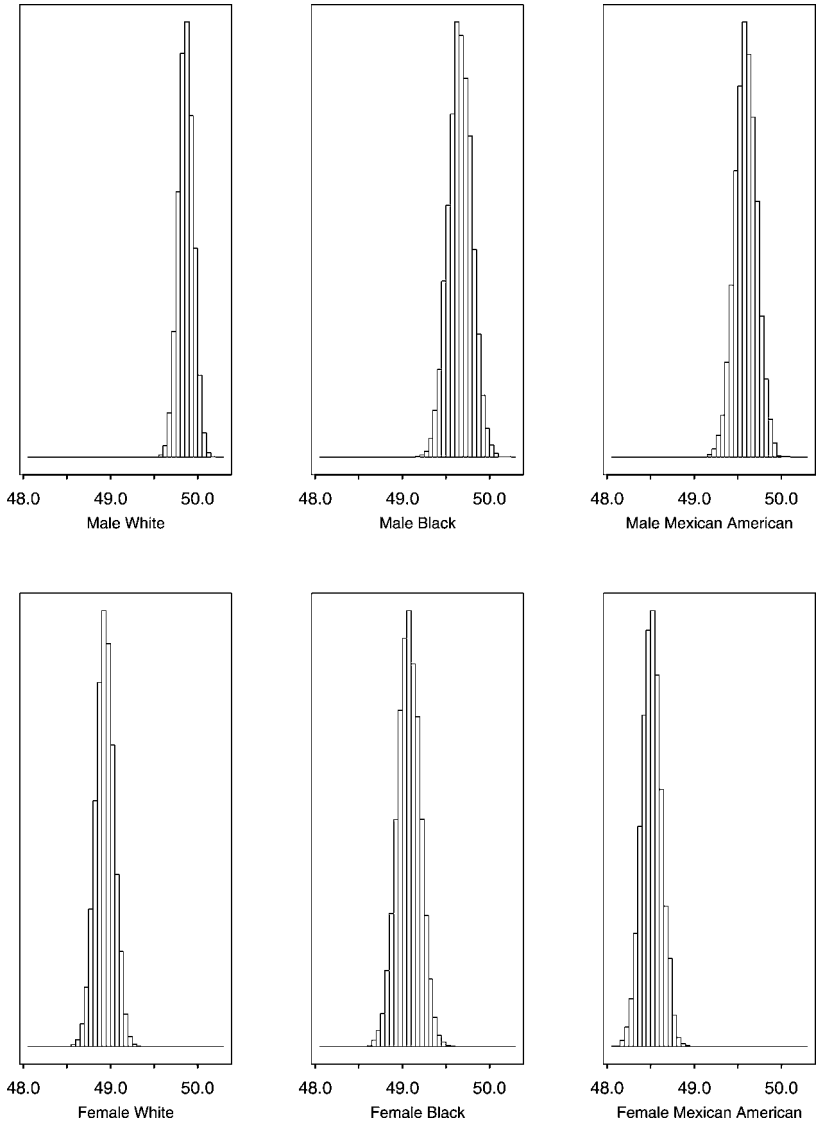

Figure 3. Histograms of Posterior Simulated Values of HEADC.

Table 4. Posterior Stratified Means of HEADC From n = 10,000 Iterations

Strata	Estimates	95% interval
Male white	49.86	(49.69, 50.03)
Male black	49.66	(49.40, 49.92)
Male Mexican-American	49.59	(49.34, 49.84)
Female white	48.94	(48.73, 49.14)
Female black	49.07	(48.82, 49.33)
Female Mexican-American	48.51	(48.29, 48.74)

interval (Table 4). Histograms of the posterior means are shown in Figure 3.

We found the estimated posterior means and confidence intervals for HEADC in Tables 1 and 4 to be similar, thus yielding similar inferences. Further, both MEMI and likelihood estimates under the assumed model required comupatationally intensive simulation methods. Therefore, the decision of whether to use MEMI or likelihood inference will depend on the particular goals of the study. In a narrow sense, one may define the problem of inference in relation to θ, the unknown parameter of the data model. In a broader sense, however, the problem of inference often goes beyond making statements about just the parameter θ. Data analysts will typically want to apply a variety of exploratory and modeling techniques to a dataset. When the data are incomplete, the analyst's task becomes considerably more difficult. Thus MEMI is one way to allow data analysts to perform general inference and multipurpose data analysis, while acknowledging the uncertainty due to missing data and response error.

5. DISCUSSION

The procedures discussed in this article are appropriate for continuous data with intermittent errors. The concept of MEMI is quite general, however, and may be extended to other types of data such as categorical, longitudinal, or mixed, and also to other error mechanisms. It should be possible to produce MEMI systems for all of those situations in which multiple imputation has been successfully implemented. Thus MEMI has many potential applications and extensions. To conclude, we discuss the assumptions of MEMI and suggest how these may be extended and improved.

MEMI operates under the assumption that the specified data model is approximately correct; departures from the multivariate normal model are treated as errors. Further, deviations of observed data from normality are used to derive posterior probabilities of response error. A natural question to ask is how departures from normality in the ideal-data distribution will affect the performance of the method. Although transforming the data is one way of ensuring normality, it may not always work. A large-scale simulation study conducted by Schafer et al. (1996) produced encouraging results for multiple imputation under plausible nonnormal populations. We plan to build on this study.

We have assumed MAR for the nonresponse mechanism in this implementaion of MEMI. However, in surveys such as NHANES III, it is conceivable that the nonresponse may be directly related to response error. For example, when the NHANES medical staff detect a measurement that seems erroneous, they may decide to not record it, thus producing a missing value. Here the magnitude of the response error is the direct cause of nonresponse, so that the missing-data mechanism is not MAR. Further, when follow-up data are available, it may be possible to relax the MAR assumption, as well as to estimate the nonresponse mechanism. Thus future work should include exploration of other possible assumptions about nonresponse.

We have assumed an a priori model of mutual independence for the response-error indicators $\mathbf{z}_i = (z_{i1}, \ldots, z_{ip})^T$, implying that the errors are expected to occur on a variable-by-variable basis. Such a model may place high probability on error patterns with no or single errors and may not sufficiently account for error patterns with multiple errors. Therefore, it would be useful to generalize the model of independence using log-linear modeling, to account for multiple errors while keeping the number of unknown parameters to be estimated under control.

We also assumed a semincontinuous distribution for the intermittent errors (ϵ_{ij}^*'s) in NHANES III body measurement data, with a normal model for the errors. In reality, the errors may be nonnormal or asymmetric. In some situations, Box–Cox transformations may be sufficient. If auxiliary data from a follow-up or reinterview study are available, then a measurement-error model may be directly estimated (Kuha and Temple 1999). It also may be possible to relax assumptions on the means of the error distributions, if we can add informative priors for the means.

Finally, a useful and interesting byproduct of MEMI inference is the estimated fraction of missing information, $\hat{\lambda}$, which provides a measure of inferential uncertainty about an estimand due to nonresponse and response error. Methods for decomposing $\hat{\lambda}$ into two parts, one part attributable to response error and the other part attributable to missing data, should be a valuable diagnostic tool in survey design. Therefore, it also will be desirable to separate these two sources of uncertainty in the future.

[Received April 2003. Revised August 2003.]

REFERENCES

Barcaroli, G., and Venturi, M. (1993), "An Integrated System for Edit and Imputation of Data: An Application to the Italian Labor Force Survey," in *Proceedings of the 49th Session of the International Statistical Institute*, Florence, Italy.

Biemer, P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A., and Sudman, S. (eds.) (1991), *Measurement Errors in Surveys*, New York: Wiley.

Biemer, P. P., and Stokes, L. (1991), "Approaches to the Modeling of Measurement Error," in *Measurement Errors in Surveys*, eds. P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman, New York: Wiley.

Bollen, K. A. (1989), *Structural Equations With Latent Variables*, New York: Wiley.

Box, G. E. P., and Tiao, G. C. (1992), *Bayesian Inference in Statistical Analysis*, New York: Wiley.

Carlin, B. P., and Louis, T. A. (2000), *Bayes and Empirical Bayes Methods for Data Analysis* (2nd ed.), Boca Raton, FL: Chapman & Hall/CRC.

Ezzati-Rice, T. M., Khare, M., and Schafer, J. L. (1993), "Multiple Imputation of Missing Data in NHANES III," in *Proceedings of the Annual Research Conference*, U.S. Bureau of the Census, pp. 459–487.

Fellegi, I. P., and Holt, D. (1976), "A Systematic Approach to Automatic Edit and Imputation," *Journal of the American Statistical Association*, 71, 17–35.

Fuller, W. A. (1987), *Measurement Error Models*, New York: Wiley.

Ghosh-Dastidar, M. (1998), "Multiple Edit/Multiple Imputation for Multivariate Continuous Data," unpublished doctoral dissertation, Pennsylvania State University.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (eds.) (1996), *Introducing Markov Chain Monte Carlo*, London: Chapman & Hall.

Granquist, L., and Kovar, J. (1999), "Editing of Survey Data: How Much Is Enough?" in *Survey Measurement and Quality*, eds. L. E. Lyberg, P. P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz, and D. Trewin, New York: Wiley.

Groves, R. (1989), *Survey Errors and Survey Costs*, New York: Wiley.

Kovar, J. G., and Whitridge, P. (1990), "Generalized Edit and Imputation System: Overview and Applications," *Revista Brasileira de Estadistica*, 51, 85–100.

Kuha, J., and Temple, J. (1999), "Covariate Measurement Error in Quadratie Regression," economics discussion paper, Nuffield College, Oxford. Available at *http://www.nuff.ox.ac.uk/economics/papers/1999/index1999.htm*.

Lange, K. L., Little, R. J. A., and Taylor, J. M. G. (1989), "Robust Statistical Modeling Using the t-Distribution," *Journal of the American Statistical Association*, 84, 881–896.

Lessler, J. T., and Kalsbeek, W. D. (1992), *Nonsampling Errors in Surveys*, New York: Wiley.

Little, R. J. A. (1988), "Robust Estimation of the Mean and Covariance Matrix From Data With Missing Values," *Applied Statistician*, 37, 23–38.

Little, R. J. A., and Smith, P. J. (1987), "Editing and Imputation for Quantitative Survey Data," *Journal of the American Statistical Association*, 82, 58–68.

Lord, F., and Novick, M. R. (1968), *Statistical Theories of Mental Test Scores*, Reading, MA: Addison-Wesley.

Meng, X. L. (1994), "Multiple-Imputation Inferences With Uncongenial Sources of Input" (with discussion), *Statistical Science*, 9, 538–573.

National Center for Health Statistics (1994), "Plan and Operation of the Third National Health and Nutrition Examination Survey," *Vital and Health Statistics*, Ser. 1.

Olsen, M. K., and Schafer, J. L. (2001), "A Two-Part Random-Effects Model for Semicontinuous Longitudinal Data," *Journal of the American Statistical Association*, 96, 730–745.

Peel, D., and McLachlan, G. J. (2000), "Robust Mixture Modelling Using the t Distribution," *Statistics and Computing*, 10, 339–348.

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

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

——— (1996), "Multiple Imputation After 18+ Years" (with discussion), *Journal of the American Statistical Association*, 91, 473–489.

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

Schafer, J. L., Ezzati-Rice, T. M., Johnson, W., Khare, M., Little, R. J. A., and Rubin, D. B. (1996), "The NHANES III Multiple Imputation Project," in *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 696–701.

Thompson, K. J., and Sigman, R. S. (1999), "Statistical Methods for Developing Ratio Edit Tolerances for Economic Data," *Journal of Official Statistics*, 15, 517–535.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions*, New York: Wiley.

Winkler, W. E., and Draper, L. (1997), "The SPEER Edit System," in *Statistical Data Editing*, Vol. II, eds. J. Kovar and L. Granquist, Geneva, Switzerland U.N. Commission for Europe, pp. 51–55.