Inference With Imputed Conditional Means

Joseph L. SCHAFER and Nathaniel SCHENKER

In this article we present analytic techniques for inference from a dataset in which missing values have been replaced by predictive means derived from an imputation model. The derivations are based on asymptotic expansions of point estimators and their associated variance estimators, and the resulting formulas can be thought of as first-order approximations to standard multiple-imputation procedures with an infinite number of imputations for the missing values. Our method, where applicable, may require substantially less computational effort than creating and managing a multiply imputed database; moreover, the resulting inferences can be more precise than those derived from multiple imputation, because they do not rely on simulation. Our techniques use components of the standard complete-data analysis, along with two summary measures from the fitted imputation model. If the imputation and analysis phases are carried out by the same person or organization, then the method provides a quick assessment of the variability due to missing data. If a data producer is supplying the imputed data set to outside analysts, then the necessary summary measures could be supplied to the analysts, enabling them to apply the method themselves. We emphasize situations with iid samples, univariate missing data, and complete-data point estimators that are smooth functions of means, but also discuss extensions to more complicated situations. We illustrate properties of our methods in several examples, including an application to a large dataset on fatal accidents maintained by the National Highway Traffic Safety Administration.

KEY WORDS: Linearization; Missing data; Multiple imputation; Nonresponse; Taylor series.

1. INTRODUCTION

1.1 Overview

A standard technique for handling missing data in a large dataset is to impute (i.e., fill in) a plausible value for each missing datum, and then analyze the resulting data as if they were complete. Imputation is attractive because it facilitates standard complete-data methods of analysis. In addition, imputations are often created by persons or organizations closely connected to the process of data collection, who may have more information available to model the missing values than those who ultimately perform the analyses. But one drawback of imputation followed by the use of complete-data methods of analysis is that the resulting inferences (e.g., confidence intervals and p values) may be seriously misleading, because uncertainty due to missing data has not been addressed (e.g., Little and Rubin 1987, chap. 3).

In response to this shortcoming, Rubin (1987, 1996) proposed the paradigm of multiple imputation for incorporating missing-data uncertainty. In multiple imputation, the data are completed several times by imputing multiple random draws of the missing values from a predictive distribution. Methods for creating these draws were described by Schafer (1997). A standard complete-data analysis is applied to each completed data set separately, and the results are combined to produce a single set of summary statistics (estimates, standard errors, etc.) that reflects variability across imputations. Rubin (1987, chap. 3) called this type of analysis a "repeated-imputation" analysis. Other methods for incorporating multiple imputations of missing values into analyses are also possible (Fay 1996). A common theme of these approaches is that missing-data uncertainty is reflected through multiple, plausible versions of the missing values.

In this article we develop an analytic method to produce appropriate variance estimates with just a single, nonrandom imputation of predictive means for the missing values. Our method is based on asymptotic expansions of point estimators and their associated variance estimators and produces a first-order approximation to Rubin's repeated-imputation inference with an infinite number of imputations. The distinction between our method and multiple imputation is somewhat like the distinction between linearization and replication/resampling methods for variance estimation in sample surveys (e.g., Wolter 1985), in that we assess variability through mathematical formulas, rather than through iterative computation of estimates from alternative simulated versions of the dataset.

At the outset, we present methods for independent and identically distributed (iid) samples, possibly multivariate, but with missing values on just a single variable. (Extensions to more complicated situations are discussed in Sec. 6.) We assume that the complete-data point estimator is a smooth scalar function of the sample means. For this limited but important class of problems, our method can be much easier to use than multiple imputation, as it does not require the creation or management of a multiply imputed dataset. Instead of drawing multiple imputations, we replace missing data with one set of predictive means computed from a regression-type imputation model. We apply the standard complete-data analysis to the imputed dataset, computing standard errors in the usual manner. The resulting standard errors, which are downwardly biased, are then corrected by two summary measures derived from the imputation model.

Because our method approximates a repeated-imputation inference with an infinite number of imputations, it can be more efficient than multiple imputation with a small number of imputations, especially when the fraction of missing information is high (Rubin 1987, chap. 4). Moreover, our

Joseph L. Schafer is Associate Professor, Department of Statistics, Pennsylvania State University, University Park, PA 16802 (E-mail: *jls@stat.psu.edu*). Nathaniel Schenker is Senior Scientist for Research and Methodology, National Center for Health Statistics, Hyattsville, MD 20782 (E-mail: *nhs1@cdc.gov*). The authors' names are listed in alphabetical order.

Journal of the American Statistical Association
March 2000, Vol. 95, No. 449, Theory and Methods

method typically requires much less computational effort than creating and maintaining a multiply imputed database. For logistical reasons, our method is especially attractive when the imputation and analysis phases are carried out by the same person or organization, because the summary measures needed to correct the standard errors will be readily available. The assumed form of the estimator applies to many of the basic summary measures computed and published by statistical agencies. For these reasons, we had originally envisioned our method being used internally within a data collection agency, rather than by external users performing secondary analyses. It has been pointed out by the associate editor, however, that if data producers supply imputed datasets to outside users, then the necessary summary measures could be supplied as well, allowing analysts to compute their own corrected standard errors in simple analyses. Even when these summary measures are not available—as in our motivating example on fatal highway accidents presented in Section 1.2—a partial correction of standard errors is often possible.

Our assumptions about the inference problem and the missing data are described in Section 2. Comments on various imputation strategies (mean, single random, and multiple imputation) are given in Section 3, followed by the theoretical justification for our approach in Section 4. In Section 5 we present abstract examples to illustrate the approach and to demonstrate its properties, and we continue our discussion of the motivating example from Section 1.2. Extensions and concluding remarks are given in Section 6.

1.2 Motivating Example

The Fatal Accident Reporting System (FARS), maintained by the National Highway Traffic Safety Administration (NHTSA), is an annual registry of fatal accidents occurring on U.S. highways. One important variable in FARS is the blood alcohol content (BAC) of persons actively involved in the accident. Because it is often impractical to collect specimens for BAC determination at the scene of an accident, 50% or more of the BAC values are missing. The fractions of missing information for this variable tend to be substantially lower than 50%, however, because powerful predictors of BAC are typically available; these include vehicle class, age and sex of the driver, time of day, and informal assessments by police as to whether alcohol appeared to have been involved.

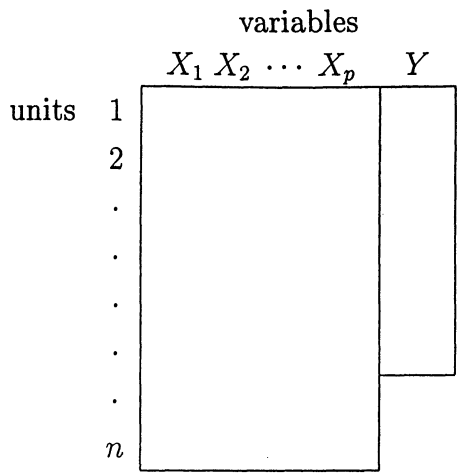

Figure 1. Rectangular Dataset With One Variable Partially Missing.

In the early 1980s, NHTSA developed a procedure for estimating individuals' probabilities of membership in three classes: BAC $= 0, 0 <$ BAC $< .10$, and BAC $\geq .10$. The cutoff of .10 represented the legal limit for drunk driving used by most states at that time. Probabilities were estimated under a three-class discriminant model incorporating predictors found to be significantly related to BAC (Klein 1986). Data files were created in which BAC was replaced by three probabilities corresponding to the three classes. Individuals with known BAC were assigned a probability of 1 for the observed BAC class and 0's in the other two classes, whereas individuals with unknown BAC were assigned nonzero probabilities for all three classes estimated under the discriminant model. Averages of these probabilities have been used extensively in published summaries of the FARS data, but without any quantitative assessment of the uncertainty introduced by missing data. Using the techniques of Section 4, it is possible to reconstruct some simple measures of missing-data uncertainty from the existing FARS data files. We do this for FARS data from 1993 in Section 5.4, where we show that the additional variance due to missing data in analyses of FARS data can be substantial.

2. SETUP AND ASSUMPTIONS

2.1 Pattern of Missing Data

Consider an iid sample with n observational units, in which a single variable Y is sometimes missing, whereas other variables $X_1, \ldots, X_p$ are completely observed. Figure 1 presents a schematic diagram of such a dataset. (Generalizations to multivariate Y and complex samples are discussed in Sec. 6.) Let $\mathbf{X}$ denote the $n \times p$ matrix of observed data for $X_1, \ldots, X_p$, and let $\mathbf{y}$ denote the $n \times 1$ vector of Y values. Then $\mathbf{y}$ can be partitioned into sets of observed and missing components, $\mathbf{y}_{\text{obs}}$ and $\mathbf{y}_{\text{mis}}$, with lengths n_1 and $n_0 = n - n_1$. The rate at which Y is observed is $r_1 = n_1/n$, whereas the missingness rate is $r_0 = 1 - r_1$. We assume that r_0 is bounded away from 1 as $n \to \infty$.

2.2 Estimation With Complete Data

Let Q denote a scalar quantity to be estimated. If the data were complete, then typical analyses would be based on a point estimate, $\hat{Q} = \hat{Q}(\mathbf{X}, \mathbf{y})$, and an associated estimate of variance for $\hat{Q}$, $U = U(\mathbf{X}, \mathbf{y})$. We consider point estimators that are smooth functions of means of the variables. Let

$$\hat{Q} = g(T_{X_1}, \ldots, T_{X_p}, T_{\mathbf{y}}), \tag{1}$$

where $T_{X_j} = n^{-1} \sum_{i=1}^{n} X_{ij}, j = 1, \ldots, p, T_{\mathbf{y}} = n^{-1} \sum_{i=1}^{n} y_i, X_{ij}$ denotes the value of X_j for unit i, y_i denotes the value of Y for unit i, and g is smooth and well behaved. Typically, the estimand Q will be the same function g of the expectations of the means, $Q = g(ET_{X_1}, \ldots, ET_{X_p}, ET_{\mathbf{y}})$, where the expectations are taken over repeated sampling of $\mathbf{X}$ and $\mathbf{y}$; hence $\hat{Q}$ can be thought of as a method-of-moments estimate of Q. The form (1) in-

cludes many estimators typically used in survey practice and elsewhere, such as means and proportions and ratios of means, but does not include medians, variances, or correlations. A referee has suggested a possible extension to a more general class of estimators; see Section 6.1.

We assume that the complete-data variance estimator U has the form

$$U = n^{-1}\left(\frac{\partial g(\mathbf{T})}{\partial \mathbf{T}}\right)^T \mathbf{S}\left(\frac{\partial g(\mathbf{T})}{\partial \mathbf{T}}\right), \tag{2}$$

where $\mathbf{T} = (T_{X_1}, \ldots, T_{X_p}, T_{\mathbf{y}})^T$ and $\mathbf{S} = (n-1)^{-1}(\mathbf{Z}^T\mathbf{Z} - n\mathbf{T}\mathbf{T}^T)$, with $\mathbf{Z} = (\mathbf{X}, \mathbf{y})$. That is, U is the classical variance estimator for $\hat{Q}$ based on the sample covariance matrix and the δ method.

Inferences for estimands of the type that we consider are typically based on a normal reference distribution. In accord with this practice, we assume that

$$U^{-1/2}(Q - \hat{Q}) \xrightarrow{\mathcal{L}} \mathrm{N}(0, 1) \tag{3}$$

as $n \to \infty$. When the units in the dataset constitute an iid sample, (3) follows from standard central limit theorem arguments and Slutsky's theorem.

2.3 Modeling Missing Data

When values of Y are missing, $X_1, \ldots, X_p$ provide useful information for predicting the missing values to the extent that these variables are related to Y. If Y is continuous, for example, we might fit a normal regression model to the cases for which Y is observed and use the fitted model to predict $\mathbf{y}_{\text{mis}}$. This approach implicitly assumes that the conditional distribution of Y given $X_1, \ldots, X_p$ is the same whether Y is missing or observed, which is appropriate if the missingness mechanism is ignorable (Rubin 1976). Most procedures for handling missing data in surveys and elsewhere are based on an assumption of ignorability. The observed data provide no information to support or contradict this assumption; such evidence must come from sources external to the observed data. Nonignorable methods are possible, but every missing-data procedure must be based on some assumption that cannot be verified from $(\mathbf{X}, \mathbf{y}_{\text{obs}})$ alone. In this article we assume that missingness is ignorable and that a probability model for $\mathbf{y}_{\text{mis}}$ given $(\mathbf{X}, \mathbf{y}_{\text{obs}})$ has been correctly specified. Also implicit in our development is the assumption that the model used for imputation is "congenial" to the procedure used for the subsequent data analysis (Meng 1994), meaning roughly that the imputation and analysis phases are based on models that agree. Congeniality is not a necessary condition for the procedure to work well, however, and Section 5.2 explores its performance in an example of uncongeniality.

A typical specification for the missing-data model will include unknown but estimable parameters $\boldsymbol{\theta}$. Let $\hat{\boldsymbol{\theta}}$ denote an efficient estimate of $\boldsymbol{\theta}$ based on $(\mathbf{X}, \mathbf{y}_{\text{obs}})$ under the assumed model. Also, let $\boldsymbol{\Gamma}$ denote an estimate of $V(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$, also based on $(\mathbf{X}, \mathbf{y}_{\text{obs}})$. For example, $\hat{\boldsymbol{\theta}}$ may be a maximum likelihood (ML) estimate, and $\boldsymbol{\Gamma}$ may be the inverse of the observed information matrix evaluated at $\hat{\boldsymbol{\theta}}$. We assume that $\boldsymbol{\Gamma} = O(n^{-1})$ and that

$$\boldsymbol{\Gamma}^{-1/2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \xrightarrow{\mathcal{L}} \mathrm{N}(0, \mathbf{I}), \tag{4}$$

where $\mathbf{I}$ denotes the identity matrix. This implicitly assumes that the missing-data model is sufficiently regular that standard ML asymptotic theory (e.g., Cox and Hinkley 1974, chap. 9) applies, that the fraction of missing information is bounded away from 1, and that the dimension of $\boldsymbol{\theta}$ is fixed. We assume further that the model for missing data imposes an uncorrelated (given $\boldsymbol{\theta}$) error structure on the missing values. More precisely, let "mis" denote the set of indices i such that y_i is missing. Then for all $i, i' \in$ mis, we assume that

$$E(y_i|\mathbf{X}, \mathbf{y}_{\text{obs}}, \boldsymbol{\theta}) = \mu_i(\boldsymbol{\theta}), \qquad V(y_i|\mathbf{X}, \mathbf{y}_{\text{obs}}, \boldsymbol{\theta}) = \sigma_i^2(\boldsymbol{\theta}),$$

and

$$\operatorname{cov}(y_i, y_{i'}|\mathbf{X}, \mathbf{y}_{\text{obs}}, \boldsymbol{\theta}) = 0, \tag{5}$$

where μ_i and σ_i^2 are smooth, well-behaved functions of $\boldsymbol{\theta}$.

Our assumptions about the missing-data model are not in practice overly restrictive. They are satisfied by normal linear regression and analysis of variance models, logistic regression, and log-linear and other generalized linear models—most of the commonly used statistical models that are appropriate for predicting a univariate Y from variables $X_1, \ldots X_p$ in a rectangular dataset. In Section 6 we discuss extensions of our results to allow for intracluster correlation in complex samples.

3. APPROACHES TO IMPUTING FOR MISSING DATA

Once a model for $\mathbf{y}_{\text{mis}}$ given $(\mathbf{X}, \mathbf{y}_{\text{obs}})$ has been specified, several approaches may be taken to impute for $\mathbf{y}_{\text{mis}}$. We comment briefly on three such approaches.

3.1 Conditional Mean Imputation

Let $\boldsymbol{\mu}(\boldsymbol{\theta})$ denote the vector with elements $\mu_i(\boldsymbol{\theta}), i \in$ mis; that is, $\boldsymbol{\mu}(\boldsymbol{\theta}) = E(\mathbf{y}_{\text{mis}}|\mathbf{X}, \mathbf{y}_{\text{obs}}, \boldsymbol{\theta})$, where expectation is with respect to the missing-data model. An approach that seeks to fill in the missing data with one set of "best" values might choose $\boldsymbol{\mu}(\hat{\boldsymbol{\theta}})$. Little and Rubin (1987, sec. 3.4) referred to this technique as imputing conditional means.

Conditional mean imputation can be efficient for point estimation of Q; in fact, we demonstrate in Section 4.3 that $\hat{Q}(\mathbf{X}, \mathbf{y}_{\text{obs}}, \boldsymbol{\mu}(\hat{\boldsymbol{\theta}}))$ is a first-order approximation to the "best" estimate of Q. Inferences can be seriously distorted with conditional mean imputation, however. The analog to (3) with $\mathbf{y}_{\text{mis}}$ replaced by $\boldsymbol{\mu}(\hat{\boldsymbol{\theta}})$ does not hold in general, because the mean-imputed variance estimate is usually biased downward. The average of $U(\mathbf{X}, \mathbf{y}_{\text{obs}}, \boldsymbol{\mu}(\hat{\boldsymbol{\theta}}))$ over repeated samples tends to be less than the sampling variance of $Q - \hat{Q}(\mathbf{X}, \mathbf{y}_{\text{obs}}, \boldsymbol{\mu}(\hat{\boldsymbol{\theta}}))$ (e.g., Little and Rubin 1987, sec. 3.4). Users of conditional mean-imputed data should also be aware that that imputed values may fall outside the set of possible data values; for example, when the variable being imputed is discrete.

3.2 Single Random Imputation

A second strategy is to impute at random from an esti-

mate of the distribution of $\mathbf{y}_{\text{mis}}$. For example, one might impute $\mathbf{y}^*_{\text{mis}}$, a random draw from $\mathcal{L}(\mathbf{y}_{\text{mis}}|\mathbf{X}, \mathbf{y}_{\text{obs}}, \hat{\boldsymbol{\theta}})$, and base inferences on $\hat{Q}(\mathbf{X}, \mathbf{y}_{\text{obs}}, \mathbf{y}^*_{\text{mis}})$ and $U(\mathbf{X}, \mathbf{y}_{\text{obs}}, \mathbf{y}^*_{\text{mis}})$.

For point estimation, single random imputation is less efficient than conditional mean imputation because the random imputation mechanism introduces extra noise; that is, $V(\hat{Q}(\mathbf{X}, \mathbf{y}_{\text{obs}}, \mathbf{y}^*_{\text{mis}})) > V(\hat{Q}(\mathbf{X}, \mathbf{y}_{\text{obs}}, \boldsymbol{\mu}(\hat{\boldsymbol{\theta}}))$, where the variance is with respect to repeated sampling and imputation. The variance estimate $U(\mathbf{X}, \mathbf{y}_{\text{obs}}, \mathbf{y}^*_{\text{mis}})$ tends to be larger than its conditional mean–imputed counterpart $U(\mathbf{X}, \mathbf{y}_{\text{obs}}, \boldsymbol{\mu}(\hat{\boldsymbol{\theta}}))$. Because the variance being estimated is also larger, however, there is still typically a downward bias; on average, $U(\mathbf{X}, \mathbf{y}_{\text{obs}}, \mathbf{y}^*_{\text{mis}})$ still underestimates the actual variance of $Q - \hat{Q}(\mathbf{X}, \mathbf{y}_{\text{obs}}, \mathbf{y}^*_{\text{mis}})$ (e.g., Rubin 1987, prob. 1.12).

3.3 Multiple Imputation

Multiple imputation (Rubin 1987) addresses the shortcomings of conditional mean and single random imputation, while retaining much of the convenience of imputation as a procedure for handling missing data. With multiple imputation, $\mathbf{y}_{\text{mis}}$ is replaced by M random draws from a predictive distribution. For proper inference, the distribution from which the imputed values are drawn must incorporate variability due to the uncertainty about the parameter θ of the missing-data model as well as due to the randomness of $\mathbf{y}_{\text{mis}}$ given θ. Using Bayesian notation, we can write the predictive density of $\mathbf{y}_{\text{mis}}$ as

$$p(\mathbf{y}_{\text{mis}}|\mathbf{X}, \mathbf{y}_{\text{obs}}) = \int p(\mathbf{y}_{\text{mis}}|\mathbf{X}, \mathbf{y}_{\text{obs}}, \theta)\; p(\theta|\mathbf{X}, \mathbf{y}_{\text{obs}})\, d\boldsymbol{\theta}, \quad (6)$$

which makes the two sources of variability explicit. Multiple imputation results in M completed datasets and M complete-data analyses; the results of these M analyses are then combined to produce a single overall inference, the repeated-imputation inference, which includes uncertainty due to missing data. When the fraction of missing information is moderate, reliable inferences can be obtained with only a few imputations, say $M = 5$.

4. CORRECTED ANALYSIS FOR CONDITIONAL MEAN IMPUTATION

We now develop a method of inference for Q from a dataset in which missing values have been replaced by conditional means. The method is essentially a linear approximation to a repeated-imputation analysis with an infinite number of imputations ($M \to \infty$).

4.1 Bayesian Interpretation

The usual frequentist interpretation of (3) regards Q as fixed and $\hat{Q}$ and U as random. A Bayesian interpretation, however, regards $\hat{Q}$ and U as fixed (given complete data) and Q as random. Exploiting the latter interpretation, we regard $\hat{Q}$ and U as the approximate complete-data posterior mean and variance of Q; that is, $\hat{Q} = E(Q|\mathbf{X}, \mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}})$ and $U = V(Q|\mathbf{X}, \mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}})$.

Under sufficient regularity conditions, posterior means and variances behave as in (3) (e.g., Cox and Hinkley 1974, chap. 10), and with large samples, the difference between frequentist and Bayesian inferences will be small. We also exploit the Bayesian interpretation of (4) and regard $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\Gamma}$ as posterior moments of $\boldsymbol{\theta}$ given the observed data; that is, $\hat{\boldsymbol{\theta}} = E(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}_{\text{obs}})$ and $\boldsymbol{\Gamma} = V(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}_{\text{obs}})$.

When the data are complete, our state of knowledge about Q is summarized by $\hat{Q}$ and U. With incomplete data, however, inferences should be based on the posterior moments given only the data actually observed, $E(Q|\mathbf{X}, \mathbf{y}_{\text{obs}})$ and $V(Q|\mathbf{X}, \mathbf{y}_{\text{obs}})$. Note that

$$E(Q|\mathbf{X}, \mathbf{y}_{\text{obs}}) = E(\hat{Q}|\mathbf{X}, \mathbf{y}_{\text{obs}}) \quad (7)$$

and

$$V(Q|\mathbf{X}, \mathbf{y}_{\text{obs}}) = V(\hat{Q}|\mathbf{X}, \mathbf{y}_{\text{obs}}) + E(U|\mathbf{X}, \mathbf{y}_{\text{obs}}), \quad (8)$$

where the moments on the right side of these equations are evaluated over the posterior predictive distribution $\mathcal{L}(\mathbf{y}_{\text{mis}}|\mathbf{X}, \mathbf{y}_{\text{obs}})$, the distribution corresponding to (6) from which multiple imputations would be drawn. Thus, to obtain approximate posterior moments of Q, we need only approximate the mean and variance of $\hat{Q}$ and the mean of U over the predictive distribution of $\mathbf{y}_{\text{mis}}$.

4.2 Approximate Moments of $\hat{Q}$ and U

The following approximations to the posterior moments of $\hat{Q}$ and U, whose derivations are outlined in the Appendix, follow from first-order Taylor series expansions of the functions g and $\mu_i, i \in \text{mis}$, and large-sample results from the theory of sample surveys (e.g., Wolter 1985, chap. 6). Under the assumptions outlined in Section 2,

$$E(\hat{Q}|\mathbf{X}, \mathbf{y}_{\text{obs}}) = \hat{Q}(\mathbf{X}, \mathbf{y}_{\text{obs}}, \boldsymbol{\mu}(\hat{\boldsymbol{\theta}})) + O_p(n^{-1}), \quad (9)$$

$$V(\hat{Q}|\mathbf{X}, \mathbf{y}_{\text{obs}}) = \left(\frac{\partial g(\hat{\mathbf{T}})}{\partial T_{\mathbf{y}}}\right)^2 n^{-2} \sum_{i \in \text{mis}} \sigma_i^2(\hat{\boldsymbol{\theta}}) + \left(\frac{\partial g(\hat{\mathbf{T}})}{\partial T_{\mathbf{y}}}\right)^2 \times D_{\mu}(\hat{\boldsymbol{\theta}})^T \boldsymbol{\Gamma} D_{\mu}(\hat{\boldsymbol{\theta}}) + O_p(n^{-3/2}), \quad (10)$$

and

$$E(U|\mathbf{X}, \mathbf{y}_{\text{obs}}) = U(\mathbf{X}, \mathbf{y}_{\text{obs}}, \boldsymbol{\mu}(\hat{\boldsymbol{\theta}})) + \left(\frac{\partial g(\hat{\mathbf{T}})}{\partial T_{\mathbf{y}}}\right)^2 \times n^{-2} \sum_{i \in \text{mis}} \sigma_i^2(\hat{\boldsymbol{\theta}}) + O_p(n^{-3/2}), \quad (11)$$

where $\hat{\mathbf{T}}$ is shorthand for the complete-data statistic $\mathbf{T}$ calculated with $\boldsymbol{\mu}(\hat{\boldsymbol{\theta}})$ substituted for $\mathbf{y}_{\text{mis}}$, and where $D_{\mu}(\boldsymbol{\theta}) = n^{-1} \sum_{i \in \text{mis}} ([\partial \mu_i(\boldsymbol{\theta})]/\partial \boldsymbol{\theta})$. Note that the moments in (9)–(11) do not necessarily exist for any finite n; each should be interpreted as the moment of a limiting distribution, not as the limit of a sequence of moments. For example, (9) means that $\hat{Q}$ can be written as the sum of a random variable with mean $\hat{Q}(\mathbf{X}, \mathbf{y}_{\text{obs}}, \boldsymbol{\mu}(\hat{\boldsymbol{\theta}}))$ and a random variable that is $O_p(n^{-1})$.

4.3 Point Estimation

It follows from (7) and (9) that the complete-data point estimate with conditional means imputed for missing values of Y is a first-order approximation to the posterior mean of Q,

$$E(Q|\mathbf{X}, \mathbf{y}_{\text{obs}}) \approx \hat{Q}(\mathbf{X}, \mathbf{y}_{\text{obs}}, \boldsymbol{\mu}(\hat{\boldsymbol{\theta}})). \tag{12}$$

Thus in large samples it is desirable to use $\hat{Q}(\mathbf{X}, \mathbf{y}_{\text{obs}}, \boldsymbol{\mu}(\hat{\boldsymbol{\theta}}))$, as it is an efficient estimate of Q. Note, however, that this result assumes that the complete-data point estimate is a smooth function of linear statistics. Thus the result does not hold for an arbitrary estimator $\hat{Q}$, such as a sample median. In Section 6.1 we describe briefly how a multivariate extension of our results can be applied to more general types of estimators, including sample variances and regression coefficients.

4.4 Variance Estimation

It follows from (8), (10), and (11) that a first-order approximation to the posterior variance of Q is

$$V(Q|\mathbf{X}, \mathbf{y}_{\text{obs}}) \approx U(\mathbf{X}, \mathbf{y}_{\text{obs}}, \boldsymbol{\mu}(\hat{\boldsymbol{\theta}})) + C_1 + C_2, \tag{13}$$

where

$$C_1 = 2\left(\frac{\partial g(\hat{\mathbf{T}})}{\partial T_{\mathbf{y}}}\right)^2 n^{-2} \sum_{i \in \text{mis}} \sigma_i^2(\hat{\boldsymbol{\theta}}) \tag{14}$$

and

$$C_2 = \left(\frac{\partial g(\hat{\mathbf{T}})}{\partial T_{\mathbf{y}}}\right)^2 D_{\mu}(\hat{\boldsymbol{\theta}})^T \boldsymbol{\Gamma} D_{\mu}(\hat{\boldsymbol{\theta}}). \tag{15}$$

In (13), $U(\mathbf{X}, \mathbf{y}_{\text{obs}}, \boldsymbol{\mu}(\hat{\boldsymbol{\theta}}))$ is the "naive" estimate, discussed in Section 3.1, that treats the conditional mean–imputed dataset as complete data. The first correction term, C_1, is a component of variance that accounts for uncertainty in $\mathbf{y}_{\text{mis}}$ given the imputed means. The second correction term, C_2, is an additional component of variance that accounts for uncertainty in the imputed means; that is, uncertainty due to the estimation of the parameters in the missing-data model.

4.5 Computing C_1 and C_2

Three basic components are needed to compute C_1 and C_2. The first component, $\partial g(\hat{\mathbf{T}})/\partial T_{\mathbf{y}}$, which is needed for both C_1 and C_2, should be easy for the data analyst to obtain, because it is also a component of the complete-data variance estimator (2). The second component, $\sum_{i \in \text{mis}} \sigma_i^2(\hat{\boldsymbol{\theta}})$, needed for C_1, should also be easy to obtain, because it involves only point estimates of parameters from the imputation model. For example, if missing values of Y are modeled by ordinary linear regression, then $\sum_{i \in \text{mis}} \sigma_i^2(\hat{\boldsymbol{\theta}}) = n_0 \hat{\sigma}^2$, where $\hat{\sigma}^2$ is the estimated residual variance. If Y is binary and missing values are modeled as Bernoulli with means $\pi_i, i \in \text{mis}$ (e.g., by logistic regression), then $\sum_{i \in \text{mis}} \sigma_i^2(\hat{\boldsymbol{\theta}}) = \sum_{i \in \text{mis}} \hat{\pi}_i(1 - \hat{\pi}_i)$. In Bernoulli and Poisson models, the variances $\sigma_i^2(\boldsymbol{\theta})$ can be expressed as functions of the means $\mu_i(\boldsymbol{\theta})$, and $\sum_{i \in \text{mis}} \sigma_i^2(\hat{\boldsymbol{\theta}})$ thus can be computed from the imputed dataset alone. In other situations (e.g., negative binomial or gamma models with the same shape parameter), the variances may depend only on the means and a single additional parameter.

The third component needed for $C_2, D_{\mu}(\hat{\boldsymbol{\theta}})^T \boldsymbol{\Gamma} D_{\mu}(\hat{\boldsymbol{\theta}})$, must be provided to the data analyst by the imputer. This component involves $\boldsymbol{\Gamma}$, the estimated covariance matrix of the parameters in the imputation model. Whereas $\boldsymbol{\Gamma}$ is obtainable directly from the estimation process for many models (e.g., using the observed information matrix), it can be difficult to obtain for more complex models. In this latter situation, one option is to use a method such as the jackknife or bootstrap to approximate $\boldsymbol{\Gamma}$. This approach was taken by Belin et al. (1993) and Dorinski and Griffin (1997) in the context of a hierarchical logistic regression model. A second option is to replace C_2 by an appropriate upper bound, if one is available for the specific application. A third option is to omit C_2. When the fraction of missing information is moderate, the proportion of variance in (13) contributed by C_2 can be quite small, as shown by the examples in Section 5.

5. EXAMPLES

5.1 Estimating a Binomial Proportion

Let $\mathbf{y}$ denote an iid sample of size n of binary (0-1) variables. Suppose that there are no other variables ($\mathbf{X}$), and let the estimated be $Q = p$, the population proportion of Y values equal to 1. With complete data (and n not small), standard inferences for p may be based on the point estimate $\hat{Q} = \bar{y} \equiv 1/n \sum_i y_i$ and the variance estimate $U = \{1/[n(n-1)]\} \sum_i (y_i - \bar{y})^2$.

Suppose that Y values are missing completely at random, so that $\mathbf{y}_{\text{obs}}$ is just a simple random sample from $\mathbf{y}$. Imputation is unnecessary in this situation, because correct inferences follow by simply ignoring the missing values. Because the correct answer is known, however, this example provides a simple check of the validity of our approach, and it provides insight into the relative contributions of C_1 and C_2 in variance estimation.

If $\mathbf{y}_{\text{mis}}$ is modeled as a vector of iid Bernoulli(θ) random variables, then in the notation of Section 2.3, $\mu_i(\theta) = \theta$ and $\sigma_i^2(\theta) = \theta(1 - \theta)$. ML estimation based on $\mathbf{y}_{\text{obs}}$ yields $\hat{\theta} = 1/n_1 \sum_{i \in \text{obs}} y_i$ (where "obs" denotes the set of indices i such that y_i is observed) and $\Gamma = \hat{\theta}(1 - \hat{\theta})/n_1$. Substitution into (12)–(15) and algebraic manipulation yield

$$\hat{Q}(\mathbf{y}_{\text{obs}}, \boldsymbol{\mu}(\hat{\theta})) = \bar{y}_1 \equiv \frac{1}{n_1} \sum_{i \in \text{obs}} y_i,$$

$$U(\mathbf{y}_{\text{obs}}, \boldsymbol{\mu}(\hat{\theta})) \approx r_1^2 \frac{1}{n_1(n_1 - 1)} \sum_{i \in \text{obs}} (y_i - \bar{y}_1)^2, \tag{16}$$

$$C_1 \approx 2 r_1 r_0 \frac{1}{n_1(n_1 - 1)} \sum_{i \in \text{obs}} (y_i - \bar{y}_1)^2, \tag{17}$$

and

$$C_2 \approx r_0^2 \frac{1}{n_1(n_1-1)} \sum_{i \in \text{obs}} (y_i - \bar{y}_1)^2. \tag{18}$$

[the approximations in (16)–(18) are due to conventions regarding the use of degrees of freedom versus sample size in the denominator of a sample variance.] Thus the results in Sections 4.3 and 4.4 suggest $\bar{y}_1$ as the point estimate of p and

$$U(\mathbf{y}_{\text{obs}}, \boldsymbol{\mu}(\hat{\theta})) + C_1 + C_2 \approx \frac{1}{n_1(n_1-1)} \sum_{i \in \text{obs}} (y_i - \bar{y}_1)^2 \tag{19}$$

as the associated variance estimate. These same estimates would be obtained by applying standard complete-data methods to $\mathbf{y}_{\text{obs}}$, and thus our method yields the correct answer.

Equations (16)–(19) show that the proportionate contributions of $U(\mathbf{y}_{\text{obs}}, \boldsymbol{\mu}(\hat{\theta}))$, C_1, and C_2 to the correct variance estimate are approximately $r_1^2, 2r_1r_0$, and r_0^2. This suggests that even if the missingness rate is moderate, the naive method of imputing conditional means for $\mathbf{y}_{\text{mis}}$ and using a complete-data estimate of variance can be very misleading. For example, if $r_0 = 20\%$, then $U(\mathbf{y}_{\text{obs}}, \boldsymbol{\mu}(\hat{\theta}))$ is approximately 36% smaller than the correct variance estimate. Another important implication, mentioned at the end of Section 4.5, is that if the missingness rate is moderate, then nearly valid inferences can be drawn without accounting for the variability due to estimating θ. For example, if $r_0 = 20\%$, then omitting C_2 shrinks the estimate by only 4%.

5.2 Inference When the Imputation Model is Uncongenial to the Analysis Procedure

Continuing the example of Section 5.1, consider now an additional variable X_1 that is independent of Y and that divides the sample into halves (say, $X_1 = a$ and $X_1 = b$) with identical response rates r_1 in each half. Suppose that the imputer, knowing that X_1 and Y are independent, follows the procedure of Section 5.1 and imputes the same value $\hat{\theta}$ for every element of y_{mis}. Finally, suppose that the analyst, unaware of this independence, draws inferences about the proportion of Y values equal to 1 in the $X_1 = a$ subpopulation by applying our method to the corresponding half sample. The imputer's model is then uncongenial to the analysis (Meng 1994), because the imputer has used knowledge that X_1 and Y are independent, whereas the analyst has not.

If $m = n/2, m_1 = n_1/2, m_0 = n_0/2$, and $\bar{y}_{1,a}$ is the sample mean of the observed Y values within the $X_1 = a$ half sample, then the analyst computes

$$\hat{Q}(\mathbf{X}, \mathbf{y}_{\text{obs}}, \boldsymbol{\mu}(\hat{\theta})) = r_1\bar{y}_{1,a} + r_0\hat{\theta},$$

$$U(\mathbf{X}, \mathbf{y}_{\text{obs}}, \boldsymbol{\mu}(\hat{\theta})) = \frac{1}{m(m-1)} \left[\sum_{X=a, i \in \text{obs}} \mathbf{y}_i^2 + m_0\hat{\theta}^2 - m(r_1\bar{y}_{1,a} + r_0\hat{\theta})^2 \right],$$

$$C_1 = \frac{2}{m^2} m_0\hat{\theta}(1-\hat{\theta}),$$

and

$$C_2 = r_0^2 \frac{\hat{\theta}(1-\hat{\theta})}{n_1}.$$

Let us compare the analysis based on these quantities with that based on the obvious incomplete-data estimator $\bar{y}_{1,a}$. Conditional on the sample sizes, $E(\bar{y}_{1,a}) = \theta$ and $E[\hat{Q}(\mathbf{X}, \mathbf{y}_{\text{obs}}, \boldsymbol{\mu}(\hat{\theta}))] = \theta$, so both estimators are unbiased. Moreover, conditional on the sample sizes,

$$\frac{V[\hat{Q}(\mathbf{X}, \mathbf{y}_{\text{obs}}, \boldsymbol{\mu}(\hat{\theta}))]}{V(\bar{y}_{1,a})} = \frac{1+r_1^2}{2} \tag{20}$$

and

$$\frac{E[U(\mathbf{X}, \mathbf{y}_{\text{obs}}, \boldsymbol{\mu}(\hat{\theta})) + C_1 + C_2]}{V(\bar{y}_{1,a})} \approx \frac{1 + r_1 + r_1r_0}{2}, \tag{21}$$

where the approximation in (21) ignores $O(n^{-1})$ terms.

Expressions (20) and (21) demonstrate three important properties of our procedure in this example of uncongeniality. First, the ratio (20) is less than 1, indicating that the complete-data point estimator applied to the conditional mean-imputed dataset is more efficient than the obvious incomplete-data estimator $\bar{y}_{1,a}$. This can be explained by the extra information carried by the imputations due to the superior knowledge of the imputer. Second, the ratio (21) is also less than 1, indicating that interval estimates based on our procedure tend to be narrower than those produced by the obvious incomplete-data procedure in which $\bar{y}_{1,a}$ is used together with an unbiased estimator of its variance. Finally, (21) is greater than (20), indicating that our variance estimator $U(\mathbf{X}, \mathbf{y}_{\text{obs}}, \boldsymbol{\mu}(\hat{\theta})) + C_1 + C_2$ overestimates the true variance of $\hat{Q}(\mathbf{X}, \mathbf{y}_{\text{obs}}, \boldsymbol{\mu}(\hat{\theta}))$. Thus interval estimates based on our procedure, although tending to be narrower than those from the obvious incomplete-data procedure, also tend to be conservative.

The associate editor has suggested that in general, such conservatism might help to correct for anticonservatism in situations for which C_2 cannot be computed, as described in Section 4.5. If C_2 is omitted in our current example, then (21) becomes

$$E[U(\mathbf{X}, \mathbf{y}_{\text{obs}}, \boldsymbol{\mu}(\hat{(\theta)})) + C_1]/V(\bar{y}_{1,a}) \approx r_1^2 + 2r_1r_0,$$

which is still greater than (20) as long as $r_0 < 2/3$. This is consistent with the statement at the end of Section 4.5 regarding the small contribution of C_2 when the fraction of missing information is moderate.

The three properties just demonstrated for our procedure have also been attributed to the standard repeated-imputation procedure for multiply imputed data under uncongeniality when the imputer's and analyst's models are both valid, but the imputer assumes more than the analyst (Meng 1994; Rubin 1996). In fact, (20) and (21) agree with results of Meng (1994, sec. 3.1) for multiple imputation in a similar example involving a subpopulation mean. This agreement is not surprising given that our procedure approximates the repeated-imputation inference with an infinite number of imputations. Fay (1991, 1992, 1993, 1996)

has criticized multiple imputation for its conservatism in situations with uncongeniality. Meng (1994) and Rubin (1996), on the other hand, have argued that a procedure that is conservative but more efficient is preferable to a procedure that is not conservative but less efficient.

5.3 Estimating a Ratio of Means: The Fieller–Creasey Problem

Let $(x_i, y_i), i = 1, \ldots, n$ be independent observations of variables (X, Y) from a bivariate normal distribution with means μ_X and μ_Y, variances σ_X^2 and σ_Y^2, and correlation ρ, and suppose that the estimand is the ratio of means, $Q = \mu_X/\mu_Y$. With complete data, inferences for Q may be based on the point estimate $\hat{Q} = \bar{x}/\bar{y}$ and the associated variance estimate $U = 1/n\bar{y}^2(s_X^2 - 2\hat{Q}s_{XY} + \hat{Q}^2 s_y^2)$ (e.g., Cochran 1977, sec. 6.4), where $\bar{x}$ and $\bar{y}$ are the sample means, s_X^2 and s_Y^2 are the sample variances, and s_{XY} is the sample covariance.

Suppose that Y is missing for some cases. The obvious model for predicting Y from X is $y_i = \beta_0 + \beta_1 x_i + \sigma e_i, i = 1, \ldots, n$, where $e_1, \ldots, e_n$ are iid standard normal random variables. If the model is fitted by least squares applied to the complete cases $(x_i, y_i, i \in \text{obs})$, then an approximate posterior distribution for $\boldsymbol{\theta} = (\beta_0, \beta_1, \log(\sigma^2))$ is normal with mean $\hat{\boldsymbol{\theta}}$ and variance $\mathbf{I}_{\text{cc}}^{-1}$, where $\mathbf{I}_{\text{cc}}$ is the observed information matrix based on the complete cases. Conditional mean imputation replaces y_i by $\hat{\beta}_0 + \hat{\beta}_1 x_i, i \in \text{mis}$. By (14) and (15), the correction terms are $C_1 = 2(\tilde{Q}/\tilde{y})^2(n_0\hat{\sigma}^2/n^2)$ and $C_2 = (\tilde{Q}/\tilde{y})^2 D_{\boldsymbol{\mu}}(\hat{\boldsymbol{\theta}})^T \boldsymbol{\Gamma} D_{\boldsymbol{\mu}}(\hat{\boldsymbol{\theta}})$, where $\tilde{Q}$ and $\tilde{y}$ are the complete-data quantities calculated with conditional means imputed for $\mathbf{y}_{\text{mis}}$, $D_{\boldsymbol{\mu}}(\hat{\boldsymbol{\theta}}) = n^{-1}(n_0, \Sigma_{i\in\text{mis}} x_i)^T$, $\boldsymbol{\Gamma} = \hat{\sigma}^2(\mathbf{A}^T\mathbf{A})^{-1}$, and $\mathbf{A}$ is the $n_1 \times 2$ matrix with 1's in the first column and $(x_i, i \in \text{obs})^T$ as the second column.

We conducted a simulation in which we generated random samples $(x_i, y_i), i = 1, \ldots 250$ from a bivariate normal distribution with $\mu_X = \mu_Y = 5, \sigma_X^2 = \sigma_Y^2 = 1$, and $\rho = .8$. We imposed on Y random missingness depending on X through a probit mechanism in which the probability of missingness increased with X. Specifically, we generated observations $w_1, \ldots, w_{250}$ of a third variable W, distributed as normal with mean 0, variance 1, correlation of .8 with X, but zero partial correlation with Y given X. We then took y_i to be missing if w_i exceeded $\Phi^{-1}(1-\alpha)$, where Φ is the standard normal cumulative distribution function and α is the missingness probability. Thus, missingness on Y tended to occur for large values of X (and thus Y), although the missing data were missing at random (Rubin 1976). Missingness probabilities were set at $\alpha = .05, .10, .25$, and .50.

We carried out 10,000 Monte Carlo trials, calculating estimates and nominal 95% confidence intervals for Q by the following methods: (a) the standard complete-data analysis applied to the dataset with single random imputations for the missing values (see Sec. 3.2); (b) the repeated-imputation analysis applied to the multiply imputed data set (e.g., Rubin and Schenker 1986) with $M = 5, M = 10$, and $M = 50$ imputations; (c) the complete-data analysis applied to the conditional mean–imputed dataset (see Sec. 3.1); (d) the analysis developed for conditional mean imputation in Section 4, but with only the first correction term, C_1, added into the variance; and (e) the analysis developed for conditional mean imputation with both correction terms C_1 and C_2. The repeated-imputation intervals were based on a Student t distribution, using the rules given by Rubin and Schenker (1986); all other intervals used a normal approximation. At each trial, we also performed the standard complete-data analysis with no data missing—which, of course, is possible only in an artificial situation such as this where the "missing" values are seen.

Results from the simulation are summarized in Table 1. For each method, we report the simulated root mean squared error of the point estimate, the average width of the nominal 95% confidence interval, and the interval's rate of coverage. For point estimation, conditional mean imputation is more efficient (i.e., has lower mean squared error) than any random imputation method, single or multiple, because it does not introduce random noise into the data. But the efficiency of multiple imputation approaches that of conditional mean imputation as the number of imputations M grows. Except when the missingness rate is very high (50%), the most dramatic improvement occurs as we move from single random imputation to multiple imputation with $M = 5$.

For interval estimation, note that it is desirable to have an interval that is as narrow as possible with coverage probability close to the nominal rate of 95%. All methods perform reasonably well at a missingness rate of 5%. However, the coverage of intervals from single random imputation and uncorrected conditional mean imputation deteriorates rapidly as the missingness rate increases to 10% and beyond, in a manner consistent with the discussions of Sections 3.1 and 3.2. Multiple imputation shows good coverage at all rates of missingness, even with only $M = 5$ imputations. Corrections for simulation error due to the smallness of M are an inherent part of the multiple-imputation interval, so the intervals tend to become narrower as M increases. Increasing M beyond 10 has minimal effect for 5% or 10% missingness, so little is to be gained by taking additional imputations when missingness rates are low. At higher rates, however, additional imputations shrink the intervals more substantially. This agrees with the observation of Rubin (1987, p. 114), who noted that highly efficient inferences can be achieved with only a few imputations when fractions of missing information are moderate.

Like multiple imputation, our new method of conditional mean imputation with correction terms C_1 and C_2 has proper coverage at all rates of missingness. Our method produces narrower intervals than multiple imputation, however, because it does not rely on simulation. Notice that adding only the first correction term C_1 does almost as well as the full correction (C_1 and C_2) for missingness rates up to 25%, but breaks down by 50%. Once again, omitting C_2 has little effect unless the missingness rate is high, which is consistent with the discussion at the end of Section 4.5. Table 2 shows the average proportion of the total estimated variance, $U(\mathbf{X}, \mathbf{y}_{\text{obs}}, \boldsymbol{\mu}(\hat{\boldsymbol{\theta}})) + C_1 + C_2$, due to each of the three terms $\hat{U} = U(\mathbf{X}, \mathbf{y}_{\text{obs}}, \boldsymbol{\mu}(\hat{\boldsymbol{\theta}})), C_1$, and C_2. Unlike the

Table 1. Simulated Root Mean Squared Error (RMSE) of the Point Estimate, Average Width of the Nominal 95% Interval, and Coverage Rate of the Interval for the Fieller–Creasy Example

Missingness probability and method	*RMSE × 1,000*	*Width × 1,000*	*Coverage (%)*
No missing data	8.07	31.3	94.8
5% missing			
Single random imputation	8.47	31.3	93.5
Multiple imputation (M = 5)	8.36	32.4	95.0
Multiple imputation (M = 10)	8.34	32.3	94.9
Multiple imputation (M = 50)	8.32	32.2	94.8
Conditional mean imputation (uncorrected)	8.32	30.6	93.7
Conditional mean imputation $+C_1$	8.32	32.0	94.7
Conditional mean imputation $+C_1 + C_2$	8.32	32.2	94.8
10% missing			
Single random imputation	8.94	31.3	91.8
Multiple imputation (M = 5)	8.74	33.9	94.7
Multiple imputation (M = 10)	8.69	33.6	94.7
Multiple imputation (M = 50)	8.64	33.4	94.7
Conditional mean imputation (uncorrected)	8.64	29.9	91.4
Conditional mean imputation $+C_1$	8.64	32.7	94.2
Conditional mean imputation $+C_1 + C_2$	8.64	33.3	94.7
25% missing			
Single random imputation	10.64	31.3	86.2
Multiple imputation (M = 5)	10.26	41.5	94.8
Multiple imputation (M = 10)	10.12	39.7	94.5
Multiple imputation (M = 50)	9.98	38.7	94.4
Conditional mean imputation (uncorrected)	9.94	27.6	83.3
Conditional mean imputation $+C_1$	9.94	34.7	91.6
Conditional mean imputation $+C_1 + C_2$	9.94	38.4	94.6
50% missing			
Single random imputation	15.16	31.3	69.4
Multiple imputation (M = 5)	15.12	68.5	94.3
Multiple imputation (M = 10)	14.65	61.3	95.0
Multiple imputation (M = 50)	14.31	57.0	95.1
Conditional mean imputation (uncorrected)	14.20	23.3	58.4
Conditional mean imputation $+C_1$	14.20	37.7	81.6
Conditional mean imputation $+C_1 + C_2$	14.20	55.9	94.9

example of Section 5.1, the fraction due to C_2 is higher than the square of the missingness probability. The missingness mechanism in this example tends to impose missing values at points with high leverage; consequently, the actual fraction of missing information is somewhat higher than the fraction of missing Y values. In general, performance is related more directly to the fraction of missing information than to the fraction of missing data.

5.4 Missing Data on Blood Alcohol Content in the Fatal Accident Reporting System (Continued)

Returning to the example of Section 1.2, consider the problem of estimating the proportion of drivers falling into BAC classes $j = 1, 2, 3$ in some domain $\mathcal{D}$ of interest; for example, passenger-car drivers age 21–29 involved in a fatal accident in 1993. Let $\hat{\pi}_{ij}$ be the probability (possibly 0 or 1) assigned to individual i of belonging to class j. Then the estimated proportion of the drivers in $\mathcal{D}$ belonging to class j is $\tilde{Q}_j = n_{\mathcal{D}}^{-1} \sum_{i \in \mathcal{D}} \hat{\pi}_{ij}$, where $n_{\mathcal{D}}$ is the number of individuals observed in $\mathcal{D}$. The estimate $\tilde{Q}_j$ can be viewed as a conditional mean-imputed version of the ordinary proportion $\hat{Q}_j = n_{\mathcal{D}}^{-1} \sum_{i \in \mathcal{D}} y_{ij}$, where $y_{ij} = 1$ if individual i falls into BAC class j and $y_{ij} = 0$ otherwise.

Table 2. Percentage of Total Estimated Variance $\hat{U} + C_1 + C_2$ due to Each Term for the Fieller–Creasy Example

	Rate of missingness			
Term	*5%*	*10%*	*25%*	*50%*
$\hat{U}$	90.4	80.4	51.7	17.7
C_1	8.5	15.9	29.8	28.4
C_2	1.1	3.7	18.5	54.0

The appropriate variance estimate to attach to $\tilde{Q}_j$ depends on whether we regard the complete data $\mathbf{y} = \{y_{ij} : i \in \mathcal{D}\}$ as a complete enumeration of the existing population or as a realized sample from a hypothetical superpopulation. With the latter view, the complete-data proportion $\hat{Q}_j$ is approximately normally distributed about the unknown superpopulation proportion, with estimated variance $U_j = 1/[n_{\mathcal{D}}(n_{\mathcal{D}} - 1)] \sum_{i \in \mathcal{D}} (y_{ij} - \hat{Q}_j)^2$. The naive variance estimate calculated from the mean-imputed data, $\hat{U}_j = 1/[n_{\mathcal{D}}(n_{\mathcal{D}} - 1)] \sum_{i \in \mathcal{D}} (\hat{\pi}_{ij} - \tilde{Q}_j)^2$, could substantially understate the uncertainty associated with $\tilde{Q}_j$. This understatement can be partially corrected by adding to $\hat{U}_j$ the term $C_1 = 2n_{\mathcal{D}}^{-2} \sum_{i \in \mathcal{D}} \hat{\pi}_{ij}(1 - \hat{\pi}_{ij})$. A full correction would also require the additional term C_2, which cannot be obtained from the FARS data files alone, but requires quantities derived from the discriminant models that produced the estimates $\hat{\pi}_{ij}$.

Some results of this partial correction applied to 1993 FARS data are shown in Table 3 for drivers of motorcycles and drivers of passenger cars. Shown are the esti-

Table 3. Estimated Percentage of 1993 FARS Drivers by Age in Three BAC Classes, With Standard Error (SE) and Percent Contribution of C_1 to the Estimated Variance

		BAC = 0			0 < BAC < .10			BAC ≥ .10		
Age	$n_{\mathcal{D}}$	Estimate	SE	C_1 (%)	Estimate	SE	C_1 (%)	Estimate	SE	C_1 (%)
					Drivers of motorcycles					
12–20	356	77.4	2.2	16.1	10.8	1.6	21.6	11.8	1.7	15.8
21–29	897	54.5	1.7	14.6	12.2	1.1	24.8	33.2	1.6	16.9
30–39	687	44.2	1.9	14.0	11.6	1.2	24.0	44.2	1.9	14.2
40–49	337	52.0	2.7	13.1	10.2	1.7	21.4	37.8	2.6	12.4
50–59	113	63.5	4.5	14.7	11.0	3.0	19.6	25.4	4.1	15.3
60+	64	84.9	4.5	25.9	4.7	2.7	32.5	10.3	3.8	27.8
					Drivers of passenger cars					
12–20	5083	77.1	.6	9.1	7.9	.4	23.1	15.0	.5	13.1
21–29	7500	59.6	.6	10.5	9.1	.3	31.0	31.3	.5	12.1
30–39	5581	63.4	.6	9.2	7.0	.3	31.5	29.6	.6	10.2
40–49	3540	73.9	.7	10.7	5.5	.4	32.0	20.6	.7	10.6
50–59	2257	81.7	.8	12.3	4.1	.4	32.1	14.2	.7	11.3
60+	5525	92.1	.4	17.0	2.7	.2	27.2	5.2	.3	17.0

mated percentages falling into each class, standard errors $\text{SE} = \sqrt{\hat{U} + C_1}$, and the contribution of C_1 to the estimated variance $\hat{U} + C_1$. These standard errors should be regarded as lower bounds, because they omit the terms C_2 due to parameter uncertainty in the discriminant model. Notice that C_1 accounts for up to 33% of $\hat{U} + C_1$. Analogous to the example in Section 5.1, suppose that $C_1/(\hat{U} + C_1)$ approximates $2r_1r_0/(r_1^2 + 2r_1r_0)$, where r_0 is the fraction of missing information and $r_1 = 1 - r_0$. This would lead us to suspect that the fractions of missing information in this example are no more than about $r_0 = .2$, and the omitted terms C_2 thus might account for no more than about $.2^2 = 4\%$ of the total variance.

6. DISCUSSION

6.1 Possible Extensions

Several extensions of our results are possible. First, suppose that several variables $Y_1, \ldots, Y_q$ are subject to missing values and are observed or missing together. (Y_j missing implies all $Y_{j'}, j' \neq j$ missing as well.) Then our results apply with only small differences in notation; details have been provided in a technical report (Schafer and Schenker 1997). A referee has pointed out that this multivariate extension could be applied to quantities such as variances by treating different functions of Y_j (e.g., Y_j and Y_j^2) as separate variables. Note that in this situation, the imputation model should properly reflect deterministic relationships among variables such as Y_j and Y_j^2.

The referee has also suggested that the fully observed variables (X) need not enter the complete-data point estimator (1) through their means, because our development treats X as fixed (see the Appendix). If more general functions of X were allowed, then the complete-data variance estimator (2) as well as C_1 and C_2 [(14) and (15)] would need modification. A related generalization of the complete-data point estimator is to allow variables to enter the estimator as weighted means or totals, where the weights may depend on the observed data $(\mathbf{X}, \mathbf{y}_{\text{obs}})$. This extension accommodates more complicated estimands, including regression coefficients, as well as sampling weights from a complex sample design. With a complex sample design, the imputation model may generalize (5) to include covariances that account for intracluster correlations. Moreover, the complete-data variance estimator would be generalized to account for the design, C_1 and C_2 would incorporate weights, and C_1 would incorporate covariances of the missing values if these were included in the imputation model. (See Schafer and Schenker 1997 for details and for an application of this generalization to undercount estimation in the 1990 census, and Dorinski and Griffin 1997 for an application to undercount estimation in the 1995 census test.)

Finally, our results are readily extendable to multidimensional estimands. For this extension, the function defining the estimands becomes a vector, and the results involve straightforward matrix generalizations of our expressions.

6.2 Other Remarks

The Bayesian development used in Section 4 and also in Rubin's (1987) justification for multiple imputation is not the only paradigm for obtaining corrected inferences from imputed datasets. Some new methods with design-based rather than Bayesian origins have been presented by Fay (1996), Rao (1996), and others. Similar results can often be obtained using different paradigms. For example, Särndal (1990) derived "model-assisted" frequentist methods for correcting variance estimates when estimating the population total with a singly imputed dataset. He showed that for a simple random sample, computing the standard variance estimate from a mean-imputed dataset yields an estimate that is only r_1^2 as large as it should be, which is the same result that we derived in Section 5.1 for estimating proportions.

Our derivation is based on the assumption that the same variables (i.e., $X_1, \ldots X_p$ and Y) are used in both the imputation process and in the calculation of the complete-data estimates $\hat{Q}$ and U (see, e.g., Sec. 2). This implicitly reflects the principle that information to be used in the analysis of an imputed dataset should not be omitted from the imputa-

tion model; Rubin (1987, chap. 4) discussed this point. We have also assumed that the complete-data estimates, $\hat{Q}$ and U, may be regarded as the posterior mean and variance of Q under a Bayesian model that is congenial to the imputation model. Although this assumption holds approximately in many applied situations, it may not hold in other situations, such as when certain design-based complete-data estimators are used for complex surveys. Such violations often produce the type of uncongeniality discussed in Section 5.2; the general results of Meng (1994) suggest that the phenomena seen in that example (i.e., efficiency relative to the standard incomplete-data procedure along with conservatism) will occur.

APPENDIX: APPROXIMATING THE MOMENTS OF $\hat{Q}$ AND U

We now sketch proofs of results (9)–(11) of Section 4. Our proofs use standard arguments in Taylor linearization (e.g., Wolter 1985, chap. 6). Care must be taken to ensure that Taylor expansions are taken with respect to quantities whose dimensions remain fixed in the asymptotic sequence. Because moments are calculated with respect to the posterior distribution given $(\mathbf{X}, \mathbf{y}_{\text{obs}})$, we are conditioning on $(\mathbf{X}, \mathbf{y}_{\text{obs}})$ throughout, but for simplicity this is suppressed in the notation. Functions of $\mathbf{X}$ and $\mathbf{y}_{\text{obs}}$ (e.g., $\hat{\boldsymbol{\theta}}$) are considered fixed, whereas functions of $\mathbf{y}_{\text{mis}}$ or $\boldsymbol{\theta}$ are considered random.

Note first that the only random argument of $\hat{Q} = g(\mathbf{T})$ is $T_{\mathbf{y}}$. We can write

$$\begin{aligned} T_{\mathbf{y}} - \hat{T}_{\mathbf{y}} &= n^{-1} \sum_{i \in \text{mis}} (y_i - \mu_i(\hat{\boldsymbol{\theta}})) \\ &= n^{-1} \sum_{i \in \text{mis}} \varepsilon_i + n^{-1} \sum_{i \in \text{mis}} (\mu_i(\boldsymbol{\theta}) - \mu_i(\hat{\boldsymbol{\theta}})), \end{aligned} \tag{A.1}$$

where the $\varepsilon_i = y_i - \mu_i(\boldsymbol{\theta})$ are independent random variables with mean 0 and variance $\sigma_i^2(\boldsymbol{\theta})$. Thus the first term in (A.1) is $O_p(n^{-1/2})$. For the second term, note that

$$\begin{aligned} \mu_i(\boldsymbol{\theta}) - \mu_i(\hat{\boldsymbol{\theta}}) &= \left(\frac{\partial \mu_i(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} \right)^T (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + O_p(n^{-1}) \\ &= O_p(n^{-1/2}). \end{aligned} \tag{A.2}$$

Because mis has $O(n)$ elements, the second term in (A.1) is also $O_p(n^{-1/2})$, and thus $T_{\mathbf{y}} = \hat{T}_{\mathbf{y}} + O_p(n^{-1/2})$. To establish (9), expand $\hat{Q} = g(\mathbf{T})$ in a Taylor series about $T_{\mathbf{y}} = \hat{T}_{\mathbf{y}}$,

$$\begin{aligned} &\hat{Q}(\mathbf{X}, \mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}) - \hat{Q}(\mathbf{X}, \mathbf{y}_{\text{obs}}, \boldsymbol{\mu}(\hat{\boldsymbol{\theta}})) \\ &\quad = g(\mathbf{T}) - g(\hat{\mathbf{T}}) \\ &\quad = \left(\frac{\partial g(\hat{\mathbf{T}})}{\partial T_{\mathbf{y}}} \right) (T_{\mathbf{y}} - \hat{T}_{\mathbf{y}}) + O_p(n^{-1}), \end{aligned}$$

and note that $E(T_{\mathbf{y}}) = \hat{T}_{\mathbf{y}} + O_p(n^{-1})$ by (A.1) and (A.2).

To establish (10), write $V(\hat{Q}) = EV(\hat{Q}|\boldsymbol{\theta}) + VE(\hat{Q}|\boldsymbol{\theta})$. Let $\tilde{T}_{\mathbf{y}}(\boldsymbol{\theta}) = n^{-1}[\sum_{i \in \text{obs}} y_i + \sum_{i \in \text{mis}} \mu_i(\boldsymbol{\theta})]$, so that $\tilde{T}_{\mathbf{y}}(\hat{\boldsymbol{\theta}}) = \hat{T}_{\mathbf{y}}$, and let $\tilde{\mathbf{T}}(\boldsymbol{\theta}) = (T_{X_1}, \ldots, T_{X_p}, \tilde{T}_{\mathbf{y}}(\boldsymbol{\theta}))^T$. For any fixed $\boldsymbol{\theta}$, $T_{\mathbf{y}} - \tilde{T}_{\mathbf{y}}(\boldsymbol{\theta})$ has mean 0 and variance $n^{-2} \sum_{i \in \text{mis}} \sigma_i^2(\boldsymbol{\theta})$. Thus the expansion

$$\begin{aligned} &\hat{Q}(\mathbf{X}, \mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}) - \hat{Q}(\mathbf{X}, \mathbf{y}_{\text{obs}}, \boldsymbol{\mu}(\boldsymbol{\theta})) \\ &\quad = \left(\frac{\partial g(\tilde{\mathbf{T}}(\boldsymbol{\theta}))}{\partial T_{\mathbf{y}}} \right) (T_{\mathbf{y}} - \tilde{T}_{\mathbf{y}}(\boldsymbol{\theta})) + O_p(n^{-1}) \end{aligned}$$

implies that

$$V(\hat{Q}|\boldsymbol{\theta}) = \left(\frac{\partial g(\tilde{\mathbf{T}}(\boldsymbol{\theta}))}{\partial T_{\mathbf{y}}} \right)^2 n^{-2} \sum_{i \in \text{mis}} \sigma_i^2(\boldsymbol{\theta}) + O_p(n^{-3/2}). \tag{A.3}$$

Note that the leading term in (A.3) is of order n^{-1}. Expanding $nV(\hat{Q}|\boldsymbol{\theta})$ about $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ gives

$$EV(\hat{Q}|\boldsymbol{\theta}) = \left(\frac{\partial g(\hat{\mathbf{T}})}{\partial T_{\mathbf{y}}} \right)^2 n^{-2} \sum_{i \in \text{mis}} \sigma_i^2(\hat{\boldsymbol{\theta}}) + O_p(n^{-3/2}). \tag{A.4}$$

Also, for any fixed $\boldsymbol{\theta}$, $E(\hat{Q}|\boldsymbol{\theta}) = g(\tilde{\mathbf{T}}(\boldsymbol{\theta})) + O_p(n^{-1})$. Expanding this expression for $E(\hat{Q}|\boldsymbol{\theta})$ about $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ gives $VE(\hat{Q}|\boldsymbol{\theta}) = ([\partial g(\tilde{\mathbf{T}}(\hat{\boldsymbol{\theta}}))]/\partial \boldsymbol{\theta})^T \boldsymbol{\Gamma} ([\partial g(\tilde{\mathbf{T}}(\hat{\boldsymbol{\theta}}))]/\partial \boldsymbol{\theta}) + O_p(n^{-3/2})$. But by the chain rule,

$$\begin{aligned} \frac{\partial g(\tilde{\mathbf{T}}(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} &= \frac{\partial g(\tilde{\mathbf{T}}(\boldsymbol{\theta}))}{\partial T_{\mathbf{y}}} \frac{\partial \tilde{T}_{\mathbf{y}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\ &= \frac{\partial g(\tilde{\mathbf{T}}(\boldsymbol{\theta}))}{\partial T_{\mathbf{y}}} n^{-1} \sum_{i \in \text{mis}} \left(\frac{\partial \mu_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right), \end{aligned}$$

so

$$VE(\hat{Q}|\boldsymbol{\theta}) = \left(\frac{\partial g(\hat{\mathbf{T}})}{\partial T_{\mathbf{y}}} \right) D_{\mu}(\hat{\boldsymbol{\theta}})^T \boldsymbol{\Gamma} D_{\mu}(\hat{\boldsymbol{\theta}}) + O_p(n^{-3/2}). \tag{A.5}$$

Combining (A.4) and (A.5) establishes (10).

Finally, to establish (11), note that $\mathbf{S} = n^{-1}(\mathbf{Z}^T\mathbf{Z} - n\mathbf{T}\mathbf{T}^T) + O_p(n^{-1})$, so

$$\begin{aligned} nU = &\left(\frac{\partial g(\mathbf{T})}{\partial \mathbf{T}} \right)^T n^{-1} \mathbf{Z}^T \mathbf{Z} \left(\frac{\partial g(\mathbf{T})}{\partial \mathbf{T}} \right) \\ &- \left(\frac{\partial g(\mathbf{T})}{\partial \mathbf{T}} \right)^T (\mathbf{T}\mathbf{T}^T) \left(\frac{\partial g(\mathbf{T})}{\partial \mathbf{T}} \right) + O_p(n^{-1}), \end{aligned} \tag{A.6}$$

where the leading terms are $O_p(1)$. The second term of (A.6) depends on $\mathbf{y}_{\text{mis}}$ only through $T_{\mathbf{y}}$, so by expansion about $T_{\mathbf{y}} = \hat{T}_{\mathbf{y}}$, the expectation of the second term is

$$- \left(\frac{\partial g(\hat{\mathbf{T}})}{\partial \mathbf{T}} \right)^T (\hat{\mathbf{T}}\hat{\mathbf{T}}^T) \left(\frac{\partial g(\hat{\mathbf{T}})}{\partial \mathbf{T}} \right) + O_p(n^{-1}). \tag{A.7}$$

The first term of (A.6), however, depends on $\mathbf{y}_{\text{mis}}$ through $T_{\mathbf{y}}, \mathbf{X}^T\mathbf{y}$, and $\mathbf{y}^T\mathbf{y}$. Let $Z(\boldsymbol{\theta})$ denote a mean-imputed version of $\mathbf{Z} = (\mathbf{X}, \mathbf{y})$ with $\mathbf{y}_{\text{mis}}$ replaced by $\boldsymbol{\mu}(\boldsymbol{\theta})$. For any $\boldsymbol{\theta}$,

$$n^{-1}\mathbf{Z}^T\mathbf{Z} - n^{-1}\mathbf{Z}(\hat{\boldsymbol{\theta}})^T\mathbf{Z}(\hat{\boldsymbol{\theta}}) = \mathbf{A} + \mathbf{B}, \tag{A.8}$$

where $\mathbf{A} = n^{-1}\mathbf{Z}^T\mathbf{Z} - n^{-1}\mathbf{Z}(\boldsymbol{\theta})^T\mathbf{Z}(\boldsymbol{\theta})$ and $\mathbf{B} = n^{-1}\mathbf{Z}(\boldsymbol{\theta})^T\mathbf{Z}(\boldsymbol{\theta}) - n^{-1}\mathbf{Z}(\hat{\boldsymbol{\theta}})^T\mathbf{Z}(\hat{\boldsymbol{\theta}})$. The matrix $\mathbf{A}$ has 0's everywhere except the last row and column, whose entries are

$$n^{-1} \sum_{i \in \text{mis}} x_{ij}(y_i - \mu_i(\boldsymbol{\theta})), \qquad j = 1, \ldots, p \tag{A.9}$$

and

$$n^{-1} \sum_{i \in \text{mis}} (y_i - \mu_i(\boldsymbol{\theta}))^2. \tag{A.10}$$

The conditional expectations of (A.9) and (A.10) given $\boldsymbol{\theta}$ are 0 and $n^{-1}\sum_{i \in \text{mis}} \sigma_i^2(\boldsymbol{\theta})$, so $E(\mathbf{A}) = EE(\mathbf{A}|\boldsymbol{\theta})$ is a matrix with $n^{-1}\sum_{i \in \text{mis}} \sigma_i^2(\hat{\boldsymbol{\theta}}) + O_p(n^{-1})$ in the lower right corner and 0's elsewhere. Similarly, $\mathbf{B}$ has 0's except in the last row and column, whose entries are

$$n^{-1} \sum_{i \in \text{mis}} x_{ij}(\mu_i(\boldsymbol{\theta}) - \mu_i(\hat{\boldsymbol{\theta}})), \qquad j = 1, \ldots, p \tag{A.11}$$

and

$$n^{-1}\sum_{i\in\text{mis}}(\mu_i(\boldsymbol{\theta})-\mu_i(\hat{\boldsymbol{\theta}}))^2. \tag{A.12}$$

By expansion (A.2), the expectations of (A.11) and (A.12) vanish up to terms of $O_p(n^{-1})$, so $E(B) = O_p(n^{-1})$, and thus (A.8) implies that

$$E(n^{-1}\mathbf{Z}^T\mathbf{Z}) = n^{-1}\mathbf{Z}(\hat{\boldsymbol{\theta}})^T\mathbf{Z}(\hat{\boldsymbol{\theta}}) + E(\mathbf{A}) + O_p(n^{-1}). \tag{A.13}$$

Finally, (A.13) and the fact that $\mathbf{T} - \hat{\mathbf{T}} = O_p(n^{-1/2})$ imply that the expectation of the first term in (A.6) is

$$\left(\frac{\partial g(\hat{\mathbf{T}})}{\partial \mathbf{T}}\right)^T n^{-1}\mathbf{Z}(\hat{\boldsymbol{\theta}})^T\mathbf{Z}(\hat{\boldsymbol{\theta}})\left(\frac{\partial g(\hat{\mathbf{T}})}{\partial \mathbf{T}}\right) \tag{A.14}$$

plus the remainder

$$\left(\frac{\partial g(\hat{\mathbf{T}})}{\partial \mathbf{T}}\right)^T E(\mathbf{A})\left(\frac{\partial g(\hat{\mathbf{T}})}{\partial \mathbf{T}}\right) + O_p(n^{-1/2}).$$

But because of the pattern of 0's in $E(\mathbf{A})$, this remainder simplifies to

$$\left(\frac{\partial g(\hat{\mathbf{T}})}{\partial T_{\mathbf{y}}}\right)^2 n^{-1}\sum_{i\in\text{mis}}\sigma_i^2(\hat{\boldsymbol{\theta}}) + O_p(n^{-1/2}). \tag{A.15}$$

Substituting (A.14), (A.15), and (A.7) into (A.6) proves the result.

[Received June 1997. Revised March 1999.]

REFERENCES

Belin, T. R., Diffendal, G. J., Mack, S., Rubin, D. B., Schafer, J. L., and Zaslavsky, A. M. (1993), "Hierarchical Logistic Regression Models for Imputation of Unresolved Enumeration Status in Undercount Estimation," *Journal of the American Statistical Association*, 88, 1149–1159.

Cochran, W. G. (1977), *Sampling Techniques* (2nd ed.), New York: Wiley.

Cox, D. R., and Hinkley, D. V. (1974), *Theoretical Statistics*, London: Chapman and Hall.

Dorinski, S. M., and Griffin, R. (1997), "Accounting for Variance due to Imputation in the Integrated Coverage Measurement Survey," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 748–753.

Fay, R. E. (1991), "A Design-Based Perspective on Missing Data Variance," *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of the Census, 429–440.

——— (1992), "When are Inferences From Multiple Imputation Valid?," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 227–232.

——— (1993), "Valid Inferences From Imputed Survey Data," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 41–48.

——— (1996), "Alternative Paradigms for the Analysis of Imputed Survey Data," *Journal of the American Statistical Association*, 91, 490–498.

Klein, T. M. (1986), "A Method for Estimating Posterior BAC Distributions for Persons Involved in Fatal Traffic Accidents," Report DOT-HS-807-094, National Highway Traffic Safety Administration, Dept. of Transportation.

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

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

Rao, J. N. K. (1996), "On Variance Estimation With Imputed Survey Data," *Journal of the American Statistical Association*, 91, 499–506.

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," *Journal of the American Statistical Association*, 91, 473–489.

Rubin, D. B., and Schenker, N. (1986), "Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse," *Journal of the American Statistical Association*, 81, 366–374.

Särndal, C-E. (1990), "Methods for Estimating the Precision of Survey Estimates When Imputation has Been Used," in *Proceedings of the Statistics Canada Symposium*, pp. 337–347.

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

Schafer, J. L., and Schenker, N. (1997), "Inference With Imputed Conditional Means," Technical Report 97-05, Pennsylvania State University, Dept. of Statistics.

Wolter, K. M. (1985), *Introduction to Variance Estimation*, New York: Springer-Verlag.