

Inference with Imputed Conditional Means

Joseph L. Schafer and Nathaniel Schenker *

June 4, 1997

Abstract

In this paper, we develop analytic techniques that can be used to produce appropriate inferences from a data set in which imputation for missing values has been carried out using predictive means. Our 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 the estimators that would be used with multiple imputation. The procedures developed can be used either for univariate missing data or for multivariate missing data in which the variables are either missing or observed together, and they are designed for situations in which the complete-data estimator is a smooth function of linear statistics. We illustrate properties of our methods in several examples, including abstract problems as well as applications to large data sets from studies carried out by the federal government.

Key Words: Linearization; Missing data; Multiple Imputation; Nonresponse; Taylor series.

1 Introduction

A standard technique for handling missing data in a large data set such as that obtained from a sample survey is to impute (i.e., fill in) a plausible value for each missing datum, and then to analyze the resulting data set as if there were no missing data. Imputation is attractive because it results in a completed data set and thus allows the use of standard complete-data methods of analysis. In addition, imputations are often created by people who are connected with the collection of the data and who may therefore have more

*Joseph L. Schafer is Assistant Professor, Department of Statistics, Pennsylvania State University, University Park, PA 16802-6202. Nathaniel Schenker is Associate Professor, Department of Biostatistics, UCLA School of Public Health, Los Angeles, CA 90095-1772. The authors' names are listed in alphabetical order.

information available to model nonresponse than ultimate data analysts. A drawback of imputation followed by the use of complete-data methods of analysis, however, 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 taken into account (e.g., Little and Rubin 1987, Chapter 3).

Multiple imputation (Rubin 1987) provides a general framework for incorporating the uncertainty due to missing data into inference. The idea is to create several completed data sets by imputing draws of the missing values several times from their predictive distribution. Then the standard complete-data analysis is applied to each completed data set, and the results are combined in such a way that the variance across imputations is included.

In this paper, we develop an analytic method that can be used to produce appropriate variance estimates in special cases with just a single imputation of predictive means. Our method can be useful in situations where generating and/or managing a multiply-imputed data set is difficult. The method: (1) can be used when there is a single variable subject to missing data (an extension is outlined for multivariate missing data as well), and when the complete-data estimator is a smooth function of linear statistics; (2) is based on asymptotic expansions of point estimators and their associated variance estimators, and can be thought of as a first-order approximation to what would be obtained from an infinite number of multiple imputations; (3) is computationally simplest when missing data can be modeled with a single-parameter error distribution, such as Bernoulli or Poisson.

Section 2 discusses our assumptions about the inference problem and the missing data. Section 3 contains comments on various imputation strategies (mean, single random, and multiple imputation), and it is followed by a presentation of 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, as well as applications to missing-data problems that have occurred in practice. Finally, a concluding discussion is given in Section 6.

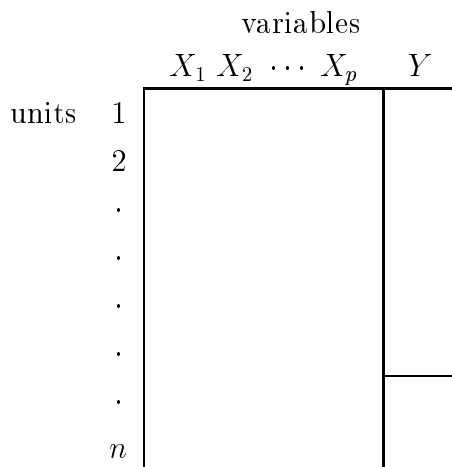


Figure 1: Rectangular data set, one variable subject to missing data.

2 Setup and Assumptions

2.1 Pattern of Missing Data

Consider a data set with n observational units, in which a single variable Y is subject to missing data, whereas other variables X_1, \dots, X_p are completely observed; Figure 1 presents a schematic diagram of such a data set. (In Section 4.5, we outline a generalization of our results to multivariate Y .) Let X be the $n \times p$ matrix of observed data for X_1, \dots, X_p , and let y denote the $n \times 1$ vector of Y -values. Then y can be partitioned into sets of observed and missing components, y_{obs} and y_{mis} , with lengths n_1 and $n_0 = n - n_1$, respectively. 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 one as $n \rightarrow \infty$.

2.2 Estimation with Complete Data

Let Q denote a scalar quantity to be estimated. If the data were complete, typical analyses would be based on a point estimate, $\hat{Q} = \hat{Q}(X, y)$, and an associated estimate of variance for \hat{Q} , $U = U(X, y)$. The point estimators \hat{Q} that we consider in this paper are smooth functions of linear statistics. Our emphasis is on simple random samples from infinite populations and smooth functions of means, although in Section 5.4, we mention an extension

to complex samples from finite populations. Let

$$\hat{Q} = g(T_{X_1}, \dots, T_{X_p}, T_y), \quad (1)$$

where $T_{X_j} = n^{-1} \sum_{i=1}^n X_{ij}$, $j = 1, \dots, p$, $T_y = n^{-1} \sum_{i=1}^n y_i$, X_{ij} denotes the value of X_j for unit i , y_i denotes the value of Y (observed or missing) 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 linear statistics,

$$Q = g(ET_{X_1}, \dots, ET_{X_p}, ET_y),$$

where the expectations are taken over repeated sampling of X and y ; hence \hat{Q} can be thought of as a method-of-moments estimate of Q . The form (1) includes many estimators typically used in survey practice and elsewhere, such as means and proportions, ratios of means, etc., but does not include medians, variances, or correlations.

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

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

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

Typically, inferences are based on a normal reference distribution. In accord with standard practice, then, we will assume that

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

as $n \rightarrow \infty$. When the units in the data set constitute a simple random sample from an infinite population, as we are assuming here, (3) is easily verified by appealing to standard central limit theorem arguments and Slutsky's theorem.

2.3 Modeling Missing Data

When values of the variable Y are missing, X provides useful information for predicting the missing values to the extent that the variables X_1, \dots, X_p 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 y_{mis} . This approach implicitly assumes that the conditional distribution of Y given X_1, \dots, X_p when Y is missing is the same as it is when Y is observed; this is appropriate if the nonresponse 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, of course, provide no information to support or contradict this assumption; such support must come from a source external to the observed data. Other approaches are possible, but every missing-data procedure must be based on some assumption that cannot be verified from (X, y_{obs}) alone. In this paper, we assume that nonresponse is ignorable and that a probability model for y_{mis} given (X, y_{obs}) has been correctly specified. A typical specification for this model will include unknown but estimable parameters, which we will refer to as θ .

Let $\hat{\theta}$ denote an efficient estimate of θ based on (X, y_{obs}) under the assumed model for missing data. Also, let Γ denote an estimate of $V(\theta - \hat{\theta})$, also based on (X, y_{obs}) . For example, $\hat{\theta}$ may be a maximum likelihood (ML) estimate, and Γ may be the inverse of the observed information matrix evaluated at $\hat{\theta}$. We assume that $\Gamma = O(n^{-1})$ and that

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

where 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, Chapter 9) applies, that the fraction of missing information is bounded away from one, and that the dimension of θ is fixed. We assume further that the model for missing data imposes an uncorrelated (given θ) error structure on the missing values. More precisely, let mis denote the set of indices i such that $y_i \in y_{mis}$. Then for all $i, i' \in mis$, we assume that

$$E(y_i | X, y_{obs}, \theta) = \mu_i(\theta),$$

$$V(y_i | X, y_{obs}, \theta) = \sigma_i^2(\theta),$$

and

$$\text{Cov}(y_i, y_{i'} \mid X, y_{obs}, \theta) = 0,$$

where μ_i and σ_i^2 are smooth, well-behaved functions of θ .

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, \dots, X_p in a rectangular data set. In Section 5.4, we will modify our results to allow for intra-cluster correlation in complex samples.

3 Approaches to Imputing for Missing Data

Once a model for y_{mis} given (X, y_{obs}) has been specified, there are several approaches that can be taken to impute for y_{mis} . We now comment briefly on three such approaches.

3.1 Conditional Mean Imputation

Let $\mu(\theta)$ denote the vector with elements $\mu_i(\theta)$, $i \in mis$; that is,

$$\mu(\theta) = E(y_{mis} \mid X, y_{obs}, \theta).$$

An approach that seeks to fill in the missing data with one set of “best” values might choose $\mu(\hat{\theta})$. Little and Rubin (1987, Section 3.4) refer 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 $\hat{Q}(X, y_{obs}, \mu(\hat{\theta}))$ is a first-order approximation to the “best” estimate of Q . Inferences can be seriously distorted with conditional mean imputation, however. The analogue to (3) with conditional means imputed for y_{mis} does not hold in general because $U(X, y_{obs}, \mu(\hat{\theta}))$ is usually biased downward,

$$EU(X, y_{obs}, \mu(\hat{\theta})) < V(Q - \hat{Q}(X, y_{obs}, \mu(\hat{\theta})));$$

see Little and Rubin (1987, Section 3.4) for examples.

3.2 Single Random Imputation

Rather than imputing predicted means for y_{mis} , another strategy is to impute at random from an estimate of the distribution of y_{mis} . For example, one might impute y_{mis}^* , a random draw from $\mathcal{L}(y_{mis} \mid X, y_{obs}, \hat{\theta})$, and base inferences on $\hat{Q}(X, y_{obs}, y_{mis}^*)$ and $U(X, y_{obs}, y_{mis}^*)$.

For point estimation, single random imputation is less efficient than conditional mean imputation because the random imputation mechanism introduces additional variance into the point estimator,

$$V(\hat{Q}(X, y_{obs}, y_{mis}^*)) > V(\hat{Q}(X, y_{obs}, \mu(\hat{\theta}))).$$

The variance estimate $U(X, y_{obs}, y_{mis}^*)$ tends to be larger than its conditional-mean-imputed counterpart $U(X, y_{obs}, \mu(\hat{\theta}))$. Since the variance being estimated is also larger, however, there is still typically a downward bias,

$$EU(X, y_{obs}, y_{mis}^*) < V(Q - \hat{Q}(X, y_{obs}, y_{mis}^*));$$

see, for example, Rubin (1987, Problem 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, y_{mis} is replaced by M random draws from a predictive distribution. To obtain proper inferences, the distribution from which the values of y_{mis} are drawn must incorporate variability due to both uncertainty about both the parameter θ of the missing-data model and the randomness of y_{mis} given θ . Using Bayesian notation, we can write the predictive density of y_{mis} as

$$p(y_{mis} \mid X, y_{obs}) = \int p(y_{mis} \mid X, y_{obs}, \theta)p(\theta \mid X, y_{obs})d\theta, \quad (5)$$

which makes the two sources of variability explicit. Multiple imputation results in M completed data sets and M complete-data analyses; the results of these M analyses are then combined to produce a single overall inference that 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 Methods for Conditional Mean Imputation

We now develop a method of drawing inferences for Q from a data set in which the missing values of Y have been replaced by conditional means. Our method can be considered a linear approximation to a full analysis using multiply-imputed data. Our results are restricted to simple random samples from an infinite population; extensions are indicated in Section 5.4.

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 , respectively,

$$\begin{aligned}\hat{Q} &= E(Q \mid X, y_{obs}, y_{mis}), \\ U &= V(Q \mid X, y_{obs}, y_{mis}).\end{aligned}$$

Under sufficient regularity conditions, posterior means and variances behave as in (3) (e.g., Cox and Hinkley 1974, Chapter 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 $\hat{\theta}$ and Γ as posterior moments of θ given the observed data,

$$\hat{\theta} = E(\theta \mid X, y_{obs}),$$

$$\Gamma = V(\theta | X, y_{obs}).$$

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, $E(Q | X, y_{obs})$ and $V(Q | X, y_{obs})$. Note that

$$E(Q | X, y_{obs}) = E(\hat{Q} | X, y_{obs}) \quad (6)$$

and

$$V(Q | X, y_{obs}) = V(\hat{Q} | X, y_{obs}) + E(U | X, y_{obs}), \quad (7)$$

where the moments on the right-hand side of these equations are evaluated over the posterior predictive distribution $\mathcal{L}(y_{mis} | X, y_{obs})$, the distribution corresponding to (5) from which multiple imputations would be drawn. To obtain approximate posterior moments of Q , then, we need only to approximate the mean and variance of \hat{Q} and the mean of U over the predictive distribution of y_{mis} .

4.2 Approximate Moments of \hat{Q} and U

Approximations to the posterior moments of \hat{Q} and U are now given. These results, for which derivations are outlined in the appendix, follow from first-order Taylor series expansions of the functions g and μ_i , $i \in mis$, and large-sample results from the theory of sample surveys (e.g., Wolter 1985, Chapter 6).

Under the assumptions outlined in Section 2:

$$E(\hat{Q} | X, y_{obs}) = \hat{Q}(X, y_{obs}, \mu(\hat{\theta})) + O_p(n^{-1}); \quad (8)$$

$$V(\hat{Q} | X, y_{obs}) = \left(\frac{\partial g(\hat{T})}{\partial T_y}\right)^2 n^{-2} \sum_{i \in mis} \sigma_i^2(\hat{\theta}) + \left(\frac{\partial g(\hat{T})}{\partial T_y}\right)^2 D_\mu(\hat{\theta})^T \Gamma D_\mu(\hat{\theta}) + O_p(n^{-3/2}), \quad (9)$$

where \hat{T} is shorthand for the complete-data statistic T calculated with $\mu(\hat{\theta})$ substituted for y_{mis} , and where $D_\mu(\theta) = n^{-1} \sum_{i \in mis} \left(\frac{\partial \mu_i(\theta)}{\partial \theta}\right)$; and finally,

$$E(U | X, y_{obs}) = U(X, y_{obs}, \mu(\hat{\theta})) + \left(\frac{\partial g(\hat{T})}{\partial T_y}\right)^2 n^{-2} \sum_{i \in mis} \sigma_i^2(\hat{\theta}) + O_p(n^{-3/2}). \quad (10)$$

Note that the moments in equations (8)–(10) do not necessarily exist for any finite n ; each moment should be interpreted as the moment of a limiting distribution, not as the limit of a sequence of moments. For example, (6) means that \hat{Q} can be written as the sum of a random variable with mean $\hat{Q}(X, y_{obs}, \mu(\hat{\theta}))$ and a second random variable that is $O_p(n^{-1})$.

4.3 Point Estimation

It follows from (6) and (8) that the complete-data point estimate with conditional means imputed for the missing values of Y is a first-order approximation to the posterior mean of Q ,

$$E(Q \mid X, y_{obs}) \approx \hat{Q}(X, y_{obs}, \mu(\hat{\theta})). \quad (11)$$

In large samples, then, it is desirable to use $\hat{Q}(X, y_{obs}, \mu(\hat{\theta}))$, 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. The result does not hold for an arbitrary estimator \hat{Q} , such as a sample variance. In fact, equation (10) demonstrates that a conditional-mean-imputed sample variance is biased downward.

4.4 Variance Estimation

It follows from (7), (9), and (10) that a first-order approximation to the posterior variance of Q is

$$V(Q \mid X, y_{obs}) \approx U(X, y_{obs}, \mu(\hat{\theta})) + C_1 + C_2, \quad (12)$$

where

$$C_1 = 2 \left(\frac{\partial g(\hat{T})}{\partial T_y} \right)^2 n^{-2} \sum_{i \in mis} \sigma_i^2(\hat{\theta}) \quad (13)$$

and

$$C_2 = \left(\frac{\partial g(\hat{T})}{\partial T_y} \right)^2 D_\mu(\hat{\theta})^T \Gamma D_\mu(\hat{\theta}). \quad (14)$$

In (12), $U(X, y_{obs}, \mu(\hat{\theta}))$ is the “naive” estimate, discussed in Section 3, that treats the conditional-mean-imputed data set as complete data. The first correction term, C_1 , is a

component of variance that accounts for uncertainty in y_{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, i.e., uncertainty due to the estimation of the parameters in the missing-data model.

The term C_1 can be very simple to compute. For example, if the estimand Q is the population mean of Y , and the missing values of Y are modeled by ordinary linear regression, then C_1 has the form $C_1 = 2n^{-2} \sum_{i \in mis} \hat{\sigma}^2$, where $\hat{\sigma}^2$ is the estimated residual variance of the regression. If Y is a binary variable, and the missing values of Y are modeled as Bernoulli with means π_i , $i \in mis$ (e.g., by logistic regression), then C_1 has the form $C_1 = 2n^{-2} \sum_{i \in mis} \hat{\pi}_i(1 - \hat{\pi}_i)$ for estimating the mean of Y . As this latter example illustrates, when the elements of y_{mis} are modeled with an error distribution that has a single parameter (e.g., Bernoulli or Poisson), then the variances $\sigma_i^2(\theta)$ can be expressed as functions of the means $\mu_i(\theta)$, and therefore C_1 can be computed from the imputed data set alone. When the error distribution has additional parameters, however, estimates of these parameters need to be retained to calculate C_1 , as illustrated by the former example involving modeling of y_{mis} by ordinary linear regression in which $\hat{\sigma}^2$ is needed.

The second correction term, C_2 , is usually more difficult to compute than C_1 , because it involves the variance of the parameters θ in the missing-data model. When the fraction of missing information is moderate, however, the proportion of variance in (12) contributed by C_2 can become small enough that C_2 can be omitted, simplifying calculations; see Section 5 for examples of this phenomenon.

4.5 Generalization to Multivariate Missing Data

Our results for univariate Y generalize to a special case of multivariate Y with only small differences in notation. Suppose as in Section 2 that X_1, \dots, X_p are completely observed. Suppose further that the variables Y_1, \dots, Y_q are only partially observed in such a way that they are either observed or missing together; that is, Y_j is missing for a unit if and only if

$Y_{j'}$, $j' \neq j$ is also missing, $1 \leq j, j' \leq q$. Let Y be the $n \times q$ matrix of Y values, with rows y_i^T , $i = 1, \dots, n$. Then Y can be partitioned into observed and missing components, Y_{obs} and Y_{mis} .

The complete-data point estimate is now of the form

$$\hat{Q}(X, Y) = g(T), \quad (15)$$

where $T = (T_{X_1}, \dots, T_{X_p}, T_{Y_1}, \dots, T_{Y_q})^T$ is the vector of complete-data means for all variables. The complete-data variance is still of the form (2), where $S = (n-1)^{-1}(Z^T Z - nTT^T)$ is now a $(p+q) \times (p+q)$ matrix and $Z = (X, Y)$. The missing-data model is now a multivariate model, in which it is assumed that the y_i are independently distributed given (X, Y_{obs}, θ) for all $i \in mis$ with mean vectors

$$\mu_i(\theta) = E(y_i | X, Y_{obs}, \theta)$$

and covariance matrices

$$\Sigma_i(\theta) = V(y_i | X, Y_{obs}, \theta)$$

that are smooth functions of θ . We still assume that the dimension of θ is fixed and that

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

for some $\hat{\theta}$ and $\Gamma = O(n^{-1})$ that are functions of (X, Y_{obs}) .

Results (8), (9), and (10) extend immediately to

$$E(\hat{Q} | X, Y_{obs}) = \hat{Q}(X, Y_{obs}, \mu(\hat{\theta})) + O_p(n^{-1}), \quad (16)$$

$$\begin{aligned} V(\hat{Q} | X, Y_{obs}) &= \left(\frac{\partial g(\hat{T})}{\partial T_Y} \right)^T \left(n^{-2} \sum_{i \in mis} \Sigma_i(\hat{\theta}) \right) \left(\frac{\partial g(\hat{T})}{\partial T_Y} \right) \\ &\quad + \left(\frac{\partial g(\hat{T})}{\partial T_y} \right)^T D_\mu(\hat{\theta}) \Gamma D_\mu(\hat{\theta})^T \left(\frac{\partial g(\hat{T})}{\partial T_y} \right) + O_p(n^{-3/2}), \end{aligned} \quad (17)$$

and

$$E(U | X, Y_{obs}) = U(X, Y_{obs}, \mu(\hat{\theta})) + \left(\frac{\partial g(\hat{T})}{\partial T_y} \right)^T \left(n^{-2} \sum_{i \in mis} \Sigma_i(\hat{\theta}) \right) \left(\frac{\partial g(\hat{T})}{\partial T_y} \right) + O_p(n^{-3/2}), \quad (18)$$

where $\mu(\theta) = E(Y_{mis} \mid X, Y_{obs}, \theta)$ is the $n_0 \times q$ matrix with $\mu_i(\theta)^T$, $i \in mis$ as its rows, \hat{T} is shorthand for the complete-data statistic T calculated with $\mu(\hat{\theta})$ substituted for Y_{mis} , $T_Y = (T_{Y_1}, \dots, T_{Y_q})^T$, and $D_\mu(\theta) = n^{-1} \sum_{i \in mis} \left(\frac{\partial \mu_i(\theta)}{\partial \theta} \right)$ is a $q \times \dim(\theta)$ matrix of derivatives.

5 Examples

We now present several examples to illustrate the methods derived in Section 4. We begin with two abstract problems to demonstrate properties of our methods, and then we discuss three applications of our methods, including one (Section 5.4) that involves an extension to complex surveys.

5.1 Estimating a Binomial Proportion

Let y denote a simple random sample of size n of binary (0-1) variables from an infinite population, and let the estimand be $Q = p$, the population proportion of Y -values that are equal to one. With complete data (and n not small), standard inferences for p would be based on the point estimate $\hat{Q} = \bar{y} \equiv \frac{1}{n} \sum_i y_i$ and the associated variance estimate $U = \frac{1}{n(n-1)} \sum_i (y_i - \bar{y})^2$.

Suppose that Y is subject to missingness completely at random, so that y_{obs} is just a simple random sample from y . Imputation is not actually necessary in this simple situation, since valid inferences may be drawn by ignoring the missing values. Since the correct answer is known, however, this example provides a simple check of the validity of our approach. In addition, the example provides insight into the relative importance of the components of variance C_1 and C_2 discussed in Section 4.4.

If the missing data y_{mis} are modeled as a vector of i.i.d. Bernoulli(θ) random variables, then in the notation of Section 2.3, $\mu_i(\theta) = \theta$ and $\sigma_i^2(\theta) = \theta(1 - \theta)$. Maximum likelihood estimation based on y_{obs} yields $\hat{\theta} = \frac{1}{n_1} \sum_{i \in obs} y_i$ (where obs denotes the set of indices such that $y_i \in y_{obs}$ and $\Gamma = \hat{\theta}(1 - \hat{\theta})/n_1$). Substitution into equations (11)–(14) and algebraic

manipulation yield

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

$$U(X, y_{obs}, \mu(\hat{\theta})) \approx r_1^2 \frac{1}{n_1(n_1 - 1)} \sum_{i \in obs} (y_i - \bar{y}_1)^2, \quad (19)$$

$$C_1 \approx 2r_1r_0 \frac{1}{n_1(n_1 - 1)} \sum_{i \in obs} (y_i - \bar{y}_1)^2, \quad (20)$$

and

$$C_2 \approx r_0^2 \frac{1}{n_1(n_1 - 1)} \sum_{i \in obs} (y_i - \bar{y}_1)^2; \quad (21)$$

the equalities are approximate 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

$$U(X, y_{obs}, \mu(\hat{\theta})) + C_1 + C_2 \approx \frac{1}{n_1(n_1 - 1)} \sum_{i \in obs} (y_i - \bar{y}_1)^2 \quad (22)$$

as the associated variance estimate. These are the same point and variance estimates that would be obtained by applying standard complete-data methods to just the observed data, and thus our methods yield the correct estimates in this example.

Equations (19)–(22) show that the proportionate contributions of $U(X, y_{obs}, \mu(\hat{\theta}))$, C_1 , and C_2 to the correct variance estimate are approximately r_1^2 , $2r_1r_0$, and r_0^2 , respectively. One important implication is that, even if the missingness rate is moderate, simply imputing conditional means for y_{mis} and then using the complete-data estimate of variance can result in very misleading inferences. For example, if $r_0 = 20\%$, $U(X, y_{obs}, \mu(\hat{\theta}))$ is approximately 36% smaller than the correct variance estimate. Another important implication, mentioned at the end of Section 4.4, is that if the nonresponse 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 from our variance estimate results in only an approximate 4% decrease from the correct estimate.

5.2 Estimating a Ratio of Means: The Fieller-Creasey Problem

Let $x_i, y_i, i = 1, \dots, n$ be i.i.d. observations 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, standard inferences for Q would be based on the point estimate $\hat{Q} = \bar{x}/\bar{y}$ and the associated variance estimate $U = \frac{1}{n\bar{y}^2}(s_X^2 - 2\hat{Q}s_{XY} + \hat{Q}^2s_Y^2)$ (e.g., Cochran 1977, Section 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 subject to missingness. A standard model to predict the variable Y from the variable X specifies that $y_i = \beta_0 + \beta_1 x_i + \sigma e_i, i = 1, \dots, n$, where the errors e_1, \dots, e_n are i.i.d. standard normal random variables. If the model is fitted by least squares applied to the complete cases ($x_i, y_i, i \in obs$), then an approximate posterior distribution for $\theta = (\beta_0, \beta_1, \log(\sigma^2))$ is normal with mean $\hat{\theta}$ and variance I_{cc}^{-1} , where I_{cc} is the observed information matrix based on the complete cases.

By equations (13) and (14), we have $C_1 = 2 \left(\frac{\hat{Q}}{\hat{y}} \right)^2 \frac{n_0}{n^2} \hat{\sigma}^2$ and $C_2 = \left(\frac{\hat{Q}}{\hat{y}} \right)^2 D_\mu(\hat{\theta})^T \Gamma D_\mu(\hat{\theta})$, where \hat{Q} and \hat{y} are the complete-data quantities calculated with conditional means imputed for y_{mis} , $D_\mu(\hat{\theta}) = n^{-1} \left(\sum_{i \in mis}^{n_0} x_i \right)$, $\Gamma = \hat{\sigma}^2 (A^T A)^{-1}$, and A is the $n_1 \times 2$ matrix with ones in the first column and $(x_i, i \in obs)^T$ as the second column.

We conducted a small simulation study in which, on each Monte Carlo trial, we generated a random sample ($x_i, y_i, i = 1, \dots, 250$) from a bivariate normal distribution with $\mu_X = \mu_Y = 5, \sigma_X^2 = \sigma_Y^2 = 1$, and $\rho = .8$. We imposed on the Y -values a random pattern of missingness with the probability of missingness depending only on the values of X :

$$P(y_i \text{ is missing} \mid x_i, y_i) = P(y_i \text{ is missing} \mid x_i).$$

Specifically, we generated observations w_1, \dots, w_{250} of a third variable W , distributed as normal with mean 0, variance 1, and correlation with X of .8, but having zero partial correlation with Y given X . We 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 was the missingness

probability. Thus, missingness on Y tended to occur for large values of X (and therefore Y), although the missing data were missing at random (Rubin 1976).

We carried out 1,000 Monte Carlo trials. For each trial, we calculated nominal 95% confidence intervals for Q using six methods: (i) the standard complete-data analysis applied with no data missing, which of course is only possible in a situation such as a simulation study, in which the missing values can be treated as observed; (ii) the standard complete-data analysis applied to the conditional-mean-imputed data set (see Section 3.1); (iii) the standard complete-data analysis applied to the data set with single random imputations for the missing values (see Section 3.2); (iv) the multiple-imputation analysis (e.g., Rubin and Schenker 1991) applied to the multiply-imputed data set with $M = 5$ imputations for the missing values (see Section 3.3); (v) the analysis developed for conditional mean imputation in Section 4, but with only the first correction term, C_1 , added into the variance; (vi) the analysis developed for conditional mean imputation in Section 4, with both correction terms C_1 and C_2 added into the variance.

The coverage rates of the confidence intervals under a variety of missingness probabilities are shown in Table 1. The coverage rates with no missing data are always within .01 of the nominal level, indicating that the complete-data analysis is approximately valid and that the data were simulated correctly. All of the methods perform well when the missingness probability is either .05 or .1, but the naive conditional mean method (no correction terms added into the variance) and the single imputation method break down for the missingness probabilities of .25 and .5, consistent with the discussions of Sections 3.1 and 3.2. The conditional mean method with C_1 added into the variance breaks down for the missingness probability of .5, consistent with the discussion in Section 4.4. The multiple imputation method and the corrected conditional mean methods (C_1 and C_2 added into the variance) perform well for all of the missingness probabilities considered, although the coverage rates for multiple imputation are slightly below the nominal levels.

Once again, omitting C_2 from the estimated variance with conditional mean imputation

Table 1: Coverage Rates (in %) of Nominal 95% Intervals for the Ratio of Means

Method	Probability of Missingness			
	.05	.10	.25	.50
No missing data	94	96	95	95
Naive conditional mean imputation	93	94	84	59
Single random imputation	92	94	87	70
Multiple random imputation ($M = 5$)	93	94	90	92
Naive conditional mean imputation + C_1	94	95	93	81
Naive conditional mean imputation + $C_1 + C_2$	94	96	95	95

does not have a large effect until the missingness probability is high. Table 2 shows the average proportion of the total estimated variance, $U(X, y_{obs}, \mu(\hat{\theta})) + C_1 + C_2$, that is due to each of the three terms. Note that the fraction due to C_2 is higher than the square of the missingness probability, unlike in Section 5.1. This is due to the fact that the simulated missingness mechanism in this example tends to result in missing data at points with high leverage; consequently, the actual fraction of missing information is actually higher than the fraction of missing Y -values. Conversely, in the example of Section 5.1, with a single variable, missingness completely at random, and estimation of a simple proportion, the fraction of missing information and the fraction of missing Y -values are the same. In general, performance of methods for handling missing data is more directly related to the fraction of missing information than to the simple fraction of missing data.

5.3 Missing Data on Blood Alcohol Content in the Fatal Accident Reporting System

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

Table 2: Average Proportion of Total Estimated Variance, $U(X, y_{obs}, \mu(\hat{\theta})) + C_1 + C_2$, Due to Each Term

Term	Probability of Missingness			
	.05	.10	.25	.50
$U(X, y_{obs}, \mu(\hat{\theta}))$.905	.803	.517	.174
C_1	.084	.159	.298	.281
C_2	.011	.038	.186	.545

BAC determination at the scene of an accident, 50% or more of the BAC values are missing. The rates of missing information for this variable tend to be somewhat 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.

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 one for the observed BAC class and zeros 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 extensively used 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, however, it is possible to reconstruct some simple measures of missing-data uncertainty from the existing FARS data files.

Consider the problem of estimating the proportion of drivers falling into BAC classes $j = 1, 2, 3$ in some domain \mathcal{D} of interest (e.g. passenger-car drivers between the ages of 21 and 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 BAC class j . The estimated proportion of the drivers in \mathcal{D} belonging to class j is $\hat{Q}_j = n_{\mathcal{D}}^{-1} \sum_{i \in \mathcal{D}} \hat{\pi}_{ij}$, where $n_{\mathcal{D}}$ is the number of individuals observed in \mathcal{D} . The estimate \hat{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. The appropriate variance estimate to attach to \hat{Q}_j depends on whether we regard the complete data $y = \{y_{ij} : i \in \mathcal{D}\}$ as a complete enumeration of the existing population, or as a realized sample from a hypothetical superpopulation.

If we adopt the view that y is a sample from a superpopulation, then the complete-data proportion \hat{Q}_j is approximately normally distributed about the unknown superpopulation proportion, with estimated variance

$$U_j = \frac{1}{n_{\mathcal{D}}(n_{\mathcal{D}} - 1)} \sum_{i \in \mathcal{D}} (y_{ij} - \hat{Q}_j)^2.$$

The naive variance estimate calculated from the mean-imputed data,

$$\hat{U}_j = \frac{1}{n_{\mathcal{D}}(n_{\mathcal{D}} - 1)} \sum_{i \in \mathcal{D}} (\hat{\pi}_{ij} - \hat{Q}_j)^2,$$

could substantially understate the uncertainty associated with \hat{Q}_j . This understatement can be partially corrected by adding to \hat{U}_j the factor $C_1 = 2n_{\mathcal{D}}^{-2} \sum_{i \in \mathcal{D}} \hat{\pi}_{ij}(1 - \hat{\pi}_{ij})$. A full correction would also require the additional factor C_2 given by (14) which cannot be obtained from the FARS data files alone; calculation of C_2 would require re-fitting of the discriminant models that produced the estimates $\hat{\pi}_{ij}$.

If we adopt the view that $y = \{y_{ij} : i \in \mathcal{D}\}$ is a complete enumeration of the relevant population, then \hat{Q}_j becomes a census statistic with no sampling variance ($U_j = 0$). The only uncertainty associated with \hat{Q}_j is then the variance of \hat{Q}_j over the predictive distribution of the missing data; see (7). It follows from (9), (13), and (14) that an estimate of

Table 3: Estimated percentage of 1993 FARS drivers by age in three BAC classes, with estimated standard errors calculated under the superpopulation (SE_1) and complete-enumeration (SE_2) assumptions

(a) Drivers of motorcycles										
Age	$n_{\mathcal{D}}$	BAC = 0			$0 < \text{BAC} < .10$			$\text{BAC} \geq .10$		
		est.	SE_1	SE_2	est.	SE_1	SE_2	est.	SE_1	SE_2
12–20	356	77.4	2.2	0.6	10.8	1.6	0.5	11.8	1.7	0.5
21–29	897	54.5	1.7	0.4	12.2	1.1	0.4	33.2	1.6	0.5
30–39	687	44.2	1.9	0.5	11.6	1.2	0.4	44.2	1.9	0.5
40–49	337	52.0	2.7	0.7	10.2	1.7	0.5	37.8	2.6	0.7
50–59	113	63.5	4.5	1.2	11.0	3.0	0.9	25.4	4.1	1.1
60+	64	84.9	4.5	1.6	4.7	2.7	1.1	10.3	3.8	1.4

(b) Drivers of passenger cars										
Age	$n_{\mathcal{D}}$	BAC = 0			$0 < \text{BAC} < .10$			$\text{BAC} \geq .10$		
		est.	SE_1	SE_2	est.	SE_1	SE_2	est.	SE_1	SE_2
12–20	5083	77.1	0.6	0.1	7.9	0.4	0.1	15.0	0.5	0.1
21–29	7500	59.6	0.6	0.1	9.1	0.3	0.1	31.3	0.5	0.1
30–39	5581	63.4	0.6	0.1	7.0	0.3	0.1	29.6	0.6	0.1
40–49	3540	73.9	0.7	0.2	5.5	0.4	0.2	20.6	0.7	0.2
50–59	2257	81.7	0.8	0.2	4.1	0.4	0.2	14.2	0.7	0.2
60+	5525	92.1	0.4	0.1	2.7	0.2	0.1	5.2	0.3	0.1

this variance is $C_1/2 + C_2$, the first term of which can be calculated from the FARS data files.

Some results of these procedures applied to 1993 FARS data are shown in Table 3 for (a) drivers of motorcycles and (b) drivers of passenger cars. In this table, $SE_1 = \sqrt{\hat{U} + C_1}$ and $SE_2 = \sqrt{C_1/2}$ refer to standard errors calculated under the superpopulation and complete-enumeration assumptions, respectively. Both should be regarded as estimated lower bounds, because they omit the term C_2 due to parameter uncertainty in the discriminant model.

5.4 Extension to Complex Surveys, with Application to Census Coverage Estimation

The methods discussed thus far are appropriate when the data from the n observational units can be regarded as independent and identically distributed, as in a simple random sample from a large population. Here we present an extension (for univariate Y) to weighted estimates from complex surveys.

Let w_i denote a weight (e.g., inverse probability of sample selection) associated with unit i , and suppose that \hat{Q} is now a smooth function of weighted sums of the values of the variables,

$$\begin{aligned}\hat{Q} &= g(T_{X_1}, \dots, T_{X_p}, T_y) \\ &= g\left(\sum_{i=1}^n w_i X_{i1}, \dots, \sum_{i=1}^n w_i X_{ip}, \sum_{i=1}^n w_i y_i\right).\end{aligned}\tag{23}$$

To stabilize the arguments of g in (23) we require that $\max_i w_i = O(n^{-1})$, which can be achieved by scaling the weights to sum to one. The weights are allowed to be functions of the observed data (X, Y_{obs}) ; for example, \hat{Q} may be a poststratified estimator with poststrata defined by categories of X . The weights may not, however, be functions of the missing data Y_{mis} .

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

$$U = \left(\frac{\partial g(T)}{\partial T}\right)^T W \left(\frac{\partial g(T)}{\partial T}\right),\tag{24}$$

where $T = (T_{X_1}, \dots, T_{X_p}, T_y)^T$ and where W is the classical unbiased variance estimator for T in a stratified probability-proportional-to-size (pps) cluster sample (e.g., Wolter 1985, Chapter 1); for example, the diagonal element of W corresponding to the variance of T_{X_k} has the form

$$\sum_s \frac{1}{n_s(n_s - 1)} \sum_c \left(n_s \sum_i w_i X_{ik} - \sum_{c,i} w_i X_{ik} \right)^2,$$

where s indexes sampling strata, c indexes clusters within strata, and i indexes units within clusters. The estimate given in (24) is a very common in survey practice; it includes, as

special cases, variance estimates for simple random samples, stratified and cluster samples, unequal probability samples (such as pps designs), and many multistage designs. The derivatives in (2) account for potential nonlinearity in g via Taylor series linearization (e.g., Wolter 1985, Chapter 6).

In accord with standard survey practice, we will continue to assume that (3) holds so that inferences may be based on a normal reference distribution. Results such as (3) have been demonstrated for finite populations and complex sample designs under a variety of conditions (e.g., Wolter 1985, Appendix B). Even if such a result has not been formally demonstrated for a particular situation, survey practitioners have still found that appealing to asymptotic normality often provides a useful first-order approximation for statistical inference.

With regard to the model for missing data, we continue to make the assumptions discussed in Section 2.3, except that the parameter vector θ may now include components for modeling intra-cluster correlations. Thus, we assume that

$$E(y_i | X, y_{obs}, \theta) = \mu_i(\theta)$$

for $i \in mis$, and that

$$V(y_{mis} | X, y_{obs}, \theta) = \Sigma(\theta),$$

a block-diagonal covariance matrix, where μ_i and Σ are smooth functions of θ . A simple model might specify that, for each block of $\Sigma(\theta)$ corresponding to a cluster, the diagonal elements are constant and the off-diagonal elements are constant; this would imply that within clusters, the components of y_{mis} have constant variance and constant pairwise correlations.

Under these new conditions, the results analogous to those in equations (8)–(10) are:

$$E(\hat{Q} | X, y_{obs}) = \hat{Q}(X, y_{obs}, \mu(\hat{\theta})) + O_p(n^{-1}); \quad (25)$$

$$V(\hat{Q} | X, y_{obs}) = \left(\frac{\partial g(\hat{T})}{\partial T_y} \right)^2 w^T \Sigma(\hat{\theta}) w + \left(\frac{\partial g(\hat{T})}{\partial T_y} \right)^2 D_\mu(\hat{\theta})^T \Gamma D_\mu(\hat{\theta}) + O_p(n^{-3/2}), \quad (26)$$

where w is the vector with elements $w_i, i \in mis$ and now $D_\mu(\theta) = \sum_{i \in mis} w_i \left(\frac{\partial \mu_i(\theta)}{\partial \theta} \right)$; and finally,

$$E(U | X, y_{obs}) = U(X, y_{obs}, \mu(\hat{\theta})) + \left(\frac{\partial g(\hat{T})}{\partial T_y} \right)^2 w^T \Sigma(\hat{\theta}) w + O_p(n^{-3/2}). \quad (27)$$

Thus the analogue to equation (13) is

$$C_1 = 2 \left(\frac{\partial g(\hat{T})}{\partial T_y} \right)^2 w^T \Sigma(\hat{\theta}) w,$$

and C_2 still has the form of equation (14), but with weights incorporated into $D_\mu(\hat{\theta})$.

In 1990 the U.S. Census Bureau conducted its Post-Enumeration Survey (PES), a large-scale effort to measure errors of coverage in the 1990 Decennial Census (Hogan 1993). The PES relied upon two sampling operations, one to detect errors of erroneous omission or undercount (the P-sample) and the other to detect erroneous enumerations or overcount (the E-sample). Estimated rates of undercount and overcount were combined in a dual-system estimator (DSE) of the form

$$DSE = (CEN - I) \left(1 - \frac{E}{N^E} \right) \frac{N^P}{M},$$

where CEN is the raw population count from the census; I is the number of non-data-defined (e.g. imputed) persons in the census count; E is the E-sample weighted estimate of erroneous enumerations; M is the P-sample weighted estimate of correctly enumerated (non-missed) persons; and N^P and N^E are the weighted total number of persons from the P- and E-samples, respectively. The DSE , which represents an estimate of the true population count, was calculated within 1,392 poststrata defined by geographic, demographic and housing characteristics.

Efforts to calculate DSE in 1990 were complicated by missing data. For some P- and E-sample persons, it was operationally not possible to determine whether they had been correctly enumerated in the census. The missing enumeration statuses for these persons were replaced by probabilities as proposed in Schenker (1988), with the probabilities estimated by an elaborate set of logistic-regression models (Belin et al. 1993). This imputation

of probabilities produced a downward bias in the variance estimate for DSE which could be corrected using (26)–(27). Specifically, the naive or mean-imputed variance estimate must be augmented by $C_1 + C_2$, where C_1 represents uncertainty in prediction of the missing binary enumeration statuses given the parameters of the logistic-regression models, and C_2 represents the variability due to the estimation of model parameters. The C_1 term, which can be calculated directly from the mean-imputed survey data, is

$$C_1 = 2 \left[\left(\frac{D\hat{S}E}{\hat{M}} \right)^2 \sum_i (w_i^P)^2 \hat{Y}_i^P (1 - \hat{Y}_i^P) + \left(\frac{D\hat{S}E}{\hat{N}^E - \hat{E}} \right)^2 \sum_i (w_i^E)^2 \hat{Y}_i^E (1 - \hat{Y}_i^E) \right],$$

where \hat{Y}_i^P and w_i^P are the estimated probability of enumeration and sample weight, respectively, for P-sample person i ; \hat{Y}_i^E and w_i^E are the estimated probability of erroneous enumeration and sample weight, respectively, for E-sample person i ; and a hat ($\hat{\cdot}$) above any other quantity indicates that it has been calculated with probabilities imputed for missing values. Details of the derivation of C_1 can be provided upon request. Obtaining an analytic expression for C_2 is impractical due to the complexity of the missing-data model; a rough estimate of this component of uncertainty was obtained by bootstrap resampling of households in the P- and E-samples (Belin et al. 1993).

As part of an extensive PES evaluation program in 1991, the correction factors C_1 and C_2 were calculated from preliminary PES data along with the naive mean-imputed variance estimates for the DSE 's in all 1,392 poststrata. These results were not used in the production of any official Census Bureau statistics, but they provide an interesting overview of the distribution of missing-data uncertainty across the PES poststrata. Figure 2 shows a scatterplot of the two correction factors C_1 and C_2 , where each factor represented as a percentage of the estimated total variance (naive + $C_1 + C_2$). This plot reveals that the missing-data uncertainty is spread rather unevenly across poststrata; in some the correction factors account for as much as 80% of the total estimated variance. Figure 3 shows the correction factors expressed on a square-root scale as percent coefficients of variation ($\sqrt{C_1}/DSE$ and $\sqrt{C_2}/DSE$). In one poststratum, the coefficient of variation associated with

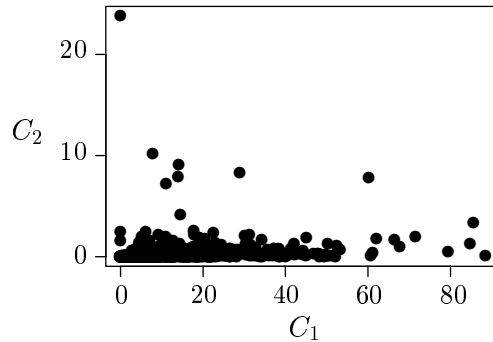


Figure 2: Correction factors C_1 and C_2 for 1,392 PES poststrata as a percentage of the total variance

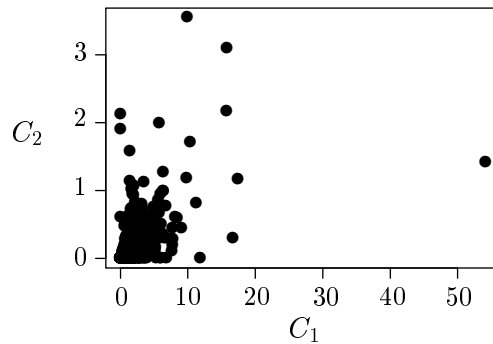


Figure 3: Correction factors C_1 and C_2 for 1,392 PES poststrata expressed as coefficients of variation (%)

C_1 exceeds 50%.

5.5 Sampling for Nonresponse in the Decennial Census

We conclude our examples by describing a potential application of our methods to the problem of estimating the census count after sampling for nonresponse.

Traditionally in the decennial census, all households that do not respond to the initial mail-out/mail-back phase are followed up by phone and/or in person to obtain responses. Consideration is being given to sampling such nonresponding households for follow-up in future censuses. The basic paradigm for sampling for nonresponse follow-up, as described by Bell and Otto (1994), is as follows: (i) a sample of blocks is drawn; (ii) nonresponding households in the sample blocks are followed up; (iii) the data from the sample blocks are

used to fit a model that predicts the number of nonrespondents in each block by race from available data for the block; (iv) the model is used to predict the number of nonrespondents by race in the out-of-sample blocks. The “census count” (estimate) for each block is then the number of respondents plus either (a) the number of nonrespondents obtained from follow-up, for the sample blocks, or (b) the predicted number of nonrespondents, for the out-of-sample blocks.

More formally, for a given block, let $P^{(n)}$ denote the number of nonrespondents (people) in the block, $N^{(n)}$ denote the number of nonresponding housing units in the block, $P^{(r)}$ denote the number of respondents in the block, and $N^{(r)}$ denote the number of responding housing units in the block. Suppose that the census count is desired for a certain area (say, a district office). Then if there were complete data available for the nonresponding housing units, the “complete-data estimate,” that is, the census count, would be

$$\hat{Q} = \sum_{i=1}^n P_i^{(r)} + \sum_{i=1}^n P_i^{(n)},$$

where i indexes blocks and n is the number of blocks in the area. Thus \hat{Q} would be the sum of two sample totals. The complete-data variance would be

$$U = 0$$

since \hat{Q} is a census count.

Let A_i denote a vector of explanatory variables for the i th block, representing demographic summaries, housing unit characteristic summaries, and summaries of census processing data. Bell and Otto (1994) suggested a log-linear Poisson regression model with an offset term to model $P^{(n)}$. Specifically, they assumed that $P_i^{(n)}$ has a Poisson distribution with expected value

$$\mu_i = N_i^{(n)} \alpha_0 (P_i^{(r)} / N_i^{(r)})^{\alpha_1} \exp(A_i^T \beta), \quad (28)$$

where $\theta = (\alpha_0, \alpha_1, \beta)$ are parameters to be estimated. Given a sample S of blocks from the area in question, $P_i^{(n)}$ is observed for $i \in S$ and missing for $i \in S^c$, where S^c denotes the

set of blocks that are not in the sample. Once model (28) has been fitted to the blocks in S , estimated conditional means $\mu_i(\hat{\theta})$ can be imputed for $P_i^{(n)}$, $i \in S^c$. It then follows from (11) that the estimated census count is

$$E(Q | X, y_{obs}) \approx \sum_{i=1}^n P_i^{(r)} + (\sum_{i \in S} P_i^{(n)} + \sum_{i \in S^c} \mu_i(\hat{\theta}));$$

here, y_{obs} corresponds to $P_i^{(n)}$ for $i \in S$, y_{mis} corresponds to $P_i^{(n)}$ for $i \in S^c$, and X corresponds to all of the other data for the blocks.

As under the complete-enumeration assumption of Section 5.3, since $U = 0$, it follows from (7) and (9) that

$$V(Q | X, y_{obs}) = V(\hat{Q} | X, y_{obs}) \approx C_1/2 + C_2,$$

where C_1 and C_2 are given by (13) and (14). Thus, expression (12) does not apply in this situation because the contribution of $E(U | X, y_{obs})$ to C_1 is absent. In the notation of Sections 2–4, $\frac{\partial g(\hat{T})}{\partial T_y} = n$ and $\sigma_i^2(\hat{\theta}) = \mu_i(\hat{\theta})$; thus $C_1 = 2 \sum_{i \in S^c} \mu_i(\hat{\theta})$. For calculating $C_2 = n^2 D_\mu(\hat{\theta})^T \Gamma D_\mu(\hat{\theta})$, the components of $D_\mu(\hat{\theta})$ are $\frac{\partial \mu_i(\hat{\theta})}{\partial \alpha_0} = \mu_i(\hat{\theta})/\alpha_0$, $\frac{\partial \mu_i(\hat{\theta})}{\partial \alpha_1} = \mu_i(\hat{\theta}) \log(P_i^{(r)}/N_i^{(r)})$, and $\frac{\partial \mu_i(\hat{\theta})}{\partial \beta_j} = \mu_i(\hat{\theta}) a_{ij}$, where a_{ij} is the j th component of A_i .

6 Discussion

In this paper, we have developed analytic approximations that can be used to produce valid inferences with imputed conditional means. The methods, which can be viewed as first-order approximations to multiple-imputation procedures, are efficient and relatively simple to implement. They are designed for situations in which the point estimator is a function of linear statistics, and in which either a single variable has missing data or several variables have missing data in a pattern such that the variables are either missing or observed together.

Two extensions of our results would be straightforward. One is the extension to multivariate Y in complex surveys, which would involve combining the results of Sections 4.5

and 5.4. The other is the extension to multidimensional estimands, in which the function g defining the estimands would be a vector and our results would involve matrix generalizations. For more complex patterns of multivariate missing data and more complicated estimators, our methods would either be invalid or difficult to extend, and thus, multiple-imputation methods would be preferable. For example, Schafer (in press) provided procedures for multiple imputation for general patterns of multivariate missing data; these methods can be applied to continuous data, categorical data, or mixtures of the two.

Multiple imputation is not the only paradigm for obtaining accurate inferences from imputed data sets. Some new methods with design-based rather than Bayesian origins are presented by Fay (1996), Rao (1996), and references therein. Moreover, similar results can sometimes be obtained using different paradigms. For example, Särndal (1992) derived “model-assisted” frequentist methods for correcting variance estimates when estimating the population total with a singly-imputed data set. He showed that for a simple random sample, computing the standard variance estimate from a mean-imputed data set yields an estimate that is only r_1^2 as large as it should be, which is the same result that we have derived in Section 5.1 for estimating proportions.

The derivation of our methods has been based on the assumption that the same variables, i.e., X_1, \dots, 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., Section 2. This implicitly reflects the notion that information that is to be used in the analysis of an imputed data set should, in principle, not be omitted from the imputation model; Rubin (1987, Chapter 4) discussed this point. Another assumption implicit in our derivations is that the complete-data estimates, \hat{Q} and U , may be regarded as the posterior mean and variance of Q ; see Section 4.1. While this assumption holds approximately in many applied situations, there are also situations in which it does not hold; for example, it is questionable when design-based complete-data estimators are used for complex surveys, as in Section 5.4. Further research is needed to investigate the performance of our methods and multiple-imputation methods

when this assumption is violated.

Appendix: Derivation of Approximate Moments of \hat{Q} and U

We now sketch proofs of results (8)–(10) of Section 4. Our proofs use standard arguments in Taylor linearization (e.g., Wolter 1985, Chapter 6). Care must be taken to ensure that all 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 (X, y_{obs}) , we are conditioning on (X, y_{obs}) throughout, but for simplicity this is suppressed in the notation that follows. Functions of X and y_{obs} (e.g., $\hat{\theta}$) are considered to be fixed, whereas functions of y_{mis} or θ are considered to be random.

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

$$\begin{aligned} T_y - \hat{T}_y &= n^{-1} \sum_{i \in mis} (y_i - \mu_i(\hat{\theta})) \\ &= n^{-1} \sum_{i \in mis} \epsilon_i + n^{-1} \sum_{i \in mis} (\mu_i(\theta) - \mu_i(\hat{\theta})), \end{aligned} \quad (29)$$

where the $\epsilon_i = y_i - \mu_i(\theta)$ are independent random variables with mean 0 and variance $\sigma_i^2(\theta)$. Thus the first term in (29) is $O_p(n^{-1/2})$. For the second term, note that

$$\begin{aligned} \mu_i(\theta) - \mu_i(\hat{\theta}) &= \left(\frac{\partial \mu_i(\hat{\theta})}{\partial \theta} \right)^T (\theta - \hat{\theta}) + O_p(n^{-1}) \\ &= O_p(n^{-1/2}); \end{aligned} \quad (30)$$

because the number of elements in mis is $O(n)$, the second term in (29) is also $O_p(n^{-1/2})$, and thus

$$T_y = \hat{T}_y + O_p(n^{-1/2}). \quad (31)$$

To establish (8), expand $\hat{Q} = g(T)$ in a Taylor series about $T_y = \hat{T}_y$,

$$\begin{aligned} \hat{Q}(X, y_{obs}, y_{mis}) - \hat{Q}(X, y_{obs}, \mu(\hat{\theta})) &= g(T) - g(\hat{T}) \\ &= \left(\frac{\partial g(\hat{T})}{\partial T_y} \right) (T_y - \hat{T}_y) + O_p(n^{-1}), \end{aligned}$$

and note that $E(T_y) = \hat{T}_y + O_p(n^{-1})$ by (29) and (30).

To establish (9), write

$$V(\hat{Q}) = EV(\hat{Q} | \theta) + VE(\hat{Q} | \theta).$$

Let

$$\tilde{T}_y(\theta) = n^{-1}[\sum_{i \in obs} y_i + \sum_{i \in mis} \mu_i(\theta)],$$

so that $\tilde{T}_y(\hat{\theta}) = \hat{T}_y$; and let $\tilde{T}(\theta) = (T_{X_1}, \dots, T_{X_p}, \tilde{T}_y(\theta))^T$. For any fixed θ , $T_y - \tilde{T}_y(\theta)$ has mean zero and variance $n^{-2} \sum_{i \in mis} \sigma_i^2(\theta)$. Thus the expansion

$$\hat{Q}(X, y_{obs}, y_{mis}) - \hat{Q}(X, y_{obs}, \mu(\theta)) = \left(\frac{\partial g(\tilde{T}(\theta))}{\partial T_y} \right) (T_y - \tilde{T}_y(\theta)) + O_p(n^{-1})$$

implies that

$$V(\hat{Q} | \theta) = \left(\frac{\partial g(\tilde{T}(\theta))}{\partial T_y} \right)^2 n^{-2} \sum_{i \in mis} \sigma_i^2(\theta) + O_p(n^{-3/2}). \quad (32)$$

Notice that the leading term in (32) is of order n^{-1} . Expanding $nV(\hat{Q} | \theta)$ about $\theta = \hat{\theta}$ leads to

$$EV(\hat{Q} | \theta) = \left(\frac{\partial g(\hat{T})}{\partial T_y} \right)^2 n^{-2} \sum_{i \in mis} \sigma_i^2(\hat{\theta}) + O_p(n^{-3/2}). \quad (33)$$

Also, for any fixed θ ,

$$E(\hat{Q} | \theta) = g(\tilde{T}(\theta)) + O_p(n^{-1}).$$

Expanding this expression for $E(\hat{Q} | \theta)$ about $\theta = \hat{\theta}$ gives

$$VE(\hat{Q} | \theta) = \left(\frac{\partial g(\tilde{T}(\hat{\theta}))}{\partial \theta} \right)^T \Gamma \left(\frac{\partial g(\tilde{T}(\hat{\theta}))}{\partial \theta} \right) + O_p(n^{-3/2}),$$

But by the chain rule,

$$\begin{aligned} \frac{\partial g(\tilde{T}(\theta))}{\partial \theta} &= \frac{\partial g(\tilde{T}(\theta))}{\partial T_y} \frac{\partial \tilde{T}_y(\theta)}{\partial \theta} \\ &= \frac{\partial g(\tilde{T}(\theta))}{\partial T_y} n^{-1} \sum_{i \in mis} \left(\frac{\partial \mu_i(\theta)}{\partial \theta} \right), \end{aligned}$$

so

$$VE(\hat{Q} | \theta) = \left(\frac{\partial g(\hat{T})}{\partial T_y} \right) D_\mu(\hat{\theta})^T \Gamma D_\mu(\hat{\theta}) + O_p(n^{-3/2}). \quad (34)$$

Combining (33) and (34) establishes (9).

Finally, to establish (10), notice that

$$S = n^{-1}(Z^T Z - nTT^T) + O_p(n^{-1}),$$

so

$$nU = \left(\frac{\partial g(T)}{\partial T}\right)^T n^{-1}Z^T Z \left(\frac{\partial g(T)}{\partial T}\right) - \left(\frac{\partial g(T)}{\partial T}\right)^T (TT^T) \left(\frac{\partial g(T)}{\partial T}\right) + O_p(n^{-1}), \quad (35)$$

where the leading terms are $O_p(1)$. The second term of (35) depends on y_{mis} only through T_y , so by expansion about $T_y = \hat{T}_y$ (see (31)) the expectation of the second term is

$$- \left(\frac{\partial g(\hat{T})}{\partial T}\right)^T (\hat{T}\hat{T}^T) \left(\frac{\partial g(\hat{T})}{\partial T}\right) + O_p(n^{-1}). \quad (36)$$

The first term of (35), however, depends on y_{mis} through T_y , $X^T y$, and $y^T y$. Let $Z(\theta)$ denote a mean-imputed version of $Z = (X, y)$ with y_{mis} replaced by $\mu(\theta)$. For any θ ,

$$n^{-1}Z^T Z - n^{-1}Z(\hat{\theta})^T Z(\hat{\theta}) = A + B, \quad (37)$$

where

$$A = n^{-1}Z^T Z - n^{-1}Z(\theta)^T Z(\theta)$$

and

$$B = n^{-1}Z(\theta)^T Z(\theta) - n^{-1}Z(\hat{\theta})^T Z(\hat{\theta}).$$

The matrix A has zeros everywhere except the last row and column, whose entries are

$$n^{-1} \sum_{i \in mis} x_{ij}(y_i - \mu_i(\theta)), \quad j = 1, \dots, p \quad (38)$$

and

$$n^{-1} \sum_{i \in mis} (y_i - \mu_i(\theta))^2. \quad (39)$$

The conditional expectations of (38) and (39) given θ are zero and $n^{-1} \sum_{i \in mis} \sigma_i^2(\theta)$, respectively, so $E(A) = EE(A | \theta)$ is a matrix with

$$n^{-1} \sum_{i \in mis} \sigma_i^2(\hat{\theta}) + O_p(n^{-1})$$

in the lower right-hand corner and zeros elsewhere. Similarly, B has zeros except in the last row and column, whose entries are

$$n^{-1} \sum_{i \in mis} x_{ij}(\mu_i(\theta) - \mu_i(\hat{\theta})), \quad j = 1, \dots, p \quad (40)$$

and

$$n^{-1} \sum_{i \in mis} (\mu_i(\theta) - \mu_i(\hat{\theta}))^2. \quad (41)$$

By expansion (30), the expectations of (40) and (41) vanish up to terms of $O_p(n^{-1})$, so $E(B) = O_p(n^{-1})$, and thus (37) implies that

$$E(n^{-1} Z^T Z) = n^{-1} Z(\hat{\theta})^T Z(\hat{\theta}) + E(A) + O_p(n^{-1}). \quad (42)$$

Finally, (42) and the fact that $T - \hat{T} = O_p(n^{-1/2})$ as follows from (31) imply that the expectation of the first term in (35) is

$$\left(\frac{\partial g(\hat{T})}{\partial T} \right)^T n^{-1} Z(\hat{\theta})^T Z(\hat{\theta}) \left(\frac{\partial g(\hat{T})}{\partial T} \right) \quad (43)$$

plus the remainder

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

But because of the pattern of zeros in $E(A)$, this remainder simplifies to

$$\left(\frac{\partial g(\hat{T})}{\partial T_y} \right)^2 n^{-1} \sum_{i \in mis} \sigma_i^2(\hat{\theta}) + O_p(n^{-1/2}). \quad (44)$$

Substituting (43), (44), and (36) into (35) proves the result.

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.
- Bell, W.R., and Otto, M.C. (1994), “Investigation of a Model-Based Approach to Estimation Under Sampling for Nonresponse in the Decennial Census,” paper presented at the 1994 Joint Statistical Meetings, Toronto.
- 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.
- Fay, R.E. (1996), “Alternative Paradigms for the Analysis of Imputed Survey Data,” *Journal of the American Statistical Association*, 91, 490-498.
- Hogan, H. (1993), “The 1990 Post-Enumeration Survey: Operations and Results,” *Journal of the American Statistical Association*, 88, 1047-1060.
- 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, Department of Transportation.
- Little, R.J.A., and Rubin, D.B. (1987), *Statistical Analysis with Missing Data*, New York: Wiley.
- 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.
- Rubin, D.B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

- Rubin, D.B., and Schenker, N. (1991), "Multiple Imputation in Health-Care Databases: An Overview and Some Applications," *Statistics in Medicine* 16 585-598.
- Schafer, J.L. (in press), *Analysis of Incomplete Multivariate Data*, London: Chapman and Hall.
- Schenker, N. (1988), "Handling Missing Data in Coverage Estimation, with Application to the 1986 Test of Adjustment Related Operations," *Survey Methodology*, 14, 87-97.
- Wolter, K.M. (1985), *Introduction to Variance Estimation*, New York: Springer-Verlag.