Survey Methodology, December 1988
Vol. 14, No. 2, pp. 209–221
Statistics Canada

Imputation Strategies for Missing Values in Post-Enumeration Surveys

DONALD B. RUBIN, JOSEPH L. SCHAFER, AND NATHANIEL SCHENKER[1]

ABSTRACT

To estimate census undercount, a post-enumeration survey (PES) is taken, and an attempt is made to find a matching census record for each individual in the PES; the rate of successful matching provides an estimate of census coverage. Undercount estimation is performed within poststrata defined by geographic, demographic, and housing characteristics, X. Portions of X are missing for some individuals due to survey nonresponse; moreover, a match status Y cannot be determined for all individuals. A procedure is needed for imputing the missing values of X and Y. This paper reviews the imputation methods used in the 1986 Test of Adjustment Related Operations (Schenker 1988) and proposes two alternative model-based methods: (1) a maximum-likelihood contingency-table estimation procedure that ignores the missing-data mechanism; and (2) a new Bayesian contingency table estimation procedure that does not ignore the missing-data mechanism. The first method is computationally simpler, but the second is preferred on conceptual and scientific grounds.

KEY WORDS: Bayesian methods; Categorical data; Coverage error; EM algorithm; Multiple imputation; Nonignorable nonresponse; Undercount.

1. INTRODUCTION

The U.S. Bureau of the Census has used a post-enumeration survey (PES) to evaluate coverage error in several past censuses, and it plans to conduct a PES after the 1990 Decennial Census as well. For each individual in the PES, an attempt is made to find a census record (*i.e.*, a match) to determine whether the person was enumerated in the census. The proportion of PES persons who were missed in the census is used as an estimate of the proportion of persons in the population who were missed. A similar matching operation is performed to match a sample of individuals from the census to the PES; this provides an estimate of the census overcount resulting from erroneous (*e.g.*, duplicate or fictitious) enumerations.

The data on matches and erroneous enumerations obtained from the PES are combined to estimate the population size via the dual-system estimator; this capture-recapture type of estimator is discussed in Marks, Seltzer and Krotki (1974), Krotki (1978), Wolter (1986), Diffendal (1988), and Fay, Passell and Robinson (1988, Chapter 5). Dual-system estimates of population size are computed within poststrata defined by geographic, demographic (age, sex, race), and housing (owner/renter, type of housing structure) characteristics.

Two problems of missing data occur in the PES and complicate the estimation process:

1. Geographic, demographic, or housing characteristics may be missing for a person, so it is not known to which poststratum that person belongs.
2. After the processing of the PES, there are some individuals with match status (dichotomous variable indicating matched/not matched to census) or erroneous enumeration status missing. This can occur, for instance, when an incomplete name is obtained in the PES, or when there is difficulty in specifying a Census Day address for someone who moved between Census Day and the PES.

[1] Donald B. Rubin and Joseph L. Schafer, Department of Statistics, Harvard University, Cambridge, MA 02138, USA; Nathaniel Schenker, Division of Biostatistics, UCLA School of Public Health, Los Angeles, CA 90024, USA.

Missing data were a major source of uncertainty in undercount estimation for the 1980 Decennial Census (Freedman and Navidi 1986; Fay, Passell and Robinson 1988, Chapter 6). Improvements in the PES design should reduce the amount of missing data in 1990 (Hogan and Wolter 1988), but a method for dealing with missing data will still be necessary.

The 1986 Test of Adjustment Related Operations (TARO), a recent test of undercount estimation and adjustment (Diffendal 1988; Schenker 1988), used a PES that was similar in design to that planned for 1990. This paper reviews the methods used to handle missing data in TARO (Schenker 1988), identifies potential weaknesses of these methods, and discusses potential alternatives.

Our goal is to indicate issues and problems, and to suggest methods for their solution. The long range plan for research is to carefully evaluate these methods. Although we only discuss imputation for missing PES data when estimating undercount, missing data also occur in the census sample used to estimate overcount. The missing-data problems in estimating overcount, however, are analogous to those in estimating undercount (Schenker 1988), and so our discussion applies to both problems.

In our discussion of alternatives to the TARO procedures, we propose a new method based on a Bayesian model that does not ignore the missing-data mechanism, and thus does not assume that the missing data are missing at random. Nonignorable models for incomplete categorical data are a recent development in the theory of handling missing data; see Fay (1986), Little and Rubin (1987, Section 11.6), and Baker and Laird (1988) for discussions and reviews of the literature. Moreover, the types of missing data that we discuss occur not only in undercount estimation, but in many other situations as well; thus our discussion is relevant to the general problem of handling missing categorical data.

Section 2 discusses the imputation methods used in TARO. In Section 3, alternative methods are described and illustrated using a simple example. Section 4 presents a concluding discussion.

2. IMPUTATION METHODS USED IN TARO

2.1 Description of Methods

For each individual in the PES, let X denote categorical variables for age, sex, race, owner/renter status, and type of housing structure; let Y denote match status (1 = match, 0 = nonmatch); and let Z denote variables indicating whether the PES interview was with a household member or a proxy, and whether the PES person moved between Census Day and the PES. In TARO, the X variables (except type of housing structure) were used in forming poststrata (Diffendal 1988); Z was observed for all PES individuals, but Y and components of X were sometimes missing (Schenker 1988).

Missing values of X and Y were imputed in two stages. (Our description is simplified for ease of presentation; see Schenker (1988) for the precise procedure). First, all missing X values were imputed using a "hot deck" scheme based on observed X variables; that is, imputed values were drawn from the observed distributions of X values. Second, after the missing values of X were filled in, a logistic regression model predicting Y from X and Z was fitted to the cases with Y observed. This logistic regression model was then used to impute probabilities of match for all missing Y values. Probabilities rather than zeros and ones were imputed to (a) increase the precision of estimation, and (b) allow the assessment of variability due to imputation (Schenker 1989).

2.2 Critique of Methods

The TARO imputation methods have many positive features. They are easily understood and use explicit modeling for the imputation of Y. They also condition on much of the observed data, rather than imputing from marginal distributions. Finally, in principle they allow the assessment of uncertainty in undercount estimates due to the missing Y values. The methods have some potential weaknesses, however, which we now describe.

The TARO imputation procedure is an "ignorable" procedure, because it ignores the missing-data mechanism. Ignorable procedures assume that the missing data are missing at random (MAR) (Rubin 1976); that is, they assume that given the observed data, the missingness is independent of the values of the missing items. For example, if X and Z are observed for all people, MAR implies that Y can be imputed using the conditional distribution of Y given X and Z for those individuals having X, Y, and Z observed.

The TARO procedure is actually a special case of an ignorable procedure, because it makes assumptions that are stronger than the general MAR assumption. The TARO procedure treated X and Y asymmetrically; that is, it imputed missing values of Y conditional on all observed data, but it imputed missing X values conditional only on the observed X's, rather than on the observed values of X, Y, and Z. Hence, in addition to the general MAR assumption, the TARO procedure also effectively assumed that, given the observed components of X, the missing components of X are conditionally independent of both Y and Z.

This additional independence assumption may not be realistic; it may be that given the observed X data, there is a residual dependence of values of missing components of X on Y and/or Z. If this is the case, then observed values of Y and Z should be used in the imputation of X. For instance, suppose a PES individual has sex missing, but is found not to match any census record ($Y = 0$) on the basis of observed age, race, and address; and suppose males tend to be undercounted in the census more than females with identical other characteristics. Then knowing that $Y = 0$ provides some evidence that the person in question is more likely to be male than if Y were 1. The most general ignorable imputation procedure would use information provided by Y and Z in imputing missing X values; this is one of the alternative imputation methods, which we discuss in Section 3.4.1.

Another feature of the TARO procedure that may be unrealistic is the ignorability assumption itself. It may be that the missing data are not MAR – *i.e.*, given the observed data, the missingness is not independent of the values of the missing items; if so, then it would be more appropriate to use a nonignorable model for the missing-data mechanism. For instance, consider a group of people with identical values of all variables except race; it may be more difficult to obtain information on race for minorities than nonminorities, and consequently the distribution of race will be different among those missing race and those with race observed. Similarly, even after all X and Z variables are controlled for, it may be that people who were not enumerated in the census are more likely to be missing Y than those who were enumerated in the census. An alternative imputation method based on a general class of nonignorable models is presented in Section 3.4.2.

3. ALTERNATIVE METHODS OF IMPUTATION IN THE PES

3.1 Introduction

Let $X = (X_1, X_2, X_3)$ denote three individual characteristics recorded by the PES (*e.g.*, age, sex, and race). The variables X_1, X_2, and X_3 are assumed to be categorical, taking I, J, and K possible values respectively. We have chosen three variables merely for illustrative purposes

and notational simplicity; all ideas developed here will extend immediately to any number of categorical variables. In practice, these X variables will probably include the demographic, geographic, and housing characteristics used to define poststrata for undercount estimation; they may also include additional PES variables, such as mover status and household member/proxy status, which are not of intrinsic interest but which may be useful for imputation purposes.

We will form IJK different classes of individuals by cross-classifying them according to X_1, X_2, and X_3. These classes may or may not be the same as the poststrata for undercount estimation; in practice the poststrata will probably be coarser than these classes. It is convenient, but not necessary, for these classes to be defined as cross-classifications of all possible values of X_1, X_2, and X_3; more complicated patterns (such as nested ones) are also possible. We will be constructing loglinear models for cross-classified contingency tables, but loglinear models may be based on other patterns as well.

Let Y be the dichotomous variable denoting match status, taking values 1 (matched to census) or 0 (not matched). If there were no missing data, the results of the PES could be summarized in a single four-dimensional contingency table with $I \times J \times K \times 2$ cells, since each individual could be fully classified according to X_1, X_2, X_3, and Y. But those individuals missing one or more variables can be only partially classified according to those variables that are observed. Those having X_1, X_2, X_3, and Y all observed will constitute a four-dimensional table, which we will call the table of *complete cases (CC)*, or the data table for missingness pattern 1 (no variables missing). Those having X_1, X_2, and X_3 observed but Y missing will constitute a three-dimensional *supplementary table* with IJK cells, which we will call the data table for missingness pattern 2. In general, there will be 2^4 such tables corresponding to all possible missingness patterns, one CC table and $2^4 - 1$ supplementary tables.

3.2 Imputation from Reference Tables

In our model-based approach to imputation, we will model the data tables for different missingness patterns as multinomial observations. Corresponding to each missingness pattern, we will define a set of cell probabilities $\Theta^t = \{\Theta^t_{ijkl}\}$, where the superscript t indexes the missingness pattern, $t = 1, \ldots, 2^4$, and the subscripts i, j, k, and l indicate the levels of X_1, X_2, X_3, and Y respectively. Because we will refer to Θ^t when imputing missing values for the t-th data table, we will call Θ^t the reference table for the t-th data table, and $\{\Theta^t\colon t = 1, \ldots, 2^4\}$ the set of reference tables.

Imputation of missing values corresponds to expanding each supplementary data table to make it fully four-dimensional, according to its corresponding reference table. For example, consider the imputation of Y for those individuals missing only Y. This is equivalent to expanding the supplementary data table for missingness pattern 2, by dividing each cell count in this table into two parts, a count of those having $Y = 1$ and a count of those having $Y = 0$, split according to the reference table Θ^2. With known Θ^2 this procedure is straightforward: we first obtain from Θ^2 the conditional distribution of Y given X for this missingness pattern, *i.e.*,

$$P(Y = 1 \mid X_1, X_2, X_3, t = 2) = \frac{\theta^2_{ijk1}}{\theta^2_{ijk0} + \theta^2_{ijk1}}, \tag{1}$$

for $i = 1, \ldots, I$, $j = 1, \ldots, J$, and $k = 1, \ldots, K$. Then, we impute $Y = 1$ for each observation in cell ijk of this table with probability given by the right-hand side of (1); alternatively, we could impute the mean of this distribution, which is just the probability of a match (1). The relative merits of random draw versus mean imputation for the PES will be discussed in Section 3.3.

Note that in the example above, the only information from Θ^2 needed for the imputation is the conditional distribution of Y given X; hence, any value of Θ^2 yielding the same values for (1) leads to the same imputation procedure. For an imputation procedure to be accurate, then, our estimate of Θ^t need not correspond to the joint distribution of Y and X for the t-th missingness pattern; the only requirement is that the conditional distribution of the missing variables given the observed ones derived from our estimate of Θ^t be close to the correct one.

In particular, if the missing-data mechanism is ignorable, one common reference table $\Theta^t = \Theta$, $t = 1, \ldots, 2^4$, provides valid imputations for all missingness patterns, even though the joint distribution of X and Y might vary across missingness patterns. The fact that only one reference table is needed follows from the definition of ignorability, which implies that the conditional distribution of missing values given observed values does not depend on the missingness pattern. The value Θ that provides valid imputations is not Θ_{CC}, the cell probabilities for the joint distribution of X_1, X_2, X_3, and Y underlying the CC table; rather, it is the the joint distribution of X_1, X_2, X_3, and Y marginalized across missingness patterns. Generally, if the missing-data mechanism is nonignorable, we will need to specify a different reference table for each missingness pattern.

In our model-based approach, the two crucial issues to be addressed are: (1) how to estimate the set of reference tables using well-established principles of efficient estimation; and (2) how to perform the imputation once these estimates are obtained. Two methods of estimation will be compared in Section 3.4; in Section 3.3 we briefly discuss various alternatives for imputation.

3.3 Single, Multiple, and Mean Imputation

Once the reference tables have been estimated, distributions for each individual's missing variables given the observed ones have been completely specified. In theory, these distributions could be used to analytically calculate correct point and interval estimates for any quantities of interest. In practice, however, these calculations are usually intractable; some other procedure is needed. Filling in the missing values by imputation is an attractive alternative, because it creates a completed dataset, which can be analyzed by complete-data methods. Little (1986) summarizes the strengths and weaknesses of various imputation methods; we shall only comment on aspects relevant to the PES.

In current practice, each missing value is typically filled in by taking a single random draw from a distribution, thereby producing a simulated complete dataset, which is analyzed in the usual complete-data fashion. Interval estimates derived from this method will be artificially too precise, because they do not reflect the uncertainties of the imputation. One remedy for this, which is coming into use, is multiple imputation (Rubin 1987), in which each missing value is replaced by m random draws from the distribution. With moderate amounts of missing information, $m = 5$ draws are enough to produce efficient point estimates and adequate interval estimates. With rates of missing information that appear likely in the PES (typically 5 - 10 percent or less, judging from TARO), $m = 2$ draws will be perfectly adequate for essentially all purposes. In a large-scale survey like the PES, however, even a small number of multiple imputations may be computationally difficult to handle.

Since the estimates of interest in the PES are the match rates within poststrata, it is probably more important to accurately reflect the variability of imputation for Y than for X; that is, it is probably more important to reflect uncertainty in overall undercount rates than uncertainty in the allocation of undercount to poststrata. Thus it may be possible to obtain adequate results by imputing a single set of X values, and then multiply imputing Y given X. Yet another possibility is to impute a single set of X values, and then impute the probability of match given X. This approach was used in TARO (Schenker 1988); it allows the imputed X's and fractional Y's to be treated like single imputations when estimating undercount rates.

Choosing an acceptable imputation procedure given a set of reference tables is the subject of ongoing research. It is hoped that the TARO approach of imputing a single value of X and then imputing $P(Y = 1 \mid X)$ will prove to be a useful compromise between the accuracy of multiple imputation and the computational ease of single imputation.

3.4 Models and Methods of Estimation

In this section, we present two alternative procedures for modeling the missing data and estimating the reference tables for imputation. The two procedures are the Ignorable Maximum-Likelihood (IML) method and a new Nonignorable Bayesian (NB) method that should be an improvement over IML if the missing data are not MAR.

3.4.1 The Ignorable Maximum-Likelihood Method

As mentioned previously, an ignorable imputation procedure needs to specify only a single reference table and apply it to all missingness patterns. One naive approach is to estimate this common reference table Θ by the cell proportions observed in the CC table. The resulting estimate $\hat{\Theta}_{CC}$ is asymptotically unbiased for Θ if the missing data are missing completely at random (MCAR), that is, if the probability of missingness for each item is completely independent of the data values, observed or missing. If the missing data are merely MAR, and not MCAR, then using $\hat{\Theta}_{CC}$ for imputation introduces biases into the data. Moreover, even when the data are MCAR, $\hat{\Theta}_{CC}$ is not efficient because it does not make use of all of the observed data to estimate Θ.

The IML method makes use of all the data, both in the CC table and in the suplementary tables, to estimate Θ. The estimated value $\hat{\Theta}_{IML}$ is chosen to maximize the likelihood ignoring the missing-data mechanism (Little and Rubin 1987, Section 5.3). In general, there is no closed form expression for $\hat{\Theta}_{IML}$; it must be obtained iteratively, for instance via the EM algorithm (Dempster, Laird and Rubin 1977; Little and Rubin 1987, Section 9.3).

The EM algorithm for contingency tables is easy to implement, and the resulting maximum likelihood estimate $\hat{\Theta}_{IML}$ is both efficient and consistent under the assumption of ignorability; thus this EM procedure for IML is attractive from both computational and theoretical perspectives. When the missing data are not MAR, however, the IML method will generally introduce biases. Since there are good reasons to believe that the missing data in the PES are not missing at random, we propose a new method of estimation that makes a different assumption.

3.4.2 Nonignorable Modeling and Nonuniqueness of the MLE

When the missing data are not MAR, it is no longer valid to ignore the missing-data mechanism; the fact that a data value is missing conveys information about its value. Hence, a model that reflects this dependence must include indicator variables for response, indicating whether data values were observed or missing. Consequently, a nonignorable model will generally estimate a separate reference table for each missingness pattern, or equivalently, an expanded reference table Θ with twice as many dimensions (*i.e.*, with an additional dimension for each missingness indicator).

Let $R = (R_1, R_2, R_3, R_Y)$ be indicator variables for whether X_1, X_2, X_3, and Y are observed, respectively; for example, $R_1 = 1$ if X_1 is observed and $R_1 = 0$ if X_1 is missing. Consider the eight-dimensional contingency table formed by cross-classifying individuals by X, Y, and R, and now let Θ be the eight-dimensional table of cell probabilities for this expanded table.

Each individual in the survey belongs to a cell of the expanded table, but because some data are missing, we only observe certain margins of this table. Because R is fully observed, any margin involving only missingness indicators is fully observed, but a margin involving Y or one of the X's might not be observed. For example, in the cross-section of the table with $R_1 = R_2 = R_3 = 1$ and $R_y = 0$, we can classify individuals by X_1, X_2, and X_3, but not by Y; therefore we observe only the marginal totals obtained by summing across Y.

The number of parameters in the fully saturated model for this table is $2^5IJK - 1$, which is larger than the number of observed sufficient statistics; hence the maximum-likelihood estimate (MLE) for Θ is not uniquely determined. In order to obtain a unique estimate for Θ, one must impose additional structure.

One possible way to obtain a unique MLE is to build a log-linear model for the expanded contingency table, with some of the higher-order interactions set equal to zero (Little 1985; Fay 1986; Little and Rubin 1987, Section 11.6). We might try to set to zero those interactions that are not estimable from the data, but the formalization of this does not always work well in practice. For example, it may at first appear that the R_1 by X_1 interaction is not estimable, because the value of X_1 is never observed when $R_1 = 0$; however, the data may contain information about the R_1 by X_1 interaction indirectly through another variable, one that is observed for some individuals having $R_1 = 1$ and some having $R_1 = 0$. An example of a quantity that is truly inestimable from the data is $P(Y = 1 \mid X_1 = i, X_2 = j, X_3 = k, R_1 = R_2 = R_3 = 1, R_y = 0)$, but this does not correspond to any single interaction term in the log-linear model parameterization. (By "truly inestimable" we mean in Rubin's (1974) sense that the parameter's posterior distribution equals its prior distribution for all priors).

In a dataset with a complicated pattern of missingness, it is not easy to find a set of log-linear terms that, if set to zero, will yield a unique MLE for Θ. The minimum number of terms that must be set to zero to produce uniqueness is $2^5IJK - 1$, the dimension of Θ, minus the number of observed sufficient statistics. Even if such a minimal set can be found, it is usually not unique, and one is faced with the task of deciding which set of terms should be excluded from the model. Rather than attempting to obtain a unique MLE by placing these kinds of prior restrictions on the log-linear model, we will instead use a Bayesian approach involving the use of a prior distribution.

3.4.3 A Nonignorable Bayesian Method

In the Bayesian paradigm, one expresses prior assumptions about the parameters formally through a prior distribution. For our situation, a proper unimodal prior, when combined with the observed-data likelihood, produces a posterior distribution for Θ that can yield a unique estimate; for example, we may take the posterior mode, $\hat{\Theta}_{NB}$, as our estimate of Θ. This method is attractive because it automatically allows precise estimation of those functions of Θ about which the data contain much information, while using the prior to select appropriate values for those quantities that are strictly inestimable from the data. If applied properly, this method will produce a nonignorable model that fits the data as well as any other model – it essentially maximizes the likelihood function, and yet is as consistent as possible with our beliefs about the nature of the missing-data mechanism as expressed in the prior distribution.

Sound scientific practice suggests that we should choose a prior distribution that favors simple structure (*i.e.*, small higher-order interactions) over complicated structure (*i.e.*, large higher-order interactions). If we choose a prior that assigns a low (but nonzero) *a priori* probability to the presence of higher-order interactions in the log-linear model, then we will be making assumptions that are similar in nature to the assumptions of the IML method – that

missing values are not radically different from their observed counterparts in their relationships with other observed variables – although in a smoother, more systematic fashion than the IML method does.

Following the notation of Bishop, Fienberg, and Holland (1975), consider the saturated log-linear model for the eight-way contingency table for R, X, and Y,

$$\begin{aligned}\log\theta_{ijk\ldots p} = \mu &+ \mu_{1(i)} + \mu_{2(j)} + \ldots + \mu_{8(p)} \\ &+ \mu_{12(ij)} + \mu_{13(ik)} + \ldots \\ &+ \mu_{123\ldots 8(ijk\ldots p)},\end{aligned} \tag{2}$$

where $\theta_{ijk\ldots p}$ is the probability that an observation falls in cell $ijk\ldots p$, and the μ's are the one-way, two-way, three-way, and higher-order interactions. We propose the simple family of independent normal prior distributions

$$\begin{aligned}\mu_i &\sim N(0,\sigma^2) \\ \mu_{ij} &\sim N(0,\sigma^2/\tau) \\ \mu_{ijk} &\sim N(0,\sigma^2/\tau^2) \\ &\vdots \\ \mu_{ijk\ldots p} &\sim N(0,\sigma^2/\tau^7),\end{aligned} \tag{3}$$

for some choice of $\sigma^2 > 0$ and $\tau > 1$. This prior distribution pulls the higher-order interactions toward zero, and hence pulls the estimate of Θ toward a more parsimonious or simpler model. We believe that this approach will produce estimates of Θ that are not too different from $\hat{\Theta}_{IML}$ when the missing data are truly MAR, but will be more robust than the IML method under departures from MAR. The only cases when IML will be superior occur when the missing data are MAR and strong higher-order interactions exist among the X's and Y.

Leonard (1975) and Laird (1978) examined log-linear models with normal prior distributions on the μ terms for complete data; our situation is complicated by the fact that only certain margins of the eight-way table are observed. Finding the posterior mode $\hat{\Theta}_{NB}$ under this model is conceptually straightforward; the EM algorithm can be applied to the posterior distribution of Θ, just as to the likelihood function. The E-step remains the same; the M-step, however, poses some computational difficulties. The posterior distribution is nearly a ridge in high-dimensional space; it is very steep in certain directions, but nearly flat in others. The second-derivitive matrix is nearly singular along this ridge; hence Newton-Raphson and other gradient methods for maximization will not work well. Difficulty arises as σ^2 becomes large, because the ridge becomes flat as $\sigma^2 \to \infty$ and a unique mode no longer exists. Difficulty also arises as the number of observations grows, because the posterior becomes very steep in certain directions and thus portions of the second-derivitive matrix become very large. More work is needed to develop effective methods for finding or approximating $\hat{\Theta}_{NB}$.

3.4.4 A Numerical Example

We now present a simple numerical example and compare the results obtained from the IML and NB methods. For simplicity, we will only use a single dichotomous X variable (taking values 0 or 1) and match status Y.

If there were no missingness, the data could be fully cross-classified by X and Y and hence summarized in a single 2×2 contingency table. With four patterns of missingness, however, the data are summarized in a CC table and three supplementary tables (Figure 1).

The CC estimate $\hat{\Theta}_{CC}$ is simply the observed proportions in Table A. The IML estimate $\hat{\Theta}_{IML}$ is found iteratively via the EM algorithm; using $\hat{\Theta}_{CC}$ as the starting value, the algorithm converges in approximately four cycles. The NB estimate $\hat{\Theta}_{NB}$ was found using a prior distribution with $\sigma^2 = 10$ and $\tau = 3$. This means that the one-way terms are *a priori* normally distributed about zero with variance 10, so there is a 95 percent probability that the log-odds for each main effect lies inside the interval $(-4\sqrt{10}, +4\sqrt{10})$. The two-way terms have variance 10/3, the three-ways have variance 10/9, and the four-ways have variance 10/27; this represents a moderate pulling of the higher-order terms toward the origin. (Finding $\hat{\Theta}_{NB}$ for varying values of σ^2 and τ proved difficult, because of the numerical instability of the particular maximization routine applied at each M-step.) The values of $\hat{\Theta}_{IML}$ and $\hat{\Theta}_{NB}$ are given in Figure 2. The expected imputations under these models are given in Figure 3, along with the expected imputations under $\hat{\Theta}_{CC}$ for comparison.

The differences between the imputation methods can be seen most clearly by comparing the expected imputations for Table D. Imputation using $\hat{\Theta}_{CC}$ simply reproduces the proportions observed in Table A. Imputation using $\hat{\Theta}_{IML}$ differs from imputation using $\hat{\Theta}_{CC}$ because Tables B and C, as well as Table A, contribute to the estimation of Θ and hence to the imputation for Table D.

Imputation using $\hat{\Theta}_{NB}$ is fundamentally different from imputation using $\hat{\Theta}_{CC}$ or $\hat{\Theta}_{IML}$ in that it assumes missingness is informative. From Table B, it surmises that missingness of Y is associated with $X = 0$. From Table C, it surmises that missingness of X is associated with $Y = 0$. It then combines this information in a smooth fashion to conclude that a larger proportion of the individuals who have both X and Y missing fall into the $(X = 0, Y = 0)$ category.

4. DISCUSSION

Our work is clearly at an early stage of development. Nevertheless, we feel that it has important potential applications, both specifically to the estimation of undercount using a PES, and generally to contingency table modeling when some data are missing. We conclude with two brief comments: first, on the need for continuing research on these procedures; and second, on the need to judge the relative propriety of models when devising an imputation procedure.

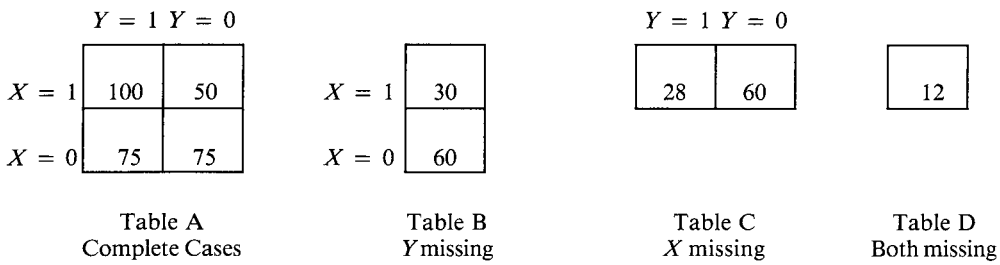

Figure 1. Observed Data

$\hat{\theta}_{IML}$

.279	.174
.239	.308

All missingness patterns

$\hat{\theta}_{NB}$

.333	.167
.250	.250

Pattern A

.193	.140
.305	.361

Pattern B

.165	.262
.153	.419

Pattern C

.211	.224
.222	.333

Pattern D

Figure 2. Reference Tables for Imputation

Observed Data

	$Y = 1$	$Y = 0$
$X = 1$	100	50
$X = 0$	75	75

Table A

$X = 1$	30
$X = 0$	60

Table B

$Y = 1$	$Y = 0$
28	60

Table C

12

Table D

Expected Imputations Under $\hat{\theta}_{CC}$

20	10
30	30

16	24
12	36

4	2
3	3

Expected Imputations Under $\hat{\theta}_{IML}$

18.5	11.5
26.2	33.8

15.1	21.7
12.9	38.3

3.35	2.09
2.86	3.69

Expected Imputations Under $\hat{\theta}_{NB}$

17.4	12.6
27.5	32.5

14.5	23.1
13.5	36.9

2.54	2.70
2.68	4.09

Figure 3. Expected Imputations Under $\hat{\theta}_{CC}$, $\hat{\theta}_{IML}$, $\hat{\theta}_{NB}$

4.1 Continuing Research

Two kinds of research efforts are needed before our NB method can become broadly applicable. First, computationally-oriented research is needed to address the ridge-like posterior distribution. Alternatives to the mode, such as the posterior mean, are worth considering. Furthermore, measures of uncertainty should also be calculated, and considering the odd non-normal shape of the posterior, these may not be simple to summarize or compute. One strategy focuses directly on drawing multiple values of θ from this posterior distribution without explicitly finding the posterior mode or the mean; these draws of θ may be used to multiply impute the missing data.

Related to the issue of measuring uncertainty is the issue of performance in repeated sampling experiments. Although we believe our Bayesian approach is fully appropriate, it is important for broad application to evaluate the operating characteristics of this procedure in the wide range of circumstances to which it might be routinely applied. For example, how well does it work in realistic cases when, unknown to the data analyst, the missing data are MAR?

These topics will be the focus of a major continuing research effort.

4.2 The Need to Judge the Relative Propriety of Models

Considering the fully saturated model for (X, Y, R) with parameter θ, any method of imputation, no matter how illogical, can be viewed as the correct procedure under some model. For example, consider imputation using $\hat{\theta}_{CC}$ as the reference table for all missingness patterns. This posits conditional distributions for the missing data, given the observed data and R, about which there is no information in the observed values. Hence, coupling these distributions with the estimable distributions (the distributions of R and the observed data) implies an estimate for θ, which maximizes the likelihood under the saturated model! It is not a very sensible answer, since it corresponds to the unique MLE under a model in which all sorts of conditional distributions given various missingness patterns R are equal to the conditional distributions given $R = (1, 1, \ldots, 1)$; however, if we consider the likelihood function only, there is no reason to prefer any other maximum-likelihood estimate to this one.

Even stranger methods of imputation, such as "impute all missing values as zero," correspond to particular models with estimated θ's that are MLE's under the saturated model, but they violate good sense. Any sensible attempt to impute missing data values is based on the belief that two individuals with similar values of observed characteristics, and similar missingness patterns, are not radically different in those characteristics that are observed for one and missing for the other. Our NB method formalizes this notion of smoothness by specifying a contingency table model with small higher-order interactions.

Choosing one imputation procedure over another, then, cannot be done on maximum-likelihood-type principles alone, but must involve consideration of the propriety of the underlying prior specifications. This is not really a serious problem; sound statistical practice has always advocated the use of smooth or parsimonious models when less smooth models fit the data equally well. Consider fitting straight lines or polynomial curves through a collection of data points; simpler models are preferable to complicated ones on scientific grounds – the same issues arise in imputation. We believe that the model, given by (2) and (3), underlying our NB method, will be reasonable in many problems, just as linear regression is a reasonable tool in many problems.

ACKNOWLEDGEMENTS

This paper reports research undertaken primarily while Nathaniel Schenker was employed by the Statistical Research Division, Bureau of the Census, Washington, DC 20233, USA. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau. This research was supported in part by Joint Statistical Agreements 87-07 and 88-02 between the U.S. Bureau of the Census and Harvard University, and in part by the U.S. National Science Foundation under grant SES-88-05433, and represents a clarification and revision of Rubin, Schafer, and Schenker (1988). The authors wish to thank the two referees for their very helpful comments.

REFERENCES

BAKER, S.G., and LAIRD, N.M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association*, 83, 62-69.

BISHOP, Y.M.M., FIENBERG, S.E., and HOLLAND, P.W. (1975), *Discrete Multivariate Analysis*, Cambridge: MIT Press.

DEMPSTER, A.P., LAIRD, N.M., and RUBIN, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society,* Series B, 39, 1-38.

DIFFENDAL, G. (1988). The 1986 test of adjustment related operations in Central Los Angeles County. *Survey Methodology*, 14, 71-86.

FAY, R.E. (1986). Causal models for patterns of nonresponse. *Journal of the American Statistical Association*, 81, 354-365.

FAY, R.E., PASSEL, J.S., and ROBINSON, J.G. (1988). *The Coverage of Population in the* 1980 *Census*. 1980 Census of Population and Housing Evaluation and Research Report PHC80-E4, Washington: U.S. Government Printing Office.

FREEDMAN, D.A., and NAVIDI, W.C. (1986). Regression models for adjusting the 1980 Census. *Statistical Science*, 1, 3-39.

HOGAN, H., and WOLTER, K. (1988). Measuring accuracy in a Post Enumeration Survey. *Survey Methodology*, 14, 99-116.

KROTKI, K.J. (1978). *Developments in Dual System Estimation of Population Size and Growth,* Edmonton: The University of Alberta Press.

LAIRD, N.M. (1978). Empirical Bayes methods for two-way contingency tables. *Biometrika*, 65, 1, 581-590.

LEONARD, T. (1975). Bayesian estimation methods for two-way contingency tables. *Journal of the Royal Statistical Society, Series B*, 37, 23-37.

LITTLE, R.J.A. (1985). Nonresponse adjustments in longitudinal surveys: models for categorical data. *Bulletin of the International Statistical Institute*, 15, 1-15.

LITTLE, R.J.A. (1986). Missing data in Census Bureau surveys. *Proceedings of the Second Annual Research Conference*, United States Bureau of the Census, 442-454.

LITTLE, R.J.A., and RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*, New York: Wiley.

MARKS, E.S., SELTZER, W., and KROTKI, K.J. (1974). *Population Growth Estimation.* New York: The Population Council.

RUBIN, D.B. (1974). Characterizing the estimation of parameters in incomplete-data problems. *Journal of the American Statistical Association*, 69, 467-474.

RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 3, 581-592.

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

RUBIN, D.B., SCHAFER, J.L., and SCHENKER, N. (1988). Imputation strategies for estimating the undercount. *Proceedings of the Fourth Annual Research Conference*, United States Bureau of the Census, 151-159.

SCHENKER, N. (1988). Handling missing data in coverage estimation, with application to the 1986 Test of Adjustment Related Operations. *Survey Methodology*, 14, 87-98.

SCHENKER, N. (1989). The use of imputed probabilities for missing binary data. *Proceedings of the Fifth Annual Research Conference*, United States Bureau of the Census (forthcoming).

WOLTER, K.M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, 81, 338-346.