Journal of Official Statistics
Vol. 3, No. 4, 1987, pp. 375–387

Interval Estimation from Multiply-Imputed Data: A Case Study Using Census Agriculture Industry Codes

Donald B. Rubin[1] *and Nathaniel Schenker*[2]

Abstract: We describe the use of multiple imputation based on logistic regression models in a project to calibrate industry and occupation codes for public-use samples from the 1970 and 1980 United States Decennial Censuses. The coverage properties of interval estimates for two estimands are examined in a case study involving multiply-imputed 1980 agriculture industry codes. The use of just a small number of imputations per missing code is shown to yield much more accurate interval estimates than single imputation. Because the problem considered here involves high fractions of missing information, it is important when creating multiple imputations to account for uncertainty in estimating the parameters of the model for nonresponse. We relate these results to the theoretical results of Rubin and Schenker (1986) in simpler situations.

Key words: Bayesian inference; logistic regression; missing data; nonresponse; public-use data; sample surveys.

1. Introduction

Imputation is a standard technique for handling item nonresponse in surveys. Its use is documented in the three volumes from the National Academy of Sciences on incomplete data in surveys (Madow, Nisselson, and Olkin (1983); Madow and Olkin (1983); Madow, Olkin, and Rubin (1983)). Imputation is especially well-suited to a public-use data base created by an organization like the United States Census Bureau for two major reasons: (1) the resulting completed data set can be analyzed using standard complete-data methods of analysis and (2) the creator of the imputations typically knows more about the reasons for nonresponse than the typical user of the data set. However, when standard complete-data methods are applied to a data set completed by imputation, the uncertainty due to using imputed rather than true values for nonrespondents is ignored. The result is inferences that are too sharp. In particular, interval estimates are too short, leading to less than nominal coverage, and p-values are too significant, leading to too many rejections of null hypotheses.

Rubin (1978) proposed multiple imputation as a method of handling nonresponse that allows assessment of uncertainty due to imputation. With multiple imputation, each missing datum is replaced with two or more values representing a distribution of likely values.

[1] Department of Statistics, Harvard University, Cambridge, MA 02138, U.S.A.
[2] Undercount Research Staff, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C. 20233, U.S.A.

This creates two or more completed data sets, each of which can be analyzed using the same complete-data method. These analyses can be combined to reflect both within-imputation and between-imputation variability as described in Section 2. Multiple imputations can also be created under several different models for nonresponse, thereby displaying the sensitivity of inferences to changes in the nonresponse model. See Heitjan and Rubin (1986), and Rubin (1986; 1987a, Ch. 6) for examples of such sensitivity analyses.

Recent theoretical and empirical work on multiple imputation is described in Herzog and Rubin (1983), Rubin and Schenker (1986), Schenker and Welsh (1986), Raghunathan (1987), Rubin (1987a), and Weld (1987). An important practical implication of this work is that multiple imputation, even with only two imputations per missing value, is decisively superior to single imputation with regard to validity of interval estimates and significance levels. The disadvantages of simple ad hoc methods other than single imputation, such as complete-case analysis, which uses only units with no missing values, are documented in Little and Rubin (1987, Ch. 3).

This paper examines coverage properties of interval estimates from multiply-imputed data in a particular Census Bureau data set. The data we use come from a project to calibrate industry and occupation codes for public-use samples across the 1970 and 1980 Decennial Censuses. Having such data calibrated across years is important for many studies, such as longitudinal analyses of employment. The 1970 public-use files (of over a million records each) use 1970 codes, and the 1980 public-use files (of similar size) use 1980 codes. Since the objective is to have 1980 codes on 1970 files, the 1980 codes can be thought of as missing on the 1970 files. There exists a special file of approximately 100 000 1970 records with both 1970 and 1980 codes, and this file can be used to predict 1980 codes for imputation on the 1970 public-use files.

Following a general technical discussion of multiple imputation in Section 2, Section 3 describes the coding calibration problem from which the case study is drawn. The techniques we study for multiply imputing agriculture industry codes are given in Section 4. Section 5 describes the Monte Carlo design of the case study. The results of the study are analyzed in Section 6. A concluding discussion is given in Section 7.

In summary, inferences for two rather different estimands are considered here, and the multiple-imputation intervals perform well for both. These important practical results are anticipated by the detailed theoretical results of Rubin and Schenker (1986) in simpler situations. Futhermore, other evaluations of the multiply-imputed industry codes support the validity of inferences based on these imputed values. Specifically, illustrative work reported in Treiman, Bielby, and Cheng (1987) suggests that inferences based on public-use files with multiply-imputed 1980 codes will be appropriate and more precise than those based on the double-coded sample with true 1980 codes. Work reported in Weld (1987) also supports the validity of p-values derived using multiply-imputed industry codes.

2. Multiple Imputation

Let X denote the covariates in a survey, which are fully observed for all units, and let $Y = (Y_{obs}, Y_{mis})$ be the desired values of the outcome variables, where Y_{obs} is observed and Y_{mis} is missing due to nonresponse. Suppose inferences about some population quantity Q are desired. A straightforward application of Bayes's Theorem shows that the posterior density of Q can be written as

$$h(Q \mid X, Y_{obs}) = \int g(Q \mid X, Y)\, f(Y_{mis} \mid X, Y_{obs})\, dY_{mis}, \qquad (2.1)$$

where f is the posterior density of the missing values and g is the complete-data posterior

density of Q. Expression (2.1) states that the posterior density of Q is the complete-data posterior density of Q averaged over the posterior density of Y_{mis}.

Multiple imputations are simulated draws from the posterior distribution of the missing data, $f(Y_{mis} \mid X, Y_{obs})$. Hence, multiple imputation can be viewed as a simulation device that allows the investigator to approximate the posterior distribution of the quantity of interest by following (2.1).

Suppose that Q is scalar and that if there were no nonresponse, inference for Q would be based on the normal reference distribution

$$Q\text{-}\hat{Q} \sim N(O, U),$$

where $\hat{Q}$ and U are standard complete-data statistics giving the estimate of Q and the variance of $Q\text{-}\hat{Q}$, respectively. In the presence of nonresponse with m sets of imputations of the missing values, Y_{mis}, there are m completed data sets and hence m values of the complete-data statistics $\hat{Q}$ and U, say $\hat{Q}_j$ and $U_j = 1,\ldots,m$.

Rubin and Schenker (1986) and Rubin (1987a) recommended the following t approximation for drawing inferences for Q based on a multiply-imputed data set:

$$Q\text{-}\bar{Q} \sim T^{1/2} t_\nu, \tag{2.2}$$

where

$$\bar{Q} = \sum_j \hat{Q}_j / m$$

is the estimate of Q based on the m imputations, and

$$T = \bar{U} + (1 + m^{-1})\, \hat{B}$$

estimates the total variance of $Q\text{-}\bar{Q}$;

$$\bar{U} = \sum_j U_j / m$$

is the average within-imputation variance of $Q\text{-}\bar{Q}$ and

$$\hat{B} = \sum_j (\hat{Q}_j - \bar{Q})^2 / (m\text{-}1)$$

is the between-imputation variance of $Q\text{-}\bar{Q}$. The degrees of freedom of the t distribution in (2.2) are given by

$$\nu = \{1 + \bar{U} / [(1 + m^{-1})\hat{B}]^2\}\ (m\text{-}1).$$

Thus a nominal 1 - α interval estimate for Q based on (2.2) is given by

$$\bar{Q} \pm t(1\text{-}\alpha/2, \nu)\, T^{1/2}, \tag{2.3}$$

where $t(1\text{-}\alpha/2, \nu)$ is the 1 - $\alpha/2$ quantile of the t_ν distribution.

When single imputation ($m = 1$) is used, imputed values are treated as observed values, and thus there is no estimate of between-imputation variability. The implication is that in practice with $m = 1$, $\hat{B}$ is set equal to zero and $t(1\text{-}\alpha/2, \nu)$ in (2.3) is replaced by the 1 - $\alpha/2$ quantile of the standard normal distribution, which would be appropriate if the imputed values were truly observed values. Hence, interval estimates based on a single-imputation version of a multiple-imputation procedure are always shorter in expectation than the multiple-imputation version, and thus theoretically have lower coverage. Our concern here is with the propriety of the multiple-imputation procedures to be applied on several Census Bureau public-use files.

3. The Industry and Occupation Coding Problem

Every decade, the responses in the United States census concerning industry and occupation are grouped into industry and occupation categories. The industry and occupation clas-

sification schemes change somewhat from census to census because new jobs emerge, old jobs disappear, and new information is obtained on certain jobs. Drastic changes, however, were made for the 1980 census, especially for occupation classifications. Although the changes in classification schemes allow a more accurate representation of employment information in each census, they create difficulties in analyzing employment data over time, since the industry and occupation codes are not directly comparable across decades. This is a serious problem because longitudinal analyses of employment data yield information on topics such as job mobility and the effects of affirmative-action programs, which are of interest to social scientists, labor economists, and the government. The Social Science Research Council and the Census Bureau jointly sponsored the Subcommittee on Comparability of Occupation Measurement (SCOM) to study the comparability problem. A detailed discussion of the problem is given in the SCOM's (1983) report to the sponsors.

Public-use samples containing data on individuals are currently available for the 1900, 1940, 1950, 1960, and 1970 censuses. A 1980 public-use sample is now being created, and the creation of a public-use sample for 1910 is under consideration. The SCOM (1983) has recommended that the public-use samples be modified so that their industry and occupation codes are comparable. The 1980 classification system has been chosen as the standard because it is regarded as more broadly used and superior to earlier systems, and because it was thought that the 1980 system would be closer to future systems.

The SCOM (1983) has suggested two possible methods for achieving comparability. The first method is to assign 1980 codes to each public-use sample by directly coding the verbal responses of the units in the sample concerning industry and occupation (the "alphabetics") using the 1980 coding scheme. The second method is to directly assign 1980 codes to only a subsample of each public-use sample prior to 1980, use the double-coded subsample (the "respondents") to estimate models for predicting 1980 codes from the old codes and covariates, and finally use the models to multiply impute 1980 codes to the units not included in the subsample (the "nonrespondents").

Directly assigning 1980 codes to each public-use sample has the obvious advantage of greater accuracy (assuming no coding errors). The SCOM (1983) has estimated, however, that directly coding the samples would be much more expensive than multiply imputing 1980 codes, especially for the 1960 and 1970 public-use samples, because the alphabetics are not included in the existing samples and would be very costly to retrieve; the 1900, 1940, and 1950 public-use samples contain alphabetics.

It was recommended by the SCOM (1983) that the two methods of achieving comparability just described be evaluated. Since a double-coded sample for 1970 of size 127 125 units already exists at the Census Bureau, a coding calibration project to multiply impute 1980 codes to a 1970 public-use sample with approximately 1.2 million units was undertaken with funding from the National Science Foundation and support from the Census Bureau (Treiman and Rubin (1983)).

Rather than being drawn from the public-use sample, the available double-coded sample is a probability sample of the population taken independently of the public-use sample. The coding calibration project differs from both the case discussed in Section 2 and the cases discussed in the literature in that analyses of the multiply-imputed data typically will be performed using only the nonrespondents' data (i.e., the public-use sample) rather than both the respondents' and nonrespondents' data. A possible implication of this dif-

ference is discussed in Section 7.

The coding calibration problem differs from standard nonresponse problems in two other major ways. First, the nonresponse in the coding calibration problem was created intentionally by probability sampling, so the reason for nonresponse is known. Second, since the double-coded sample is only about one-tenth as large as the public-use sample, the response rate in the coding calibration problem is much lower than in standard surveys. There are, however, many covariates that are observed for all units in the double-coded and public-use samples. Since these covariates contain information about population quantities, the simple fraction of missing outcome values is not an accurate indication of the amount of information lost due to nonresponse for many estimands of interest. Li (1985, Section 4.5) and Rubin (1987a, Sections 3.2 and 3.3) have addressed the general problem of determining the amount of information about a population quantity that is lost when there is nonresponse.

4. Multiply Imputing 1980 Agriculture Industry Codes Using Bayesian Logistic Regression

4.1. *Data set for case study*

The agriculture data set for our case study was extracted from the 1970 double-coded sample, and contains 3 654 units. All of the units have the 1970 industry code for "agricultural production"; the 1980 industry codes are either for "agricultural production, crops" or "agricultural production, livestock." In addition to industry and occupation codes, the data set contains the covariates age, sex, relationship to head of household, race, educational level, employment status, class of worker (for example, government or self-employed), hours worked per week, weeks worked per year, earnings, region of the country, and metropolitan versus non-metropolitan area. We chose the agriculture industry because it has many units and also because one 1970 industry splits into two large 1980 industries; the split between crops and livestock is 57% versus 43%.

4.2. *Model used for imputation*

In the actual Census Bureau problem, multiple imputations of 1980 industry codes (A= crops, B=livestock) are created for units in the 1970 public-use sample with agricultural production as their 1970 classification. These are created in the following steps, where for convenience, we let S_1 denote the double-coded sample and S_0 denote the public-use sample.

First, using the S_1 data, model the probability of having code A versus code B by a Bayesian logistic regression with an intercept and coefficients for the 15 predictor variables listed in Table 1. These variables were chosen by substantive researchers, including Census Bureau staff, as being important either predictively or substantively. The prior distribution placed on the logistic regression coefficient vector β is as follows (Rubin (1983), Rubin and Schenker (1987)). Suppose β has p components and the predictor variables form a contingency table with C cells; for our study and the coding calibration project, $p = 16$ and $C = 2304$. Then p/C prior observations are added to each cell, divided between 1980 codes A and B according to the marginal frequency of these codes in S_1. The logistic regression is fitted to S_1 by maximum likelihood after the prior observations have been added. Using this prior distribution guarantees that the posterior distribution of β is unimodal and easy to maximize for any sample size.

Splitting the prior observations in the same ratio across all cells pulls each logistic regression coefficient, except the intercept, toward zero from its maximum-likelihood estimate with no prior observations added. Adding the same number of prior observations in each cell

reflects the exchangeability of prior judgements across the C cells concerning the true splits. Both of these seem reasonable in the context of thousands of routine applications. It is shown in Rubin and Schenker (1987) that the p/C prior assigns the same average prior variance to the cell logits regardless of the design and model, and that in special cases this prior has desirable frequentist properties.

It is important to realize that the objective of the logistic regression modelling is to allow the representation of uncertainty in the prediction of 1980 codes through the use of multiple imputation. Thus, if sample sizes are small and the predictors are weak, the results of fitting the logistic regression model will involve large standard errors, and consequentially substantial variability in imputed industry codes for the same units across sets of imputations. But this is as it must be: if interval estimates are to be valid, uncertainty must be reflected. Although fitting 16 predictor variables using as few as 20 units, as described in Section 5, may often be a hopeless task for the purpose of finding good point estimates, it is straightforward and necessary for the purpose of providing multiple imputations that adequately reflect inferential uncertainty.

4.3. Creation of imputations from logistic regression output

Multiple imputations of 1980 industry codes for S_0 are created as follows. To impute one set of codes for S_0: (a) a vector β^* is drawn from the posterior distribution of β obtained by fitting the logistic regression (this is discussed further below); (b) for unit i in S_0 ($i = 1,\ldots,n_0$), the unit's covariate values and β^* are used to obtain a probability π_i of the unit having 1980 code A; and (c) for unit $i(i = 1,\ldots, n_0)$, code A is imputed with probability π_i and code B is imputed with probability $1 - \pi_i$. The desired number of imputations is created by independently repeating steps (a)-(c) m times. The actual coding calibration project is creating $m = 5$ imputations per missing datum.

The creation of m imputations requires drawing m values of β^* from the posterior distribution of β. A method of doing this approximately is to let $\beta^* \sim N(\hat{\beta}, H^{-1})$, where $\hat{\beta}$ is the posterior mode of β and H is the negative second derivative matrix of the posterior distribution of β evaluated at $\hat{\beta}$; $\hat{\beta}$ and H are obtained from the standard maximum likelihood computations used to fit the logistic regression. Rubin (1983) has suggested a refinement of the $N(\hat{\beta}, H^{-1})$ approximation that is used by the Census Bureau in difficult cases. The idea is developed further in Tanner and Wong (1987) and Rubin (1987b), but is not studied here.

5. Description of the Case Study

For our case study, we consider the agriculture data set to be the population of units in the agriculture industry and repeatedly draw double-coded and public-use samples, S_1 and S_0, respectively. Let S_1 and S_0 be independent random samples (with replacement) of sizes n_1 and n_0 from the population with the covariates recorded, where the 1980 codes (1=crops, 0=livestock) are known for S_1 but missing for S_0.

In each of 100 independent Monte Carlo trials, samples S_1 and S_0 were drawn, a logistic regression was fitted to S_1, and then multiple imputations were created for S_0. From the 100 multiply-imputed data sets, interval estimates were constructed for two estimands and evaluated by calculating the proportion of times the actual population quantities (that is, the estimands calculated for the entire agriculture data set) were included in the intervals.

The factors considered in the study will now be described.

Estimand

Two estimands are considered: the proportion of units in the population having 1980 industry

code A (crops) and the proportion of black males in the population that have code A. The first was chosen to see how multiple-imputation intervals would perform in the simplest possible inference problem. The second was chosen as a more complicated estimand involving covariates (sex and race). The true values of the estimands are 0.57 = 2083/3654 and 0.84 = 245/291, respectively.

Multiple-imputation intervals were formed on the logit scale and then inverted as follows. For the first estimand, let

$$\hat{p}_j = (X_j + 1/2)/(n_0 + 1)$$

be the jth complete-data estimate (see Section 2) of the population proportion having code A, where X_j is the number of units in S_0 on the jth imputation having code A. Multiple-imputation intervals on the logit scale are formed using the method outlined in Section 2 with

$$\hat{Q}_j = \text{logit}(\hat{p}_j) = \log[\hat{p}_j/(1 - \hat{p}_j)]$$

and

$$U_j = [(n_0 + 1)\,\hat{p}_j\,(1 - \hat{p}_j)]^{-1}.$$

The intervals are transformed to the original scale by the inverse mapping

$$p = \text{logit}^{-1}(Q) = \exp(Q)/[1 + \exp(Q)].$$

The second estimand is treated analogously, with n_0 replaced by the number of black males in S_0, and X_j being the number of black males in S_0 on the jth imputation having code A. For justification of this simple logit-based approach, see Rubin and Schenker (1987).

Nominal level of interval

The nominal levels considered are 50%, 80%, 90%, and 95%.

Method of drawing β^* for imputations

Two methods of drawing β^* are considered. The first is to fix β^* at $\hat{\beta}$ across the m imputations, which does not account for uncertainty due to estimating β. In this respect, it is analogous to the "simple random imputation" method described in Rubin and Schenker (1986), and would be called "improper" by Rubin (1987a). The second method considered is to draw β^* as iid from $N(\hat{\beta}, H^{-1})$. Since this method accounts for uncertainty due to estimating β, it is analogous to the "adjusted" methods of Rubin and Schenker (1986), and would be called "proper" by Rubin (1987a).

Number of imputations per missing value (m)

The values of m considered are 1, 2, and 5. Recall from Section 2 that when $m = 1$, a normal rather than t reference distribution is used.

Size of S_1

The sample sizes n_1 of S_1 considered are 20 and 200. Values as small as $n_1 = 20$ are of interest mainly because they occur often with occupation codes in the coding calibration project. As mentioned earlier, such small sample sizes create no technical problems for our Bayesian logistic regression computer program.

Ratio of size of S_1 to size of S_0

The values of n_1/n_0 considered are 1/5, 1/10, and 1/20. These are appropriate since the 1970 double-coded sample is about one-tenth as large as the 1970 public-use sample.

Summary of design

To summarize, the study can be described as a $2 \times 4 \times 2 \times 3 \times 2 \times 3$ factorial design. The one summary value for each cell is the coverage rate of the interval estimate over repeated samples (i.e., draws of S_1 and S_0). To correlate responses across cells of the design, steps simi-

lar to those described in Rubin and Schenker (1986, Appendix B.1) were used.

6. Results of the Study

Tables 2 and 3 present the coverage rates of the multiple-imputation intervals over 100 Monte Carlo trials for the two inference problems. Although there is a great deal of Monte Carlo variability when only 100 trials are used, computing costs precluded performing enough trials to reduce the standard errors. Nevertheless, many trends can be seen that are consistent with the theoretical results of Rubin and Schenker (1986) for simpler situations, and consequently, the results are practically important.

Multiple imputation versus single imputation

We first compare single imputation with multiple imputation. Table 4 displays the coverage rates for $m = 1$ and $m \geq 2$ averaged over the two estimands and the values of n_1 and n_1/n_0. For $m = 1$, only the method of fixing β^* at $\hat{\beta}$ when creating imputations is considered since the best estimate of β should be used when there is just one imputation. (Tables 2 and 3 show that when $m = 1$, the results for $\beta^* \sim N(\hat{\beta}, H^{-1})$ are worse than for $\beta^* = \hat{\beta}$.) Clearly, the use of multiple imputation leads to coverage rates that are substantially closer to the nominal levels than for single imputation.

When $m \geq 2$, the improper imputation method leads to coverage rates that are too low, whereas the fully Bayesian proper imputation method yields much more accurate results. The difference between the methods is analogous to the difference found in Rubin and Schenker (1986) between the simple random method, which effectively fixes population parameters at point estimates, and the adjusted methods, which draw parameters from approximate posterior distributions. Rubin and Schenker (1986) showed that the improvements due to using adjusted (proper) methods were important for low response rates.

Two versus five imputations for proper imputation

Since the cost of analyzing a multiply-imputed data set increases with the number of imputations m, it is relevant for practical situations to see whether the advantages of multiple imputation can be achieved with just a few imputations. For the simple problem studied in Rubin and Schenker (1986), two or three imputations were sufficient. The improvements in coverage rates due to increasing m were roughly linear in $1/(m\text{-}1)$. Table 5 compares the average coverage rates for $m = 2$ and $m = 5$ for our case study. For nominal levels of 90% and 95%, the use of five imputations yields intervals with better coverage rates than the use of $m = 2$. The average coverage rates for $m = 2$, however, are all within 5% of the nominal levels, and comparison with single imputation (Table 4) shows that most of the gains from multiple imputation are achieved with just $m = 2$, especially for the lower nominal levels.

Differential performance of multiple imputation for the two estimands

Table 6 displays the average coverage rates with proper multiple imputation for the two estimands considered in this study. The average coverage rates for the simpler estimand (the overall population proportion in one industry versus the other) are higher than the average coverage rates for the more complicated estimand (the proportion of black males in one industry versus the other). All of the averages, however, are within 5% of the nominal levels. Thus the proper multiple-imputation procedures perform well for both estimands, especially considering the very high apparent nonresponse rates and the small sample sizes.

7. Discussion

Overall, the results in Section 6 are encouraging, especially considering the complexity of the problem. The results also reflect and support many of the conclusions drawn in Rubin and Schenker (1986) for simpler situations. Moreover, they are consistent with other evaluations of the utility of multiply-imputed industry codes (Treiman, Bielby, and Cheng (1987); Weld (1987)).

The simple case study presented here, in which one 1970 industry code maps into two 1980 industry codes, is relevant to the general coding calibration problem for the following reasons. First, since the occupation coding problem is structurally very similar to the industry coding problem, the results for industry codes should apply to occupation codes as well. Second, although there are many 1970 codes in the coding calibration problem, a separate model is estimated for each 1970 code. Finally, when there are more than two 1980 codes corresponding to a 1970 code, the polytomous imputation is performed as a sequence of dichotomous imputations, as described in Rubin (1983).

As mentioned at the end of Section 3 and as reflected in the design of the case study, multiple-imputation analyses in the coding calibration problem are based only on S_0 rather than both S_1 and S_0. Thus multiple-imputation inferences here are based on the density

$$g(Q \mid X_0, Y_{\text{mis}})\, f(Y_{\text{mis}} \mid X, Y_{\text{obs}})\, dY_{\text{mis}} \quad (7.1)$$

instead of (2.1), where X_0 denotes the observed covariate values in S_0.

Suppose that $m = \infty$ and the distributions of covariates in S_1 and S_0 are approximately the same. Then $g(Q \mid X, Y_{\text{obs}}, Y_{\text{mis}})$ and $g(Q \mid X_0, Y_{\text{mis}})$ should have approximately the same centers. Since $g(Q \mid X, Y_{\text{obs}}, Y_{\text{mis}})$ conditions on more than does $g(Q \mid X_0, Y_{\text{mis}})$, the former should have a lower variance than the latter. Thus multiple-imputation inferences based on (7.1) (that is, on S_0) should be more conservative than the proper multiple-imputation inferences based on (2.1) (that is, on S_1 and S_0). The difference should diminish as n_1/n_0 decreases, since the effect of S_1 on $g(Q \mid X, Y_{\text{obs}}, Y_{\text{mis}})$ decreases.

A simple example from Rubin and Schenker (1986) will demonstrate the heuristic ideas given above. Suppose there are no covariates and an interval estimate of the population mean is desired. Suppose further that the data are normal, the "fully normal" method is used, $m = \infty$, and n is large enough that all "t effects" can be ignored. It follows from Rubin and Schenker (1986, Appendix A) that the nominal 95% multiple-imputation interval based on Y_{obs} and the imputed values of Y_{mis} is

$$\bar{Y}_1 \pm 1.96 \left(s_1^2/n + \frac{n_0}{n n_1} s_1^2\right)^{1/2}, \quad (7.2)$$

where $\bar{Y}_1$ and s_1^2 are the sample mean and variance of Y_{obs}. On the other hand, when the "complete-data" statistics are calculated only from the imputed values of Y_{mis}, the multiple-imputation interval is

$$\bar{Y}_1 \pm 1.96 \left(s_1^2/n_0 + \frac{n}{n_0 n_1} s_1^2\right)^{1/2}. \quad (7.3)$$

Interval (7.3) is clearly wider than interval (7.2), and the two intervals have the same centers. As n_1/n_0 becomes smaller, n_0 approaches n, and (7.3) becomes closer to (7.2).

Table 1. Predictor variables used in logistic regressions for imputing industry codes

Predictor variables	Values
Sex	-1 if male 1 if female
Race	-1 if black 1 if white or other
Sexrace	Sex × Race
Age 1	-1 if $16 \leq$ age ≤ 24 0 if age ≥ 40 1 if $25 \leq$ age ≤ 39
Age 2	-1 if $16 \leq$ age ≤ 24 0 if $25 \leq$ age ≤ 39 or age ≥ 60 1 if $40 \leq$ age ≤ 59
Age 3	-1 if $16 \leq$ age ≤ 24 0 if $25 \leq$ age ≤ 59 1 if age ≥ 60
Sexage	-1 if female and $16 \leq$ age ≤ 39 or male and age ≥ 40 1 otherwise
Class of worker 1	-1 if private industry 0 if self-employed or without pay 1 if government
Class of worker 2	-1 if private industry 0 if government 1 if self-employed or without pay
Metro	-1 if in metropolitan area 1 otherwise
Education	-1 if high school or less 1 if at least one year in college
Hours worked per week	-1 if hours/week ≤ 34 1 otherwise
Weeks worked per year	-1 if weeks/year ≤ 39 1 otherwise
Region 1	-1 if northeast or east north central 0 if west or west north central 1 if south
Region 2	-1 if northeast or east north central 0 if south 1 if west or west north central

Table 2. Coverage rates (in %) of intervals for the proportion of units in the population having 1980 industry code A (crops)

	Nominal level	$\beta^* = \hat{\beta}$			$\beta^* \sim N(\hat{\beta}, H^{-1})$			$\beta^* = \hat{\beta}$			$\beta^* \sim N(\hat{\beta}, H^{-1})$		
		m			m			m			m		
n_1/n_0		1	2	5	1	2	5	1	2	5	1	2	5
		$n_1 = 20$						$n_1 = 200$					
.2	50%	27	40	29	18	56	55	27	37	39	20	53	64
	80%	43	62	59	35	82	85	44	70	62	33	84	87
	90%	53	71	71	41	83	94	56	75	74	40	89	94
	95%	58	79	78	46	86	95	66	83	84	46	94	99
.1	50%	22	25	21	16	51	51	19	29	29	18	53	63
	80%	32	51	45	29	78	82	31	53	47	30	82	87
	90%	39	58	56	32	84	90	41	65	57	34	91	94
	95%	46	68	67	34	89	95	47	73	70	39	92	98
.05	50%	15	22	18	10	50	50	10	19	18	17	57	61
	80%	21	44	33	19	78	79	22	36	37	22	83	87
	90%	31	53	40	22	84	88	30	45	44	24	89	94
	95%	33	58	52	27	90	93	38	56	55	30	93	95

Table 3. Coverage rates (in %) of intervals for the proportion of black males in the population having 1980 industry code A (crops)

	Nominal level	$\beta^* = \hat{\beta}$			$\beta^* \sim N(\hat{\beta}, H^{-1})$			$\beta^* = \hat{\beta}$			$\beta^* \sim N(\hat{\beta}, H^{-1})$		
		m			m			m			m		
n_1/n_0		1	2	5	1	2	5	1	2	5	1	2	5
		$n_1 = 20$						$n_1 = 200$					
.2	50%	18	36	40	22	51	54	31	39	39	12	56	59
	80%	44	71	69	40	76	82	45	69	69	32	78	89
	90%	56	84	86	50	83	93	54	80	77	46	91	94
	95%	67	94	92	57	90	98	60	81	81	56	93	96
.1	50%	18	34	29	16	43	53	22	31	28	13	55	62
	80%	36	60	58	34	70	78	39	59	52	27	79	86
	90%	48	71	66	41	81	84	42	63	62	32	86	93
	95%	52	76	77	42	85	90	48	70	72	39	89	96
.05	50%	17	21	16	12	41	46	15	21	17	12	54	58
	80%	26	44	43	26	68	75	25	41	36	17	77	84
	90%	36	58	55	30	79	84	32	48	46	24	86	93
	95%	39	64	63	34	85	89	39	56	58	27	88	93

Table 4. Comparison of average coverage rates (in %) for single imputation (m=1) and improper and proper multiple imputation (m≥2)

Nominal level	$m=1$	$m\geq 2$	
	$\beta^* = \hat{\beta}$	$\beta^* = \hat{\beta}$	$\beta^* \sim N(\hat{\beta}, H^{-1})$
50%	20	28	54
80%	34	53	81
90%	43	63	88
95%	49	71	92

Table 5. Comparison of average coverage rates (in %) for m=2 and m=5 with proper multiple imputation ($\beta^* \sim N(\hat{\beta}, H^{-1})$)

Nominal level	$m=2$	$m=5$
50%	52	56
80%	78	83
90%	86	91
95%	90	95

Table 6. Average coverage rates (in %) by estimand for proper multiple imputation

Nominal level	Estimand 1	Estimand 2
50%	55	53
80%	83	79
90%	90	87
95%	93	91

Note: Estimand 1 is the proportion of units in the population having 1980 industry code A (crops). Estimand 2 is the proportion of black males in the population having code A.

8. References

Heitjan, D.F. and Rubin, D.B. (1986): Inference for Coarse Data Using Multiple Imputation. Proceedings of the 18th Symposium on the Interface of Computer Science and Statistics. T. Boardman (ed.), American Statistical Association, Washington D.C., pp. 138–143.

Herzog, T.N. and Rubin, D.B. (1983): Using Multiple Imputations to Handle Nonresponse in Surveys. In Incomplete Data in Sample Surveys, Vol. 2: Theory and Bibliographies, W.G. Madow, I. Olkin, and D.B. Rubin (eds.), Academic Press, New York, pp. 209–245.

Li, K.H. (1985): Hypothesis Testing in Multiple Imputation – With Emphasis on Mixed-Up Frequencies in Contingency Tables. Doctoral Dissertation, Department of Statistics, University of Chicago.

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

Madow, W.G., Nisselson, H., and Olkin, I. (1983): Incomplete Data in Sample Surveys, Vol. 1: Report and Case Studies. Academic Press, New York.

Madow, W.G. and Olkin, I. (1983): Incomplete Data in Sample Surveys, Vol. 3: Proceedings of the Symposium. Academic Press, New York.

Madow, W.G., Olkin, I., and Rubin, D.B. (1983): Incomplete Data in Sample Surveys, Vol. 2: Theory and Bibliographies. Academic Press, New York.

Raghunathan, T.E. (1987): Large Sample Significance Levels from Multiply-Imputed Data. Doctoral Dissertation, Department of Statistics, Harvard University.

Rubin, D.B. (1978): Multiple Imputations in Sample Surveys – A Phenomenological Bayesian Approach to Nonresponse. Proceedings of the Survey Research Methods Section of the American Statistical Association, pp. 20–34.

Rubin, D.B. (1983): Progress Report on Project for Multiple Imputation of 1980 Codes. Distributed to the United States Bureau of the Census, the National Science Foundation, and the Social Science Research Council.

Rubin, D.B. (1986): Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations. Journal of Business and Economic Statistics, 4, pp. 87–94.

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

Rubin, D.B. (1987b): Discussion of Tanner and Wong. Journal of the American Statistical Association, 82, pp. 543–546.

Rubin, D.B. and Schenker, N. (1986): Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse. Journal of the American Statistical Association, 81, pp. 366–374.

Rubin, D.B. and Schenker, N. (1987): Logit-Based Interval Estimation for Binomial Data Using the Jeffreys Prior. Sociological Methodology, 17, pp. 131–144.

Schenker, N. and Welsh, A.H. (1986): Asymptotic Results for Multiple Imputation. Technical Report 196, Department of Statistics, University of Chicago. To appear in The Annals of Statistics.

Subcommittee on Comparability of Occupation Measurement (1983): Alternative Methods for Effecting the Comparability of Occupation Measurement over Time. Report to the Social Science Research Council and the U.S. Bureau of the Census.

Tanner, M.A. and Wong, W.H. (1987): The Calculation of Posterior Distributions by Data Augmentation. Journal of the American Statistical Association, 82, pp. 528–540.

Treiman, D.J., Bielby, W., and Cheng, M. (1987): Evaluating a Multiple-Imputation Method for Recalibrating 1970 U.S. Census Detailed Industry Codes to the 1980 Standard. To appear in Sociological Methodology.

Treiman, D.J. and Rubin, D.B. (1983): Multiple Imputation of Categorical Data to Achieve Calibrated Public-Use Samples. Proposal to the National Science Foundation.

Weld, L.H. (1987): Significance Levels from Public-Use Data with Multiply-Imputed Industry Codes. Doctoral Dissertation, Department of Statistics, Harvard University.

Received April 1987
Revised September 1987