EFFECT ON SECONDARY DATA ANALYSIS OF THE USE OF IMPUTED VALUES: THE CASE WHERE MISSING DATA ARE NOT MISSING AT RANDOM

J.H. Jinn, National Chengchi University and J. Sedransk, University of Iowa and Bureau of the Census
J. Sedransk, Bureau of the Census, Statistical Research Division, Washington, D.C. 20233

Key Words: Incomplete data, Nonresponse, Double sampling.

ABSTRACT

A data set having missing observations often is completed by using imputed values, and secondary data analysts typically treat the completed data set as if it has only observed values. The objective of our research is to investigate the effect on the properties of standard statistical techniques of proceeding in this way. We assume that the missing data cannot be regarded as missing at random, and that the secondary data analyst's objective is the standard confidence interval for the population mean. We consider both standard and new imputation methods, some of which assume knowledge of the missing data process.

1. INTRODUCTION

In a sample survey or census, item nonresponse occurs when some but not all of the required information is obtained from a sampled unit. Then the survey organization may either do nothing about the missing values or try to compensate for the item nonresponse by *imputation* for all (or some) of the missing values. Here, imputation means assigning one or more values for each missing response. In large complex sample surveys, imputation is commonly used, and is, *a priori*, an appealing general purpose strategy. (For an excellent introduction to the practical aspects of imputation, see Sande 1982.) In their excellent review paper, Kalton and Kasprzyk (1982) describe the desirable features of imputation: "First... it aims to reduce biases in survey estimates from missing data... Second, by assigning values at the microlevel and thus allowing analyses to be conducted as if the data set were complete, imputation makes analyses easier to conduct and results easier to present. Complex algorithms to estimate population parameters in the presence of missing data (e.g., the EM algorithm of Dempster, Laird and Rubin 1977) are not required. Third, the results obtained from different analyses are bound to be consistent, a feature which need not apply with an incomplete data set." The alternative to imputation, that is, having an incomplete data set, leaves to the secondary data analyst the task of compensating for the missing data. This will be a formidable problem if the likelihood that a datum is missing is related to the values of the variables under study. In such circumstances standard routines in computer packages generally will fail to make appropriate adjustments for the missing data. Conversely, the survey organization is familiar with the survey process (including characteristics of the sampled units who are item nonrespondents) and some of the reasons why the data are missing. It can provide estimates of the missing values (i.e., imputations) that are consistent with a postulated model for the missing data process, leading in many cases to acceptable estimates of the missing values.

On the other hand, Kalton and Kasprzyk point out that imputation "does not necessarily lead to estimates that are less biased than those obtained from the incomplete data set; indeed, the biases could be much greater, depending on the imputation procedure and the form of estimate. There is also the risk that analysts may treat the completed data set as if all the data were actual responses, thereby overstating the precision of the survey estimates."

It is our experience that the overwhelming majority of secondary data analysts proceed as if the completed data set contains only observed responses, and it is our belief that they will continue to do so. The objective of our research is to try to discern the effect on the properties of standard statistical techniques of proceeding in this way. That is, we view imputation methodology from the perspective of the secondary data analyst who does not take cognizance of the presence of imputed values in the data set. We mimic their behavior by evaluating properties of statistics obtained from standard, "canned" computer programs using a data set having both observed and imputed values.

While imputation has been used for a long time, systematic research on properties of imputation methods is recent. Early published papers whose objectives were to determine analytical properties of estimators containing both observed and imputed data include Bailar and Bailar (1978), Bailar, Bailey and Corby (1978), Ernst (1978, 1980) and Platek, Singh and Tremblay (1978). Other, more recent, references of interest are Madow, Nisselson and Olkin (1983), Madow, Olkin and Rubin (1983), Madow and Olkin (1983) and Little (1986).

All of the research cited above investigates properties of univariate descriptive statistics such as means and totals and assumes that the missing data are missing at random (MAR). Santos (1981a,b) extends this work by studying bivariate statistics such as the sample covariance, correlation and regression coefficients in the MAR case. However, only the biases of these statistics are considered. Also Herzog and Rubin (1983) study the effects of using two simple imputation methods on the usual confidence interval for a population mean. Given the paucity of results concerning the effects of alternative imputation methods on the properties of analytical statistics, we initiated research to study properties of statistics derived from a typical analysis of a simple linear regression model. Jinn and Sedransk (1987, 1989) considered the common imputation methods, and emphasized properties of the usual confidence interval for the slope. Now, the imputation method chosen by the survey organization must be

appropriate for a wide variety of statistical applications (e.g., multiple regression, CART), and any value of the parameters in the model being analyzed by the secondary data analyst. Unfortunately, Jinn and Sedransk (1987, 1989) found that when the missing data are MAR no one of the imputation methods is sufficiently reliable that it can be recommended for general use. While the general tenor of this finding might have been anticipated, the extent and apparent generality of the deficiencies were not. Given this conclusion, we have turned to the case where imputation has the greatest promise of successful application; i.e., when the missing data cannot be regarded as MAR.

In this paper we study the effects on secondary data analysis of using data sets containing imputed values when the missing data cannot be regarded as missing at random. Given the paucity of literature, we start with a simple case; i.e., the usual confidence interval for a population mean. However, our specification of the missing data process is general. As in Chiu and Sedransk (1986), we consider a finite population of N elements, and assume that the random variable of interest, Y, can take on the values $Y_{(1)} < Y_{(2)} < \cdots < Y_{(D)}$, which are specified before sampling. Clearly, this specification is also appropriate when Y is a categorized continuous variable and $Y_{(i)}$ is a measure of central tendency for the i-th category. Let P_i denote the unknown proportion of elements in the population with $Y = Y_{(i)}$. Also, θ_i denotes the proportion of elements in the population with $Y = Y_{(i)}$ who would, if sampled, be nonrespondents; $1-\theta_i$ denotes the corresponding proportion of respondents. Table 1 summarizes the notation; e.g., $P_i(1-\theta_i)$ is the proportion of elements with $Y = Y_{(i)}$ and who would respond if sampled.

Table 1. Notation for the Nonresponse Process

Category	1	i	D
Value of Y	$Y_{(1)}$	$Y_{(i)}$	$Y_{(D)}$
Respondents	$P_1(1-\theta_1)$	$P_i(1-\theta_i)$	$P_D(1-\theta_D)$
Nonrespondents	$P_1\theta_1$	$P_i\theta_i$	$P_D\theta_D$

Assuming, for simplicity, a random sample of size n selected with replacement, r respondents are observed, with r_i having $Y = Y_{(i)}$ ($\sum_{i=1}^{D} r_i = r$). Among the n-r nonrespondents, denote by t_i the unknown number of elements having $Y = Y_{(i)}$ ($\sum t_i = n-r$). Corresponding to each imputation method there are estimates, $\{\hat{t}_i: i = 1,\cdots,D\}$, of $\{t_i: i = 1,\cdots,D\}$. Then the imputed values for the n-r nonrespondents are $\hat{t}_i$ repetitions of $Y_{(i)}$ and the completed data set consists of $(r_i+\hat{t}_i)$ repetitions of $Y_{(i)}$ with sample mean $n^{-1}\sum_{i=1}^{D} Y_{(i)}(r_i+\hat{t}_i)$. When Y is a categorized continuous variable, it may be preferable to add random residuals to the $Y_{(i)}$. One may regard this specification as applying to a single post-stratum or "adjustment cell."

This report is organized as follows. Our analytical procedure is presented in Section 2. The imputation methods are described in Section 3, and compared in Section 4. Section 5 has some concluding remarks.

2. ANALYTICAL PROCEDURE

Given the completed data set as $(r_i+\hat{t}_i)$ repetitions of $Y_{(i)}$ $(i = 1,\cdots,D)$, it is assumed that the secondary data analyst's confidence interval for the population mean, μ, is

$$\bar{y}_c \pm z_{\alpha/2}\hat{\sigma}_c/\sqrt{n} \tag{2.1}$$

where n is the overall sample size, $z_{\alpha/2}$ is the $100\{1 - (\alpha/2)\}$ percentage point of the N(0,1) distribution,

$$n\bar{y}_c = \sum_{i=1}^{D} Y_{(i)}(r_i+\hat{t}_i) \tag{2.2}$$

and

$$(n-1)\hat{\sigma}_c^2 = \sum_{i=1}^{D} (r_i+\hat{t}_i)(Y_{(i)}-\bar{y}_c)^2. \tag{2.3}$$

The most desirable way to ascertain the properties of (2.1) would be to determine whether

$$z_c = \sqrt{n}(\bar{y}_c-\mu)/\hat{\sigma}_c \tag{2.4}$$

is well-approximated by a normal distribution with mean 0 and variance 1. Since it is difficult to consider (2.4) directly we proceed in stages: (a) If the bias of $\bar{y}_c$ is large then the approximation will not be satisfactory; (b) For the interval in (2.1) to be an approximately $100(1-\alpha)\%$ confidence interval for μ, $\hat{\sigma}_c^2/n$ should estimate $\mathrm{Var}(\bar{y}_c)$; i.e.,

$$Q^2 = E\{\hat{\sigma}_c^2/n\ \mathrm{Var}(\bar{y}_c)\} \tag{2.5}$$

should not differ much from 1. We also consider the bias of $\hat{\sigma}_c^2$ and the variance of $\bar{y}_c$. Note that the probability distribution in Table 1 is used to derive these expected values.

3. IMPUTATION METHODS

As an alternative to the confidence interval in (2.1), one might delete the imputed values and use as the interval for μ

$$\bar{y}_r \pm z_{\alpha/2}\hat{\sigma}_r/\sqrt{r} \tag{3.1}$$

where

$$r\bar{y}_r = \sum_{i=1}^{D} Y_{(i)}r_i \text{ and } (r-1)\hat{\sigma}_r^2 = \sum_{i=1}^{D} r_i(Y_{(i)}-\bar{y}_r)^2.$$

A simple, standard imputation method is random imputation. Given a sample of size n with m = n-r missing values, a random sample of size m is taken with replacement from the r observed values. The selected respondents are the "donors" and their values are randomly assigned to the nonrespondents.

The remaining imputation methods are ones that assume complete or partial knowledge about the missing data process defined in Table 1. If under such circumstances one cannot develop an imputation procedure that, when used in (2.1), leads to a confidence interval with satisfactory properties, then the task of accommodating uncritical secondary data analysts seems to be insuperable for the following two reasons: (a) such a secondary analyst *will* use the confidence interval in (2.1) and (b) we shall be using in the imputations knowledge of the missing data process. Even the methods that assume complete knowledge about the missing data process will have a practical application when a careful methodological study of nonrespondents has been carried out for a population similar to the one of interest. (See Section 4 for an example.) In this paper, we have also considered the effects of misspecifying the $\{\theta_i\}$.

First, define

$$\phi_i = P_i\theta_i/\sum_j P_j\theta_j, \tag{3.2}$$

the probability that $Y = Y_{(i)}$ given that the individual is a nonrespondent, and

$$\pi_i = P_i(1-\theta_i)/\sum_j P_j(1-\theta_j), \tag{3.3}$$

the probability that $Y = Y_{(i)}$ given that the individual is a respondent.

Then, assuming ϕ_i is known, one may make the obvious assignment

$$\hat{t}_i = (n-r)\phi_i, \quad i = 1,\cdots,D \tag{3.4}$$

so that $\bar{y}_c = n^{-1}\sum_{i=1}^{D} Y_{(i)}\{r_i + (n-r)\phi_i\}$. An alternative to this mean imputation method is to select $\{\hat{t}_i\}$ as a random sample of size (n-r) from the (point) multinomial distribution with probabilities $\{\phi_i: i = 1,\cdots,D\}$.

Another possibility is to shift the observed distribution $\{(Y_{(i)},r_i): i = 1,\cdots,D\}$ to approximate the distribution for the nonrespondents. First, select the $\{\hat{t}_i\}$ as described for the random imputation method; i.e., select (n-r) observations from the point multinomial distribution with probabilities $\{(r_i/r): i = 1,\cdots,D\}$. Then, re-scale $Y_{(i)}$ to $Y'_{(i)} = Y_{(i)}\phi_i/\pi_i$ using (3.2) and (3.3). Finally, the (n-r) imputed values are $\hat{t}_i$ repetitions of $Y'_{(i)}$ so that $\bar{y}_c = n^{-1}\left\{\sum_{i=1}^{D} Y_{(i)}r_i + \sum Y'_{(i)}\hat{t}_i\right\}$. One motivation for this choice of $Y'_{(i)}$ is that $E(Y'_{(i)}\,|\,R) = E(Y_{(i)}\,|\,NR)$ where R and NR denote "respondent" and "nonrespondent." An alternative is to use $Y''_{(i)} = Y_{(i)}\phi_i r/r_i$ rather than $Y'_{(i)}$. It is our experience that the difference in the distributions for (a) respondents and (b) nonrespondents typically cannot be modelled in a simple way, for example as a location-scale change.

One may use double sampling to estimate the $\{\phi_i\}$. Assuming a subsample of b nonrespondents, let b_i denote the number in the sample with $Y = Y_{(i)}$. Then one may take

$$\hat{t}_i = (n-r)b_i/b, \tag{3.5}$$

a method related to the one in (3.4). Alternatively, a modification using random imputation can be employed: Select $\{\hat{t}_i\}$ as a random sample from the point multinomial distribution with probabilities $\{(b_i/b): i = 1,\cdots,D\}$ and use as the imputed values t_i repetitions of $Y_{(i)}$. We have also considered modified versions of the two double sampling methods just described. In each case

$$n\bar{y}_c = \sum_{i=1}^{D} Y_{(i)}(r_i + b_i + \hat{t}_i). \tag{3.6}$$

For the modified mean imputation method, $\hat{t}_i = (n-r-b)b_i/b$ while for the random imputation method, $\{\hat{t}_i\}$ is a random sample of size (n-r-b) from the point multinomial distribution with probabilities $\{(b_i/b): i = 1,\cdots,D\}$. Finally, one may modify the scale-change method to use $Y'''_{(i)} = Y_{(i)}b_i r/br_i$ rather than $Y''_{(i)}$.

In Table 2 we list and name the imputation methods that we have investigated.

For each of the fifteen imputation methods listed in Table 2 we obtained analytical expressions for $E(\bar{y}_c)$, $Var(\bar{y}_c)$ and $E(\hat{\sigma}_c^2)$; they are given in Appendix 1. The only approximations that were used were first order approximations for $E(r^{-1})$ and $E(r/r_i)$. Obtaining these expressions was often complex, involving several stages of expectations: (a) subsampling of the nonrespondents, (b) random selection of the $\{\hat{t}_i\}$ given the $\{r_i\}$, and (c) sampling the $\{r_i\}$ according to the specification in Table 1.

Before proceeding to a formal comparison of the alternative imputation methods, we indicate that there may be gains from using a data set completed using a "good" imputation method rather than by using only the observed data. For the latter case it is clear that $E(\bar{y}_r) = E(Y\,|\,\text{Respondent})$ and $E(\hat{\sigma}_r^2) = Var(Y\,|\,\text{Respondent})$. (See (3.1).) Thus, using the notation in Table 1,

$$\text{bias}(\bar{y}_r) = \sum_{i=1}^{D} Y_{(i)}\left\{\frac{P_i(1-\theta_i)}{\Sigma P_j(1-\theta_j)} - P_i\right\} \tag{3.7}$$

which is independent of the sample size. When (3.7) is not negligible, the confidence interval in (3.1) will not be satisfactory. Alternatively, assume $\hat{\phi}_1,\cdots,\hat{\phi}_D$, unbiased estimators of $\{\phi_i: i = 1,\cdots,D\}$, and consider

$$\bar{y}_c = n^{-1}\left\{\sum_{i=1}^{D} Y_{(i)}(r_i + \hat{t}_i)\right\}$$

with $\hat{t}_i = (n-r)\hat{\phi}_i$ where $\hat{\phi}_i$ may depend on $\{r_i: i = 1,\cdots,D\}$. Then it is easily shown that $E(\bar{y}_c) = \sum_{i=1}^{D} Y_{(i)}P_i = E(Y)$, as desired. Also consider $\hat{\sigma}_c^2$ as defined in (2.3). After some algebraic manipulation, and using a first order Taylor Series approximation, it can be shown that

$$E(\hat{\sigma}_c^2) \doteq Var(Y) - \frac{1}{n}(1-\frac{r}{n})Var(Y \mid Respondent).$$

Thus, for large n, $E(\hat{\sigma}_c^2) \doteq Var(Y)$ as one would like to have.

4. EVALUATION

To complement our analytical comparisons of the fifteen imputation methods presented in Section 3, we have also carried out a numerical investigation using several data sets from a methodological study at the U.S. Census Bureau following the completion of the 1982 Census of Wholesale Trade. For each Standard Industrial Classification (SIC) and for each of several variables the values of Y for the nonrespondents were obtained from administrative records. Thus we have a frequency distribution of Y for each of the "respondent" and "nonrespondent" subpopulations and also know the overall response rate. These data are then used to obtain the P_i and θ_i in Table 1, and, finally, values for $E(\bar{y}_c)$, $Var(\bar{y}_c)$, $E(\hat{\sigma}_c^2)$ and Q^2 (see (2.2), (2.3), (2.5)).

For our investigation we use two variables, 1982 total sales and 1982 annual payroll, and four SIC's (5052, Wholesale distribution of coal and other minerals and ores; 5171, Petroleum bulk stations and terminals; 5172, Petroleum and petroleum products wholesalers — except bulk stations and terminals; 5181, Wholesale distribution of beer and ale). Eliminating the open-ended classes, there are nineteen classes for sales and eleven classes for payroll. To illustrate, we present in Tables 3 and 4 the values of the P_i, θ_i and $Y_{(i)}$ for Sales 5052 and Payroll 5052: Here, $Y_{(i)}$ is the mid-point of Y in the i-th category.

As a summary of the results from this numerical study we present in Table 5 (Sales 5052) and Table 6 (Payroll 5052) the values of $B = bias(\bar{y}_c)/\{Var(\bar{y}_c)\}^{1/2}$, $P = \{E(\hat{\sigma}_c^2) - \sigma^2\}/\sigma^2$ and Q^2 for imputation methods N, R, MC, RC, SC, SC-RE, SD, DM and DR. Letting $f = b/(n-r)$ denote the fraction of nonrespondents who are subsampled (methods SD, DM, DR), we consider (n,f) = (200, 0.5), (200, 0.1), (100, 0.5), (100, 0.3) and (40, 0.5).

We have also studied the sensitivity of methods MC, RC and SC-RE to misspecification of the $\{\phi_i\}$ by considering as alternatives MI, RI and SI-RE (Table 2). The results of this investigation are reported later in this section.

First, consider the use of only the observed values (method N) leading to the confidence interval in (3.1). It is clear from Tables 5 and 6 that the value of $B = bias(\bar{y}_r)/\{Var(\bar{y}_r)\}^{1/2}$ may be very large; among the eight cases (4 SIC's, 2 variables), the largest values of B are in Table 5 while the smallest are in Table 6. Since $bias(\bar{y}_r)$ does not depend on n (see (A.1)), B decreases as n decreases.

The standard random imputation method, R, is *inferior* to method N described above: First, the bias of $\bar{y}_c$ is exactly the same for the two methods (see (A.1), (A.4)). However, the value of Q^2 is very much smaller for R than for N. First, using (A.5) and (A.6) and ignoring terms of $0(n^{-1})$, it is easily shown that for R, $Q^2 < 1$ while for N, $Q^2 = 1$. Also, for these examples (4 SIC's, 2 variables and all values of n) the value of Q^2 for method R ranged from 0.44 to 0.67. The bias of $\hat{\sigma}_c^2$ is slightly smaller for R, but the values of P are very similar for the two methods. Thus, as one might expect, using an imputation procedure motivated by the situation where missing data are MAR is inappropriate when this condition does not hold.

The remaining methods assume complete or partial knowledge of the missing data process. Method MC is a mean imputation method with $\hat{t}_i = (n-r)\phi_i$ (see (3.4) and (3.2)). First, $\bar{y}_c$ is an unbiased estimator of E(Y). Second, ignoring terms of $0(n^{-1})$, $E(\hat{\sigma}_c^2) = Var(Y)$ (see (A.9)). For these examples there is little bias in $\hat{\sigma}_c^2$.

Third, ignoring terms of $0(n^{-1})$, $Q^2 > 1$: Letting R and NR denote, respectively, the respondent and nonrespondent subpopulations, and using

$$Var(Y) = Var(Y \mid R)Pr(R) + Var(Y \mid NR)Pr(NR) + Pr(R)Pr(NR)\{E(Y \mid R) - E(Y \mid NR)\}^2 \quad (4.1)$$

and (A.8),

$$Var(\bar{y}_c) = n^{-1}\{Var(Y) - Var(Y \mid NR)Pr(NR)\}. \quad (4.2)$$

Finally, using (4.1) and (4.2) and ignoring terms of $0(n^{-1})$,

$$Q^2 = Var(Y)\{Var(Y) - Var(Y \mid NR)Pr(NR)\}^{-1}.$$

In these examples, $1.17 \le Q^2 \le 1.57$ where we consider all SIC's, both variables and all values of n. Note that the value of Q^2 varies little with n (Tables 5, 6).

Assuming that good estimates of the $\{\phi_i\}$ are available, the random imputation version of MC, RC, is excellent. It can be shown that for RC, $E(\bar{y}_c) = E(Y)$, $E(\hat{\sigma}_c^2) = Var(Y)$ and $Q^2 = 1$. Later, we give several examples to illustrate the effect of misspecification of the $\{\theta_i\}$ on RC and on the other methods that assume a knowledge of the nonresponse process.

Using a subsample of b nonrespondents to estimate the ϕ_i is a practical alternative. Assuming that all subsampled nonrespondents provide the required data, the four methods, DM, DM-B, DR and DR-B, share the attributes that the bias of $\bar{y}_c$ is zero and the bias of $\hat{\sigma}_c^2$ is $0(n^{-1})$. (See (A.34), (A.36), (A.40) and (A.42).) In these examples the bias of $\hat{\sigma}_c^2$ is negligible. After considerable algebraic manipulation it can be shown that

$1 > Q^2_{DM} = Q^2_{DM-B} \geq Q^2_{DR-B} \geq Q^2_{DR}$.

Thus, mean imputation is preferred. Also, for each of the four methods Q^2 decreases as f decreases, and when f is small, Q^2 is small. Considering all SIC's, both variables and all values of n, if the subsampling fraction, f, is 0.5, $.73 \leq Q^2_{DM} \leq .87$, while if f = 0.3, $.54 \leq Q^2_{DM} \leq .74$ and if f = 0.1, $.23 \leq Q^2_{DM} \leq .43$. (Note that for a given value of f, the value of Q^2_{DM} varies little as n varies.) Thus, double sampling with mean imputation is an adequate method if the rate of subsampled nonrespondents is rather large.

Finally, consider the scale-change methods (SC, SC-RE and SD). First, note that for SC-RE, $\bar{y}_c$ is an unbiased estimator of E(Y) and $Q^2 = 1$ (see (A.22) and (A.23)). Method SC-RE is superior to SC: First, it requires less prior knowledge than SC (π_i must be specified for SC, but is estimated for SC-RE). Second, although $E(\bar{y}_c) = E(Y)$ for each method, and there is little difference in $|P|$ between the methods, $Q^2 = 1$ for SC-RE, but $Q^2 < 1$ for SC (see Tables 5 and 6). Method SD provides an alternative to SC-RE. First, $E(\bar{y}_c) = E(Y)$. Also, as expected, the value of Q^2_{SD} is satisfactory if f is large. For these examples, if f = 0.5, $0.69 \leq Q^2_{SD} \leq 0.94$ while if f = 0.3, $0.61 \leq Q^2_{SD} \leq 0.87$ and if f = 0.1, $0.37 \leq Q^2_{SD} \leq 0.76$. Unfortunately, the values of P are sometimes quite large, ranging up to 5.0 for an example with n = 200, f = 0.1.

We have also studied the sensitivity of methods MC, RC and SC-RE to misspecification of the ϕ_i by considering as alternatives methods MI, RI and SI-RE (see Table 2). The results for SI are not presented because its counterpart, SC, has been shown to be unsatisfactory. We consider a simpler case by using only five classes for Y rather than the nineteen classes in the numerical investigation reported above (see Table 3). We take as the "correct" specification the values of the $Y_{(i)}$, P_i and θ_i given in the first three columns of Table 7 (Sales 5181). "Incorrect" specifications are constructed by assuming that the P_i are correct but the θ_i are not, and by taking the overall probability of nonresponse (0.23) to be approximately the same for each (mis)specification in Table 7. These alternative (mis)specifications of the θ_i range from mild (number 1) to major (number 4) departures from the correct specification of the θ_i.

The results for methods MC and MI are summarized in Table 8, those for RC and RI in Table 9, and those for SC-RE and SI-RE in Table 10. We proceed as in Tables 5 and 6 by presenting the values of B, P and Q^2 for n = 100, 40 and 20. Note that for this investigation estimators associated with methods MI, RI and SI-RE use an incorrect set of θ_i, but properties of these estimators (e.g., bias) are evaluated using the correct P_i and θ_i (columns 2 and 3 of Table 7).

First, note that for SI-RE, $Q^2 = 1$ (see (A.32) and (A.33)). Second, the values of Q^2 for MI and RI are substantially larger than 1 (Tables 8, 9). While the values of B are similar for the three methods, the values of P tend to be larger for RI than for the other methods. Thus, we tentatively conclude that SI-RE is preferable to MI and MI to RI.

The bias of $\bar{y}_c$ is the same for each of the methods:

$$\text{bias}(\bar{y}_c) = \left[\sum_{i=1}^{D} P_i\theta_i\right]\left\{\sum_{i=1}^{D} Y_{(i)}(\phi_i^*-\phi_i)\right\}. \qquad (4.3)$$

Since (4.3) does not depend on n, $|B| = |\{\text{bias}(\bar{y}_c)\}\{\text{Var}(\bar{y}_c)\}^{-1/2}|$ increases as n increases (Tables 8, 9, 10). Moreover, (4.3) will be large whenever large values of $|\phi_i^*-\phi_i|$ are associated with large values of $Y_{(i)}$. The first incorrect set of θ_i represents a mild departure from the *correct* set (see Table 7) and, as expected, the values of $|B|$ for this case are relatively small. (See the columns labelled B, MI, B, RI, and B, SI-RE.) Conversely, the fourth set of incorrect θ_i represents a substantial departure from the correct set and, as expected, the values of $|B|$ are large. Both the second and third sets of incorrect θ_i have the same pattern (θ_i decreasing with i) and similar values of θ_i, but differ in that for the fifth category, $0.16-\theta_i$ is much larger for incorrect set 3 than for set 2. Thus, as anticipated from (4.3), the values of $|B|$ are much larger for set 3 than for set 2.

We conclude that moderate misspecification of the values of the θ_i will not lead to extreme values of B unless there are very large discrepancies between the correct and incorrect θ_i corresponding to the categories having the largest values of $Y_{(i)}$. Smaller values of n reduce the value of $|B|$.

We summarize the results of this section as follows:

1. In many circumstances, using only the observed values, method N, is inappropriate because the bias of $\bar{y}_c$ is unacceptably large.
2. The standard random imputation method, R, is even less satisfactory than N. This illustrates the point that using an imputation procedure motivated by the situation where missing data are missing at random is inappropriate when this condition does not hold.
3. Using a subsample of the nonrespondents to estimate the $\{\phi_i\}$ is an effective method if the fraction subsampled is large. Mean imputation, method DM, is the preferred method.
4. Methods MC, RC and SC-RE are all potentially useful, but the latter appears to be preferable because $Q^2_{SC-RE} = Q^2_{SI-RE} = 1$. Moderate misspecification of the values of the $\{\theta_i\}$ should not lead to extreme values of B except when there are very large discrepancies between the correct and incorrect $\{\theta_i\}$ corresponding to the categories having the largest values of $Y_{(i)}$.

5. DISCUSSION

In this paper, we have shown that, in a simple situation, there are potentially effective methods for imputation when the missing data cannot be regarded as missing at random. There are two reasons that we regard this to be important. First, the results of Jinn and Sedransk (1987, 1989) provide a strong indication that for secondary data analysis imputation is not worthwhile when the missing data are missing at random or, presumably, nearly so. Second, the

current work suggests the possible value of good imputation methods in the case of greatest importance for effective secondary data analysis. The provision of an appropriately completed data set provides an automated way to "correct" an observed data set that cannot be regarded as a probability sample from a well-defined population.

Much additional research is needed to: (a) confirm these findings for more complicated statistical analyses, and (b) develop better imputation methods. Two areas of current research effort are extensions to (a) situations where there is true item nonresponse, and (b) linear regression analysis.

APPENDIX I

Algebraic Expressions for $E(\bar{y}_c)$, $Var(\bar{y}_c)$ and $E(\hat{\sigma}_c^2)$ for Imputation Methods in Table 2

We present in this appendix algebraic expressions for $E(\bar{y}_c)$, $Var(\bar{y}_c)$ and $E(\hat{\sigma}_c^2)$ where $\bar{y}_c$ and $\hat{\sigma}_c^2$ are defined in (2.2) and (2.3). The imputation methods are described in Section 3 and outlined in Table 2. Note that the only approximations used are first order approximations for $E(r^{-1})$ and $E\{(r_i/r)^{-1}\}$.

First, if only the observed values of Y are used (see (3.1)),

$$E(\bar{y}_r) = \sum_{i=1}^{D} \pi_i Y_{(i)} = E(Y \mid R) \tag{A.1}$$

where R denotes "respondent,"

$$Var(\bar{y}_r) \doteq \sum_{i=1}^{D} \pi_i \{Y_{(i)} - E(Y \mid R)\}^2 / n \left(1 - \sum_{i=1}^{D} P_i \theta_i\right)$$

$$= Var(Y \mid R)/n \left(1 - \sum_{i=1}^{D} P_i \theta_i\right), \tag{A.2}$$

and

$$E(\hat{\sigma}_r^2) = Var(Y \mid R). \tag{A.3}$$

For method R,

$$E(\bar{y}_c) = \sum_{i=1}^{D} \pi_i Y_{(i)} = E(Y \mid R), \tag{A.4}$$

$$Var(\bar{y}_c) \doteq n^{-1} Var(Y \mid R) \left\{ \sum_{i=1}^{D} P_i \theta_i + \left(1 - \sum_{i=1}^{D} P_i \theta_i\right)^{-1} \right\}, \tag{A.5}$$

and

$$E(\hat{\sigma}_c^2) \doteq (n-1)^{-1} Var(Y \mid R) \left\{ n - \sum_{i=1}^{D} P_i \theta_i - (1 - \sum P_i \theta_i)^{-1} \right\}. \tag{A.6}$$

For MC it can be shown that

$$E(\bar{y}_c) = \sum_{i=1}^{D} P_i Y_{(i)} = E(Y), \tag{A.7}$$

$$Var(\bar{y}_c) = n^{-1} \left[\left(1 - \sum_{i=1}^{D} P_i \theta_i\right) Var(Y \mid R) \right. \tag{A.8}$$

$$\left. + \left(\sum_{i=1}^{D} P_i \theta_i \right) \left(1 - \sum_{i=1}^{D} P_i \theta_i\right)^{-1} \left\{ \sum_{i=1}^{D} Y_{(i)} (\phi_i - P_i) \right\}^2 \right],$$

and

$$E(\hat{\sigma}_c^2) = (n-1)^{-1} \left[n\, Var(Y) - \left(1 - \sum_{i=1}^{D} P_i \theta i\right) Var(Y \mid R) \right.$$

$$\left. - \left(1 - \sum_{i=1}^{D} P_i \theta_i\right) \left(\sum_{i=1}^{D} P_i \theta_i \right) \left\{ \sum Y_{(i)} (\pi_i - \phi_i) \right\}^2 \right]. \tag{A.9}$$

For MI it can be shown that

$$E(\bar{y}_c) = E(Y) + \left(\sum_{i=1}^{D} P_i \theta_i \right) \sum_{i=1}^{D} Y_{(i)} (\phi_i^* - \phi_i) \tag{A.10}$$

where the ϕ_i^* are constants estimating the ϕ_i,

$$Var(\bar{y}_c) = n^{-1} \left[\left(1 - \sum_{i=1}^{D} P_i \theta_i\right) Var(Y \mid R) \right.$$

$$\left. + \left(1 - \sum_{i=1}^{D} P_i \theta_i\right) \left(\sum_{i=1}^{D} P_i \theta_i \right) \left\{ \sum_{i=1}^{D} Y_{(i)} (\pi_i - \phi_i^*) \right\}^2 \right], \tag{A.11}$$

and

$$E(\hat{\sigma}_c^2) = (n-1)^{-1} \left[n \sum_{i=1}^{D} u_i \left(Y_{(i)} - \sum_{j=1}^{D} Y_{(j)} u_j \right)^2 \right.$$

$$- \left(1 - \sum_{i=1}^{D} P_i \theta_i\right) \left(\sum_{i=1}^{D} P_i \theta_i \right) \left\{ \sum_{i=1}^{D} Y_{(i)} (\pi_i - \phi_i^*) \right\}^2$$

$$\left. - \left(1 - \sum_{i=1}^{D} P_i \theta_i\right) Var(Y \mid R) \right] \tag{A.12}$$

where

$$u_i = P_i(1-\theta_i) + \phi_i^* \sum_{j=1}^{D} P_j \theta_j = \sum_{j=1}^{D} P_j \theta_j \left\{ \phi_i^* \right.$$

$$\left. + \pi_i \left(1 - \sum_{j=1}^{D} P_j \theta_j\right) \left(\sum_{j=1}^{D} P_j \theta_j \right)^{-1} \right\}.$$

For RC,

$$E(\bar{y}_c) = E(Y), \tag{A.13}$$

$$Var(\bar{y}_c) = n^{-1} Var(Y), \tag{A.14}$$

and

$$E(\hat{\sigma}_c^2) = Var(Y). \tag{A.15}$$

For RI,

$$E(\bar{y}_c) = E(Y) + \left(\sum_{i=1}^{D} P_i\theta_i\right) \sum_{i=1}^{D} Y_{(i)}(\phi_i^*-\phi_i), \qquad (A.16)$$

$$\begin{aligned} Var(\bar{y}_c) = n^{-1} \Bigg[& \left(\sum_{i=1}^{D} P_i\theta_i\right) \left\{\sum_{i=1}^{D} Y_{(i)}^2\phi_i^* - \left(\sum_{i=1}^{D} Y_{(i)}\phi_i^*\right)^2\right\} \\ & + \left\{\sum_{i=1}^{D} P_i(1-\theta_i)\right\} Var(Y \mid R) \\ & + \left\{\sum_{i=1}^{D} P_i\theta_i\right\} \left\{\sum_{i=1}^{D} P_i(1-\theta_i)\right\} \left\{\sum_{i=1}^{D} Y_{(i)}(\pi_i-\phi_i^*)\right\}^2 \Bigg], \end{aligned} \qquad (A.17)$$

and

$$\begin{aligned} E(\hat{\sigma}_c^2) = (n-1)^{-1} \Bigg[& n\left\{\sum_{i=1}^{D} Y_{(i)}^2 P_i^* - \left(\sum_{i=1}^{D} Y_{(i)} P_i^*\right)^2\right\} \\ & - \left(\sum_{i=1}^{D} P_i\theta_i\right) \left\{\sum_{i=1}^{D} Y_{(i)}^2\phi_i^* - \left(\sum_{i=1}^{D} Y_{(i)}\phi_i^*\right)^2\right\} \\ & - \left\{\sum_{i=1}^{D} P_i(1-\theta_i)\right\} \left\{\sum_{i=1}^{D} Y_{(i)}^2\pi_i - \left(\sum_{i=1}^{D} Y_{(i)}\pi_i\right)^2\right\} \\ & - \left(\sum_{i=1}^{D} P_i\theta_i\right) \left\{\sum_{i=1}^{D} P_i(1-\theta_i)\right\} \left\{\sum_{i=1}^{D} Y_{(i)}(\pi_i-\phi_i^*)\right\}^2 \Bigg] \end{aligned} \qquad (A.18)$$

where

$$P_i^* = \pi_i\left\{\sum_{j=1}^{D} P_j(1-\theta_j)\right\} + \phi_i^*\left(\sum_{j=1}^{D} P_j\theta_j\right).$$

For SC,

$$E(\bar{y}_c) = E(Y), \qquad (A.19)$$

$$\begin{aligned} Var(\bar{y}_c) \doteq n^{-1} \Bigg[& \left(1-\sum_{i=1}^{D} P_i\theta_i\right) \left(\sum_{i=1}^{D} P_i\theta_i\right) \left\{\sum_{i=1}^{D} Y_{(i)}(\pi_i-\phi_i)\right\}^2 \\ & + \left(1-\sum_{i=1}^{D} P_i\theta_i\right) Var(Y \mid R) \\ & + (n-1)n^{-1} \left(\sum_{i=1}^{D} P_i\theta_i\right) \left(1-\sum_{i=1}^{D} P_i\theta_i\right)^{-1} \left\{\sum_{i=1}^{D} Y_{(i)}^2(\phi_i)^2(\pi_i)^{-1}\right. \\ & \left. - \left(\sum_{i=1}^{D} Y_{(i)}\phi_i\right)^2\right\} \\ & + 2\left(\sum_{i=1}^{D} P_i\theta_i\right) \left\{\sum_{i=1}^{D} Y_{(i)}^2\phi_i - \sum_{i=1}^{D}\sum_{j=1}^{D} Y_{(i)}Y_{(j)}\pi_i\phi_j\right\}\Bigg], \end{aligned} \qquad (A.20)$$

and

$$\begin{aligned} E(\hat{\sigma}_c^2) \doteq Var(Y) & - (n+1)(n-1)^{-1} \left(\sum_{i=1}^{D} P_i\theta_i\right) \left\{\sum_{i=1}^{D} Y_{(i)}^2\phi_i\right\} \\ & + 2(n-1)^{-1} \left(\sum_{i=1}^{D} P_i\theta_i\right) \sum_{i=1}^{D}\sum_{j=1}^{D} Y_{(i)}Y_{(j)}\pi_i\phi_j \\ & + \left\{n(n-1)^{-1} \sum_{i=1}^{D} P_i\theta_i - n^{-1}\left(\sum_{i=1}^{D} P_i\theta_i\right) \left(1-\sum_{i=1}^{D} P_i\theta_i\right)^{-1}\right\} \cdot \\ & \left\{\sum_{i=1}^{D} Y_{(i)}^2(\phi_i)^2(\pi_i)^{-1}\right\} \\ & + n^{-1}(n-1)^{-1} \left(\sum_{i=1}^{D} P_i\theta_i\right) \left(n\sum_{i=1}^{D} P_i\theta_i - 1\right) \left(1-\sum_{i=1}^{D} P_i\theta_i\right)^{-1} \cdot \\ & \left(\sum_{i=1}^{D} Y_{(i)}\phi_i\right)^2. \end{aligned} \qquad (A.21)$$

For SC-RE,

$$E(\bar{y}_c) = E(Y), \qquad (A.22)$$

$$\begin{aligned} Var(\bar{y}_c) \doteq n^{-1} \Bigg[& Var(Y) - \left(\sum_{i=1}^{D} P_i\theta_i\right) \left\{\sum_{i=1}^{D} Y_{(i)}^2\phi_i \right. \\ & \left. - \sum_{i=1}^{D} Y_{(i)}^2(\phi_i)^2(\pi_i)^{-1}\right\}\Bigg], \end{aligned} \qquad (A.23)$$

and

$$\begin{aligned} E(\hat{\sigma}_c^2) \doteq Var(Y) & - \left(\sum_{i=1}^{D} P_i\theta_i\right) \left\{\sum_{i=1}^{D} Y_{(i)}^2\phi_i \right. \\ & \left. - \sum_{i=1}^{D} Y_{(i)}^2(\phi_i)^2(\pi_i)^{-1}\right\}. \end{aligned} \qquad (A.24)$$

For SD,

$$E(\bar{y}_c) = E(Y), \qquad (A.25)$$

$$\begin{aligned} Var(\bar{y}_c) \doteq n^{-1} \Bigg[& Var(Y) - \sum_{i=1}^{D} Y_{(i)}^2 P_i\theta_i \\ & + f^{-1} \left\{\sum_{i=1}^{D} P_i\theta_i - n^{-1}\right\} \{Var(Y \mid NR)\} \\ & + (nf)^{-1} \left\{\sum_{i=1}^{D} Y_{(i)}^2\phi_i(1-\phi_i)(\pi_i)^{-1}\right\} \\ & + \left(\sum_{i=1}^{D} P_i\theta_i\right) \sum_{i=1}^{D} Y_{(i)}^2(\phi_i)^2(\pi_i)^{-1}\Bigg] \end{aligned} \qquad (A.26)$$

where $f = b(n-r)^{-1}$, the fraction of nonrespondents to be subsampled, and

$$E(\hat{\sigma}_c^2) \doteq Var(Y) - \sum_{i=1}^{D} Y_{(i)}^2 P_i\theta_i$$

$$- \left[n \sum_{i=1}^{D} P_i\theta_i - 1\right] \{n(n-1)f\}^{-1} Var(Y \mid NR)$$

$$+ n^{-1} \left\{ f^{-1} \sum_{i=1}^{D} Y_{(i)}^2 \phi_i (1-\phi_i)(\pi_i)^{-1} \right.$$

$$\left. + n \left[\sum_{i=1}^{D} P_i\theta_i\right] \sum_{i=1}^{D} Y_{(i)}^2 (\phi_i)^2 (\pi_i)^{-1} \right\}. \qquad (A.27)$$

For SI,

$$E(\bar{y}_c) = E(Y) + \left[\sum_{i=1}^{D} P_i\theta_i\right] \left[\sum_{i=1}^{D} Y_{(i)} \{\pi_i \phi_i^* (\pi_i^*)^{-1} - \phi_i\}\right], \qquad (A.28)$$

$$Var(\bar{y}_c) \doteq n^{-1} \left[\left[1 - \sum_{i=1}^{D} P_i\theta_i\right] Var(Y \mid R) \right.$$

$$+ (n-1)n^{-1} \left[\sum_{i=1}^{D} P_i\theta_i\right] \left\{1 - \sum_{i=1}^{D} P_i\theta_i\right\}^{-1} Var(Y^* \mid R)$$

$$+ 2 \left[\sum_{i=1}^{D} P_i\theta_i\right] \left\{ \sum_{i=1}^{D} Y_{(i)} Y_{(i)}^* \pi_i (1-\pi_i) - \sum_{i \neq j} Y_{(i)} Y_{(i)}^* \pi_i \pi_j \right\}$$

$$\left. + \left[1 - \sum_{i=1}^{D} P_i\theta_i\right] \left[\sum_{i=1}^{D} P_i\theta_i\right] \left\{ \sum_{i=1}^{D} \pi_i (Y_{(i)} - Y_{(i)}^*) \right\}^2 \right] \qquad (A.29)$$

where $Y_{(i)}^* = Y_{(i)} \phi_i^* (\pi_i^*)^{-1}$, and

$$E(\hat{\sigma}_c^2) \doteq \left[1 - \sum_{i=1}^{D} P_i\theta_i\right] \sum_{i=1}^{D} Y_{(i)}^2 \pi_i$$

$$+ \left\{ n(n-1)^{-1} \sum_{i=1}^{D} P_i\theta_i \right.$$

$$- n^{-1} \left[\sum_{i=1}^{D} P_i\theta_i\right] \left[1 - \sum_{i=1}^{D} P_i\theta_i\right]^{-1} \left\} \left\{ \sum_{i=1}^{D} (Y_{(i)}^*)^2 \pi_i \right\} \right.$$

$$+ \left\{ n^{-1} \left[\sum_{i=1}^{D} P_i\theta_i\right] \left[1 - \sum_{i=1}^{D} P_i\theta_i\right]^{-1} \right.$$

$$\left. - (n-1)^{-1} \left[\sum_{i=1}^{D} P_i\theta_i\right] \right\} \left[\sum_{i=1}^{D} Y_{(i)}^* \pi_i\right]^2$$

$$- \left\{ \left[1 - \sum_{i=1}^{D} P_i\theta_i\right] \sum_{i=1}^{D} Y_{(i)} \pi_i + \left[\sum_{i=1}^{D} P_i\theta_i\right] \sum_{i=1}^{D} Y_{(i)}^* \pi_i \right\}^2$$

$$- 2(n-1)^{-1} \left[\sum_{i=1}^{D} P_i\theta_i\right] \left\{ \sum_{i=1}^{D} Y_{(i)} Y_{(i)}^* \pi_i (1-\pi_i) \right.$$

$$\left. - \sum_{i \neq j} Y_{(i)} Y_{(i)}^* \pi_i \pi_j \right\}. \qquad (A.30)$$

For SI-RE,

$$E(\bar{y}_c) = E(Y) - \left[\sum_{i=1}^{D} P_i\theta_i\right] \left\{ \sum_{i=1}^{D} (\phi_i - \phi_i^*) Y_{(i)} \right\}, \qquad (A.31)$$

$$Var(\bar{y}_c) \doteq n^{-1} \left[\left[\sum_{i=1}^{D} P_i\theta_i\right] \left\{ \sum_{i=1}^{D} Y_{(i)}^2 (\phi_i^*)^2 (\pi_i)^{-1} \right.\right.$$

$$\left. - \left[\sum_{i=1}^{D} Y_{(i)} \phi_i^*\right]^2 \right\} + \left[1 - \sum_{i=1}^{D} P_i\theta_i\right] Var(Y \mid R)$$

$$\left. + \left[1 - \sum_{i=1}^{D} P_i\theta_i\right] \left[\sum_{i=1}^{D} P_i\theta_i\right] \left\{ \sum_{i=1}^{D} Y_{(i)} (\pi_i - \phi_i^*) \right\}^2 \right], \qquad (A.32)$$

and

$$E(\hat{\sigma}_c^2) \doteq \left[1 - \sum_{i=1}^{D} P_i\theta_i\right] \sum_{i=1}^{D} Y_{(i)}^2 \pi_i$$

$$+ \left[\sum_{i=1}^{D} P_i\theta_i\right] \sum_{i=1}^{D} Y_{(i)}^2 (\phi_i^*)^2 (\pi_i)^{-1}$$

$$- \left\{ \left[1 - \sum_{i=1}^{D} P_i\theta_i\right] \left[\sum_{i=1}^{D} Y_{(i)} \pi_i\right] \right.$$

$$\left. + \left[\sum_{i=1}^{D} P_i\theta_i\right] \left[\sum_{i=1}^{D} Y_{(i)} \phi_i^*\right] \right\}^2. \qquad (A.33)$$

For DM,

$$E(\bar{y}_c) = E(Y), \qquad (A.34)$$

$$Var(\bar{y}_c) = n^{-1} \left[\left[1 - \sum_{i=1}^{D} P_i\theta_i\right] Var(Y \mid R) \right.$$

$$+ f^{-1} \left[\sum_{i=1}^{D} P_i\theta_i\right] Var(Y \mid NR)$$

$$\left. + \left[1 - \sum_{i=1}^{D} P_i\theta_i\right] \left[\sum_{i=1}^{D} P_i\theta_i\right] \left\{ \sum_{i=1}^{D} (\pi_i - \phi_i) Y_{(i)} \right\}^2 \right], \qquad (A.35)$$

and

$$E(\hat{\sigma}_c^2) = Var(Y) - (n-1)^{-1}(1-f) f^{-1} \left[\sum_{i=1}^{D} P_i\theta_i\right] Var(Y \mid NR). \qquad (A.36)$$

For DM-B,

$$E(\bar{y}_c) = E(Y), \tag{A.37}$$

$$\mathrm{Var}(\bar{y}_c) = n^{-1}\left[\mathrm{Var}(Y) + (1-f)f^{-1}\left(\sum_{i=1}^{D} P_i\theta_i\right)\mathrm{Var}(Y\mid NR)\right], \tag{A.38}$$

and

$$E(\hat{\sigma}_c^2) = \mathrm{Var}(Y) - \{(n-1)f\}^{-1}(1-f)\left(\sum_{i=1}^{D} P_i\theta_i\right)\mathrm{Var}(Y\mid NR). \tag{A.39}$$

For DR,

$$E(\bar{y}_c) = E(Y), \tag{A.40}$$

$$\mathrm{Var}(\bar{y}_c) = n^{-1}\left[\left(1-\sum_{i=1}^{D} P_i\theta_i\right)\mathrm{Var}(Y\mid R) + \left\{\left(\sum_{i=1}^{D} P_i\theta_i\right)(1+f^{-1}) - (fn)^{-1}\right\}\mathrm{Var}(Y\mid NR) + \left(1-\sum_{i=1}^{D} P_i\theta_i\right)\left(\sum_{i=1}^{D} P_i\theta_i\right)\left\{\sum_{i=1}^{D}(\pi_i-\phi_i)Y_{(i)}\right\}^2\right], \tag{A.41}$$

and

$$E(\hat{\sigma}_c^2) = \mathrm{Var}(Y) - \{(n-1)f\}^{-1}\left\{\sum_{i=1}^{D} P_i\theta_i - n^{-1}\right\}\mathrm{Var}(Y\mid NR). \tag{A.42}$$

Finally, for DR-B,

$$E(\bar{y}_c) = E(Y), \tag{A.43}$$

$$\mathrm{Var}(\bar{y}_c) = n^{-1}\left[\mathrm{Var}(Y) + \mathrm{Var}(Y\mid NR)\left\{(1-f^2)f^{-1}\left(\sum_{i=1}^{D} P_i\theta_i\right) - (1-f)(nf)^{-1}\right\}\right], \tag{A.44}$$

and

$$E(\hat{\sigma}_c^2) = \mathrm{Var}(Y) - (n-1)^{-1}\mathrm{Var}(Y\mid NR)\left\{(1-f^2)f^{-1}\left(\sum_{i=1}^{D} P_i\theta_i\right) - (1-f)(nf)^{-1}\right\}. \tag{A.45}$$

Table 2. Imputation Methods

Method	Number of Imputations for the i-th Category	Value for Nonrespondents of Dependent Variable Corresponding to the i-th Category
	a. Standard Methods	
N	None. See (3.1).	$Y_{(i)}$
R	Random sample of size (n-r) from multinomial $\{r_i/r\}$.	$Y_{(i)}$
	b. $\Pr(Y = Y_{(i)} \mid$ Nonrespondent) Specified	
MC	$(n-r)\phi_i$	$Y_{(i)}$
MI	$(n-r)\phi_i^*$	$Y_{(i)}$
RC	Random sample of size (n-r) from multinomial $\{\phi_i\}$.	$Y_{(i)}$
RI	Random sample of size (n-r) from multinomial $\{\phi_i^*\}$.	$Y_{(i)}$
	c. Scale-Change Methods	
SC	Same as for R.	$Y_{(i)}\phi_i/\pi_i$
SC-RE	Same as for R.	$Y_{(i)}\phi_i r/r_i$
SD	Same as for R.	$Y_{(i)}b_i r/br_i$
SI	Same as for R.	$Y_{(i)}\phi_i^*/\pi_i^*$
SI-RE	Same as for R.	$Y_{(i)}\phi_i^* r/r_i$
	d. Double Sampling Methods	
DM	$(n-r)b_i/b$	$Y_{(i)}$
DM-B	$(n-r-b)b_i/b$	$Y_{(i)}$
DR	Random sample of size (n-r) from multinomial $\{b_i/b\}$.	$Y_{(i)}$
DR-B	Random sample of size (n-r-b) from multinomial $\{b_i/b\}$.	$Y_{(i)}$

Note: The total sample size is n, of which r are respondents with r_i having $Y = Y_{(i)}$. A subsample of b nonrespondents has b_i with $Y = Y_{(i)}$. ϕ_i^* and π_i^* are constants presumed to estimate ϕ_i and π_i. The remaining symbols are defined in Table 1 and formulas (3.2) and (3.3).

The nomenclature is: N means no imputation while R and M denote, respectively, random and mean imputation. Also, C denotes a correct and I an incorrect specification of ϕ_i or π_i. For the scale-change methods (S), RE denotes estimation of π_i by (r_i/r) while D ("double sampling") denotes estimation of π_i by (r_i/r) and ϕ_i by (b_i/b). Finally, for the double sampling methods (D) "B" implies use of $\bar{y}_c$ as in (3.6) rather than as defined in (2.2).

Table 3. Unconditional Distribution of Y, Probability of Nonresponse Given Y = y and Moments of Distributions for Respondent and Nonrespondent Populations. Sales 5052.

$Y_{(i)}$(in 000's)	P_i	θ_i
25	.058	.676
75	.049	.553
125	.043	.685
200	.059	.561
312.5	.057	.604
437.5	.046	.428
625	.069	.654
875	.050	.546
1250	.062	.373
1750	.059	.474
2250	.039	.365
2750	.033	.307
3500	.051	.280
4500	.042	.359
6250	.073	.356
8750	.037	.388
12500	.055	.243
20000	.051	.112
37500	.065	.174

NOTE: $E(Y \mid R) = 7893$, $\mathrm{Var}(Y \mid R) = 1.24\times10^8$, $\Pr(R) = 0.57$; $E(Y \mid NR) = 3101$, $\mathrm{Var}(Y \mid NR) = 4.43\times10^7$

Table 4. Unconditional Distribution of Y, Probability of Nonresponse Given Y = y and Moments of Distributions for Respondent and Nonrespondent Populations. Payroll 5052.

$Y_{(i)}$ (in 000's)	P_i	θ_i
25	.250	.571
75	.137	.376
125	.125	.385
200	.174	.411
312.5	.103	.259
437.5	.056	.269
625	.064	.283
875	.041	.287
1250	.024	.139
1750	.018	.569
2250	.008	.229

NOTE: $E(Y \mid R) = 314$, $E(Y \mid NR) = 225$
$Var(Y \mid R) = 148092$, $Var(Y \mid NR) = 126318$
$Pr(R) = 0.60$

Table 5. Sales 5052. Values[a] of B, P, Q^2 for Imputation Methods in Table 2 for Several Choices[b] of (n,f).

Method	B	P	Q^2
	a. (n,f) = (200,0.5)		
N	1.98	0.30	1.00
R	1.78	0.29	0.45
MC	0.00	0.00	1.25
RC	0.00	0.00	1.00
SC	0.00	-0.16	0.74
SC-RE	0.00	-0.16	1.00
SD	0.00	-0.10	0.69
DM	0.00	0.00	0.83
DR	0.00	0.00	0.71
	b. (n,f) = (200,0.1)		
SD	0.00	0.16	0.37
DM	0.00	-0.01	0.35
DR	0.00	-0.01	0.33
	c. (n,f) = (100,0.5)		
N	1.40	0.30	1.00
R	1.26	0.29	0.45
MC	0.00	0.00	1.25
RC	0.00	0.00	1.00
SC	0.00	-0.16	0.74
SC-RE	0.00	-0.16	1.00
SD	0.00	0.00	0.71
DM	0.00	0.00	0.83
DR	0.00	0.00	0.71
	d. (n,f) = (100,0.3)		
SD	0.00	0.05	0.61
DM	0.00	0.00	0.68
DR	0.00	-0.01	0.60
	e. (n,f) = (40,0.5)		
N	0.88	0.30	1.00
R	0.80	0.26	0.44
MC	0.00	0.01	1.26
RC	0.00	0.00	1.00
SC	0.00	-0.17	0.73
SC-RE	0.00	-0.16	1.00
SD	0.00	0.16	0.75
DM	0.00	-0.01	0.83
DR	0.00	-0.01	0.72

[a] $B = bias(\bar{y}_c)/\{Var(\bar{y}_c)\}^{1/2}$, $P = \{E(\hat{\sigma}_c^2) - \sigma^2\}/\sigma^2$,
$Q^2 = E\{\hat{\sigma}_c^2/n\ Var(\bar{y}_c)\}$.

[b] The results for (n,f) and (n,f') are the same for methods N, R, MC, RC, SC, SC-RE.

Table 6. Payroll 5052. Values[a] of B, P, Q^2 for Imputation Methods in Table 2 for Several Choices[b] of (n,f).

Method	B	P	Q^2
	a. (n,f) = (200,0.5)		
N	1.01	0.05	1.00
R	0.91	0.04	0.48
MC	0.00	0.00	1.56
RC	0.00	0.00	1.00
SC	0.00	0.08	0.55
SC-RE	0.00	0.09	1.00
SD	0.00	0.75	0.71
DM	0.00	0.00	0.73
DR	0.00	0.00	0.58
	b. (n,f) = (200,0.1)		
SD	0.00	3.38	0.55
DM	0.00	0.00	0.23
DR	0.00	0.00	0.22
	c. (n,f) = (100,0.5)		
N	0.72	0.05	1.00
R	0.64	0.04	0.48
MC	0.00	0.00	1.56
RC	0.00	0.00	1.00
SC	0.00	0.08	0.54
SC-RE	0.00	0.09	1.00
SD	0.00	1.41	0.77
DM	0.00	0.00	0.73
DR	0.00	-0.01	0.58
	d. (n,f) = (100,0.3)		
SD	0.00	2.29	0.74
DM	0.00	-0.01	0.54
DR	0.00	-0.01	0.46
	e. (n,f) = (40,0.5)		
N	0.45	0.05	1.00
R	0.41	0.02	0.47
MC	0.00	0.01	1.57
RC	0.00	0.00	1.00
SC	0.00	0.07	0.54
SC-RE	0.00	0.09	1.00
SD	0.00	3.39	0.86
DM	0.00	-0.01	0.73
DR	0.00	-0.02	0.59

[a] See footnote *a* to Table 5.
[b] See footnote *b* to Table 5.

Table 7. Unconditional Distribution of Y; Correct and Incorrect Probabilities of Nonresponse Given Y = y. Sales 5181.

$Y_{(i)}$	P_i	Correct θ_i	Incorrect θ_i 1	2	3	4
125	.09	.64	.60	.40	.50	.23
625	.22	.28	.25	.40	.40	.23
2000	.33	.18	.15	.20	.20	.23
6500	.26	.13	.10	.11	.10	.23
30000	.10	.16	.10	.11	.05	.23

NOTE: The probability of nonresponse is about 0.23 for each choice of the θ_i.

Table 8. Sales 5181. Values[a] of B, P, Q^2 for Imputation Methods MC and MI for Several Choices of n and Several Alternative[b] Choices of the θ_i.

Incorrect Set of θ_i	n	B		P		Q^2	
		MC	MI	MC	MI	MC	MI
1	100	0.00	-0.19	0.00	-0.03	1.22	1.17
	40	0.00	-0.12	0.00	-0.03	1.22	1.17
	20	0.00	-0.09	0.01	-0.02	1.23	1.17
2	100	0.00	-0.21	0.00	-0.04	1.22	1.15
	40	0.00	-0.13	0.00	-0.04	1.22	1.16
	20	0.00	-0.09	0.01	-0.03	1.23	1.16
3	100	0.00	-0.46	0.00	-0.09	1.22	1.08
	40	0.00	-0.29	0.00	-0.09	1.22	1.08
	20	0.00	-0.20	0.01	-0.09	1.23	1.08
4	100	0.00	0.51	0.00	0.04	1.22	1.28
	40	0.00	0.32	0.00	0.05	1.22	1.29
	20	0.00	0.23	0.01	0.05	1.23	1.30

[a]See footnote *a* to Table 5 for definitions of B, P and Q^2.
[b]The alternative choices of θ_i are given in Table 7.

Table 9. Sales 5181. Values[a] of B, P, Q^2 for Imputation Methods RC and RI for Several Choices of n and Second Alternative[b] Choices of the θ_i.

Incorrect Set of θ_i	n	B		P		Q^2	
		RC	RI	RC	RI	RC	RI
1	100	0.00	-0.18	0.00	0.19	1.00	1.23
	40	0.00	-0.11	0.00	0.20	1.00	1.24
	20	0.00	-0.08	0.00	0.20	1.00	1.24
2	100	0.00	-0.19	0.00	0.18	1.00	1.24
	40	0.00	-0.12	0.00	0.19	1.00	1.24
	20	0.00	-0.09	0.00	0.19	1.00	1.24
3	100	0.00	-0.44	0.00	0.14	1.00	1.25
	40	0.00	-0.28	0.00	0.14	1.00	1.25
	20	0.00	-0.20	0.00	0.15	1.00	1.26
4	100	0.00	0.45	0.00	0.26	1.00	1.22
	40	0.00	0.29	0.00	0.27	1.00	1.22
	20	0.00	0.20	0.00	0.27	1.00	1.23

[a]See footnote *a* to Table 8.
[b]See footnote *b* to Table 8.

Table 10. Sales 5181. Values[a] of B, P, Q^2 for Imputation Methods SC-RE and SI-RE for Several Choices of n and Several Alternative[b] Choices of the θ_i.

Incorrect Set of θ_i	n	B		P		Q^2	
		SC-RE	SI-RE	SC-RE	SI-RE	SC-RE	SI-RE
1	100	0.00	-0.19	-0.08	-0.12	1.00	1.00
	40	0.00	-0.12	-0.08	-0.12	1.00	1.00
	20	0.00	-0.08	-0.08	-0.12	1.00	1.00
2	100	0.00	-0.20	-0.08	-0.13	1.00	1.00
	40	0.00	-0.13	-0.08	-0.13	1.00	1.00
	20	0.00	-0.09	-0.08	-0.13	1.00	1.00
3	100	0.00	-0.45	-0.08	-0.15	1.00	1.00
	40	0.00	-0.29	-0.08	-0.15	1.00	1.00
	20	0.00	-0.20	-0.08	-0.15	1.00	1.00
4	100	0.00	0.46	-0.08	0.01	1.00	1.00
	40	0.00	0.29	-0.08	0.01	1.00	1.00
	20	0.00	0.21	-0.08	0.01	1.00	1.00

[a]See footnote *a* to Table 8.
[b]See footnote *b* to Table 8.

REFERENCES

Bailar, B.A., Bailey, L. and Corby, C.A. (1978). A comparison of some adjustment and weighting procedures of survey data. *Survey Sampling and Measurement* (Namboodiri, N.K. ed.), 175-198, New York: Academic Press.

Bailar, J.C. and Bailar, B.A. (1978). Comparison of two procedures for imputing missing survey values. *Proc. Sect. Survey Res. Meth., Amer. Statist. Assoc.,* 462-467.

Chiu, H.Y. and Sedransk, J. (1986). A Bayesian procedure for imputing missing values in sample surveys. *J. Amer. Statist. Assoc.,* 667-676.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J.R. Statist. Soc. B,* 39, 1-38.

Ernst, L.R. (1978). Weighting to adjust for partial nonresponse. *Proc. Sect. Survey Res. Meth., Amer. Statist. Assoc.,* 468-472.

Ernst, L.R. (1980). Variance of the estimated mean for several imputation procedures. *Proc. Sect. Survey Res. Meth., Amer. Statist. Assoc.,* 716-720.

Herzog, T.N. and Rubin, D.B. (1983). Using multiple imputations to handle nonresponse in sample surveys. *Incomplete Data in Sample Surveys* (Vol. 2 — Theory and Bibliographies), 209-245, New York: Academic Press.

Jinn, J.H. and Sedransk, J. (1987). Effect on secondary data analysis of different imputation methods. *Proc. Third Annual Census Bureau Research Conference.*

Jinn, J.H. and Sedransk, J. (1989). Effect on secondary data analysis of common imputation methods. *Sociological Methodology.*

Kalton, G. and Kasprzyk, D. (1982). Imputing for missing survey responses. *Proc. Sect. Survey Res. Meth., Amer. Statist. Assoc.,* 22-33.

Little, R. (1986). Missing data in Census Bureau surveys. *Proc. Second Annual Census Bureau Research Conference,* 442-454.

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

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

Platek, R., Singh, M.P. and Tremblay, V. (1978). Adjustment for nonresponse in surveys. *Survey Sampling and Measurement* (Namboodiri, N.K. ed.), 157-174, New York: Academic Press.

Sande, I.G. (1982). Imputation in surveys: Coping with reality. *The American Statistician,* 36, 145-152.

Santos, R.L. (1981a). Effects of imputation on complex statistics, Technical Report, Survey Research Center, Univ. of Michigan, Ann Arbor, Michigan.

Santos, R.L. (1981b). Effects of imputation on regression coefficients. *Proc. Sect. Survey Res. Meth., Amer. Statist. Assoc.,* 140-145.