Survey Methodology, December 1994
Vol. 20, No. 2, pp. 137–147
Statistics Canada

Bias Corrections for Survey Estimates from Data with Ratio Imputed Values for Confounded Nonresponse

E. RANCOURT, H. LEE and C.-E. SÄRNDAL[1]

ABSTRACT

Most surveys suffer from the problem of missing data caused by nonresponse. To deal with this problem, imputation is often used to create a "completed data set", that is, a data set composed of actual observations (for the respondents) and imputations (for the nonrespondents). Usually, imputation is carried out under the assumption of unconfounded response mechanism. When this assumption does not hold, a bias is introduced in the standard estimator of the population mean calculated from the completed data set. In this paper, we pursue the idea of using simple correction factors for the bias problem in the case that ratio imputation is used. The effectiveness of the correction factors is studied by Monte Carlo simulation using artificially generated data sets representing various super-populations, nonresponse rates, nonresponse mechanisms, and correlations between the variable of interest and the auxiliary variable. These correction factors are found to be effective especially when the population follows the model underlying ratio imputation. An option for estimating the variance of the corrected point estimates is also discussed.

KEY WORDS: Conditional bias; Monte Carlo simulation; Restoring estimator; Variance estimation.

1. INTRODUCTION

Occurrence of nonresponse is rather a norm than an exception in surveys. Missing data caused by nonresponse are often imputed to obtain a completed data set and the standard estimator is applied to the completed data set assuming that the underlying response mechanism is unconfounded. However, a point estimate obtained in such a way is biased when the response mechanism is confounded. The bias in this case could be very severe as pointed out in Lee, Rancourt and Särndal (1994). A response mechanism is unconfounded, according to Rubin (1987, p. 39), if it does not depend on the variable under study, otherwise it is confounded. (A formal definition suitable for this paper will be given in Section 2.)

In a Bayesian framework, a concept similar to that of an unconfounded response mechanism is termed ignorable. For bias caused by a nonignorable response mechanism, Rubin (1977, 1987) and Little and Rubin (1987) considered a method to correct the respondent mean using auxiliary variables. In this approach, a linear regression is assumed between the variable of interest y and a vector of auxiliary variables $\boldsymbol{x}$. The regression coefficient vector for the nonrespondents is assumed to have a normal prior with mean equal to the regression coefficient vector for the respondents.

Assuming a logistic model for the response probability, Greenless, Reece and Zieschang (1982) proposed a method to deal with nonignorable nonresponse using maximum likelihood estimation. Further, a linear regression model is assumed for the relationship between y and $\boldsymbol{x}$, a vector of auxiliary variables. The logistic model of the response probability includes y and $\boldsymbol{z}$, a vector of other auxiliary variables. Assuming also that the error term of the regression is normally distributed, they obtain maximum likelihood estimates of the unknown parameters of the regression model and the logistic model. Finally, for a nonrespondent, an imputed value is calculated as the mean of the distribution of y conditional on the values of $\boldsymbol{x}$ and $\boldsymbol{z}$ for the nonrespondents, and the estimated parameters. Such a method may give good results when all the model assumptions are satisfied but is likely to be highly sensitive to the specifications of the two models. The adequacy of the response probability model is usually untestable. If data are available from an external source, however, then it may be possible to test the response probability model as Greenless *et al.* did in their application to the Current Population Survey data. This method is highly computer-intensive.

In the case of categorical data, a few methods have also been proposed to deal with the problem of nonignorable nonresponse. For instance, Baker and Laird (1988) try to model the response mechanism with the help of log-linear models. As well, causal modeling is discussed in Fay (1986, 1989).

Ratio imputation is often used at Statistics Canada, especially in repeated surveys. For instance, in the Monthly Survey of Manufacturing, a missing value of the current shipment is imputed by ratio imputation using previous month shipment as the auxiliary variable value. This simple method is very appealing to subject matter specialists because it reflects month-to-month movement.

[1] E. Rancourt and H. Lee, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6; C.-E. Särndal, Département de mathématiques et de statistique, Université de Montréal, C.P. 6128, succursale A, Montréal (Québec), Canada, H3C 3J7.

In this paper, we investigate the possibility of improving the estimator applied to data containing ratio imputation with the aid of simple correction factors. Therefore, we assume that imputation has already been performed, and try to correct the estimator. We focus our attention on the estimation of the mean. The use of simple correction factors would be very appealing to the user provided it works reasonably well. Such a procedure is also easy to implement without resorting to excessive computational efforts and it enables us to avoid explicit modeling of the nonresponse mechanism. However, our approach differs from Rubin's in that we use sample dependent correction factors rather than an *a priori* chosen constant.

In Section 2, we define several simple correction factors that meet our requirements. In Section 3, we propose a variance estimator that may be used in conjunction with the corrected point estimators. The properties of the corrected point estimators were examined by a Monte Carlo simulation reported in Sections 4 and 5. Section 6 presents some concluding remarks.

2. SIMPLE BIAS CORRECTION FACTORS

Let $U = \{1, \ldots, k, \ldots, N\}$ denote the index set of a finite population and let the population mean of the variable of interest y be denoted by $\bar{y}_U = (1/N) \sum_U y_k$. We assume that $y_k > 0$ for all $k \in U$. From U, a simple random sample s of size n is drawn without replacement (SRSWOR). The unbiased estimator that would be used with 100% response is the sample mean

$$\bar{y}_s = (1/n) \sum_s y_k. \tag{2.1}$$

Let r and o be the sets of the responding and nonresponding units, respectively, so that $s = r \cup o$. We denote the SRSWOR sampling plan by $p(\cdot)$ and the response mechanism given s by $q(\cdot \mid s)$. That is, $p(s)$ is the probability that the SRSWOR sample s is drawn, and $q(r \mid s)$ is the probability that the set r responds given the sample s. Let also m and l be the sizes of r and o, respectively. For simplicity, we assume that the probability of $m = 0$ is negligible. We assume that imputation is carried out with the aid of an auxiliary variable, x, whose value, x_k, is known and positive for all $k \in s$. If $k \in o$, the missing value y_k is imputed by $\hat{y}_k$. The completed data set is denoted as $\{y_{\cdot k} : k \in s\}$ where $y_{\cdot k} = y_k$ if $k \in r$ and $y_{\cdot k} = \hat{y}_k$ if $k \in o$.

In this paper, we examine ratio imputation. This often-used imputation method is based on a simple model. That is, if the value y_k is missing, it is imputed by $\hat{B}_r x_k$, where $\hat{B}_r = (\sum_r y_k)/(\sum_r x_k)$. The model denoted ξ, is stating that, for $k \in s$.

$$y_k = \beta x_k + \epsilon_k, \quad E_\xi(\epsilon_k \mid x_k) = 0, \quad V_\xi(\epsilon_k \mid x_k) = \sigma^2 x_k,$$
$$E_\xi(\epsilon_k \epsilon_l \mid x_k, x_l) = 0, \quad k \neq l. \tag{2.2}$$

Under this model, $\hat{B}_r x_k$ is the best linear unbiased predictor of the missing value y_k, based on the respondent data $\{(y_k, x_k) : k \in r\}$. The completed data set is then composed of the values

$$y_{\cdot k} = \begin{cases} y_k, & \text{if } k \in r \\ \hat{B}_r x_k, & \text{if } k \in o. \end{cases} \tag{2.3}$$

The customary procedure is to apply the estimator formula used for 100% response to the completed data set. This gives

$$\bar{y}_{\cdot s} = \frac{1}{n} \sum_s y_{\cdot k} = \frac{\bar{y}_r}{\bar{x}_r} \bar{x}_s = \bar{y}_{\text{raimp}}, \tag{2.4}$$

where $\bar{x}_s = (1/n) \sum_s x_k$, $\bar{y}_r = (1/m) \sum_r y_k$ and $\bar{x}_r = (1/m) \sum_r x_k$. Note that raimp stands for ratio imputed.

It now becomes necessary to address the question whether the imputation can restore the full response estimator, $\bar{y}_s$, in the sense that the imputation estimator $\bar{y}_{\cdot s}$ is equal to $\bar{y}_s$ in expectation given s. Unless this can be achieved, the ratio imputation will have introduced bias. To examine this question, we must consider the response mechanism. A response mechanism $q(\cdot \mid s)$ is said to be *unconfounded* for the purpose of this paper if it is of the form $q(r \mid s) = q(r \mid x_s)$, where $x_s = \{x_k : k \in s\}$ and the response probabilities satisfy $P(k \in r \mid s) > 0$ for all $k \in s$. That is, it may depend on s and on the associated x-values. If it depends also on the y-values, so that $q(r \mid s) = q(r \mid x_s, y_s)$, then is is called *confounded*. In these definitions, the response mechanism is conditional on the realized sample s. Slightly different definitions of "confounded" and "unconfounded" are given in Rubin (1987, p. 39) where they are unconditional.

An example of an unconfounded response mechanism is

$$q(r \mid s) = \prod_{k \in r} (1 - \Theta_k) \prod_{k \in s-r} \Theta_k,$$

where $\Theta_k = 1 - P(k \in r \mid s) = 1 - e^{-\gamma x_k}$ for some positive constant γ, is the nonresponse probability of unit k. By contrast, if $\Theta_k = 1 - e^{-\gamma y_k}$, then $q(r \mid s)$ is a confounded mechanism.

A particularly simple unconfounded mechanism is the uniform response mechanism defined by $q(r \mid s) = (1 - \Theta)^m \Theta^{n-m}$. Here, units respond according to independent and identical Bernoulli $(1 - \Theta)$ trials, where Θ is the nonresponse probability common to all units.

Whether an imputation estimator $\hat{\bar{y}}_U$ of $\bar{y}_U$, including $\bar{y}_{\text{raimp}}$ given by (2.4), is considered good depends in part on the assumptions made by the analyst about the response mechanism and in part on the relation between y and x. Several possible assumptions are discussed later in this section. For any given s, the goal is that, under specified realistic assumptions, the expectation of the difference

$\hat{\bar{y}}_U - \bar{y}_s$ should be close to zero. That is, under the given assumptions, the conditional bias of $\hat{\bar{y}}_U$, $C\text{-bias}(\hat{\bar{y}}_U) = E(\hat{\bar{y}}_U - \bar{y}_s \mid s)$, should be small. We call $\hat{\bar{y}}_U$ a *restoring estimator* of $\bar{y}_U$ if $C\text{-bias}(\hat{\bar{y}}_U) = 0$ or ≈ 0, that is, if $\hat{\bar{y}}_U$ is (approximately) equal to $\bar{y}_s$ in conditional expectation. It follows that if the C-bias is (approximately) zero for any s, then the unconditional bias over all sample realizations s is also (approximately) zero.

Different analysts make different assumptions. Let us consider some typical assumptions and ask the question: What restoring estimators do these assumptions allow?

Assumption I: The response mechanism is uniform.

Under Assumption I, $\bar{y}_{\text{raimp}}$ is a restoring estimator. To see this, note that

$$C\text{-bias}(\bar{y}_{\text{raimp}}) = E_q(\bar{y}_{\text{raimp}} \mid s) - \bar{y}_s \approx 0,$$

because, given s, $\bar{y}_{\text{raimp}}$ is the classical ratio estimator of $\bar{y}_s$. Assumption I is unrealistic in most surveys. The response propensity is known to vary with observable characteristics such as size and industry (for business establishments), family size and type (for households), age, sex and income (for individuals). Under this unrealistic assumption, even a naive estimator such as the respondent mean, $\bar{y}_r = (1/m)\sum_r y_k$, is a restoring estimator:

$$C\text{-bias}(\bar{y}_r) = E_q(\bar{y}_r \mid s) - \bar{y}_s = 0.$$

However, if Assumption I holds, $\bar{y}_{\text{raimp}}$ is preferred to $\bar{y}_r$ because the ratio estimator feature leads to a smaller variance if the model ξ holds.

The analyst clearly needs to consider more realistic assumptions which allow the response probabilities to vary with background variables. The following assumption, composed of two parts, is of this kind.

Assumption II: (II-1): the response mechanism is unconfounded but otherwise arbitrary;

(II-2): the ratio model (2.2) holds.

Here (II-1) is a weaker and more realistic requirement on the response mechanism than the uniformity requirement in Assumption I. Under (II-1), the response mechanism can be of any form as long as it is unconfounded. However, Assumptions I and II are not directly comparable since II contains a model component, (II-2), which is lacking in I. Under Assumption II, $\bar{y}_{\text{raimp}}$ is a restoring estimator because

$$\begin{aligned} C\text{-bias}(\bar{y}_{\text{raimp}}) &= E_\xi\{E_q(\bar{y}_{\text{raimp}}) - \bar{y}_s \mid s\} \\ &= E_q E_\xi\left(\frac{\bar{y}_r}{\bar{x}_r}\bar{x}_s\right) - E_\xi(\bar{y}_s) \\ &= E_q(\beta\bar{x}_s) - \beta\bar{x}_s = 0. \end{aligned}$$

Note that changing the order of the expectations, $E_\xi E_q$ to $E_q E_\xi$, is allowed under Assumption II, because the response mechanism is then of the form $q(r \mid x_r)$, that is, it does not depend on the y-values. By contrast, the respondent mean $\bar{y}_r$ is not a restoring estimator because

$$C\text{-bias}(\bar{y}_r) = E_\xi\{E_q(\bar{y}_r) - \bar{y}_s \mid s\} = \beta\{E_q(\bar{x}_r \mid s) - \bar{x}_s\},$$

which is generally nonzero under Assumption II. We can, however, transform $\bar{y}_r$ into a restoring estimator by the use of a multiplicative correction factor. This leads to

$$\bar{y}_r\left\{1 + \left(1 - \frac{m}{n}\right)\left(\frac{\bar{x}_o}{\bar{x}_r} - 1\right)\right\}, \tag{2.5}$$

which is just another way of writing $\bar{y}_{\text{raimp}}$, as can easily be verified. In an example using the Bayesian approach, Little and Rubin (1987, p. 233) arrive at an estimator identical to the estimator (2.5).

Let us now consider confounded response mechanisms. They cause more difficult problems for finding a restoring estimator.

Assumption III: (III-1): the response mechanism is confounded but otherwise arbitrary;

(III-2): the ratio model (2.2) holds.

It is usually difficult, if not impossible, for the analyst to decide whether Assumption II or Assumption III is more appropriate. Examining the data will not be of much help if the only data available relate to the present point in time, as would typically be the case in a one-time survey. The assumption made (whether II or III) is then unverifiable. By contrast, if the analyst has experience with a regularly repeated survey, he or she may have legitimate reasons to believe, for example, that the nonresponse is a function of the variable of interest.

In some situations, the assumption of a confounded mechanism may be made on the following grounds. Suppose in a survey of personal finances that y, the variable under study is "savings" and that x, the auxiliary variable is "income", with values x_k known for the individuals $k \in s$. The nonresponse probability of respondent k is likely to be correlated with the savings figure y_k that he or she is asked to reveal as well as with the income figure x_k known from other sources. But since savings, not income, is the variable with which the respondent is directly confronted in the survey, the assumption that the nonresponse probability is a function of y_k may be more realistic than the assumption that it is a function of x_k. Hence a confounded mechanism may be more realistic to assume than an unconfounded mechanism.

Under Assumption III, neither $\bar{y}_r$ nor $\bar{y}_{\text{raimp}}$ are restoring estimators. The C-bias of $\bar{y}_{\text{raimp}}$ can be expressed as

$$C\text{-bias}(\bar{y}_{\text{raimp}}) = \bar{x}_s E_\xi E_q\left(\frac{\sum_r \epsilon_k}{\sum_r x_k}\right),$$

where ϵ_k is defined by the model (2.2). This C-bias is generally nonzero and can be quite large when the nonresponse rate is high and the correlation is not so strong. However, the C-bias is hard to evaluate, since the exact form of the response mechanism is left unspecified. Note that changing the order of the expectations E_ξ and E_q is not permitted under Assumption III since $q(r \mid s)$ depends on the y-values. For example, a negative C-bias is likely to occur if the respondent residual total, $\sum_r \epsilon_k$ tends to be negative.

A confounded response mechanism (as in Assumption III), introduces bias in the slope estimator $\hat{B}_r = (\sum_r y_k)/(\sum_r x_k)$. Consequently, $\hat{B}_r x_k$ is a biased imputation for a missing value y_k. To improve the situation, suppose that a missing value y_k is imputed by $C\hat{B}_r x_k$ instead of $\hat{B}_r x_k$, where C is a quantity to be specified. Then the data after imputation are given by

$$y^c_{\cdot k} = \begin{cases} y_k, & \text{if } \ k \in r \\ C\hat{B}_r x_k, & \text{if } \ k \in o \end{cases} \qquad (2.6)$$

and denoting the sample mean of these data as $\bar{y}_{c\cdot s} = (1/n)\sum_s y^c_{\cdot k}$, we get the estimator

$$\bar{y}_{c\cdot s} = \bar{y}_r\left[1 + \left(1 - \frac{m}{n}\right)\left(C\frac{\bar{x}_o}{\bar{x}_r} - 1\right)\right]. \qquad (2.7)$$

A simple correction of the type used in (2.6) was mentioned in Rubin (1986; 1987, p. 203) in the context of multiple imputation. Rubin views C as a fixed constant chosen by the user according to his or her prior knowledge. If such a choice happens to be well founded, the bias of (2.7) may be small.

Here, we shall examine choices of C that are adaptive, that is, they reflect the realized sample s and the realized response set r. Ideally, C should be such that the imputation will exactly restore the estimator $\bar{y}_s = (1/n)\sum_s y_k$ that would be used with 100% response. This C-value is determined by the equation

$$\bar{y}_s = \frac{1}{n}\sum_s y_k = \frac{1}{n}\sum_s y^c_{\cdot k} = \frac{1}{n}\left(\sum_r y_k + \sum_o C\hat{B}_r x_k\right).$$

A simple calculation shows that the optimal C-value is

$$C_{\text{opt}} = \frac{\hat{B}_o}{\hat{B}_r},$$

where $\hat{B}_o = \sum_o y_k / \sum_o x_k$ is the slope estimate if the model (2.2) could be fitted to nonrespondents. The imputed values would then be $\hat{y}_k = \hat{B}_o x_k$ for $k \in o$. Obviously, C_{opt} and $\hat{B}_o$ cannot be computed since they depend on missing y_k-values. For an unconfounded mechanism (as in Assumption II), we can expect $C_{\text{opt}} \approx 1$, given s, because

$$E_\xi E_q(C_{\text{opt}} \mid s) = E_q E_\xi\left(\frac{\hat{B}_o}{\hat{B}_r} \mid s\right) \approx 1.$$

But for a confounded mechanism (as in Assumption III), C_{opt} can be distinctly away from unity. Suppose that $C_{\text{opt}} > 1$. Note that $C_{\text{opt}} > 1$ if and only if $\sum_r e_{ks} < 0$ with $e_{ks} = y_k - \hat{B}_s x_k$, where $\hat{B}_s = (\sum_s y_k)/(\sum_s x_k)$ is the unknown slope estimate with 100% response. That is, $C_{\text{opt}} > 1$ implies that respondents' residuals e_{ks} are negative on the average. An illustration of this is shown in figure 1, where $n = 10$, $l = n - m = 5$, and all five respondents' residuals e_{ks} are negative.

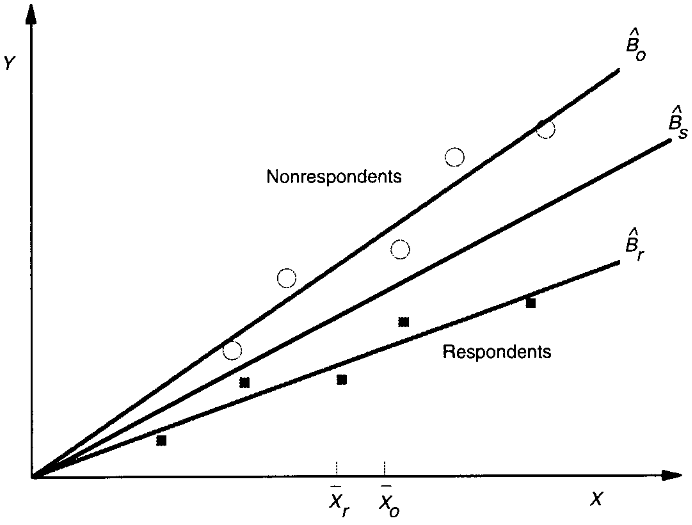

Figure 1. Example of data plot (y_k, x_k) for a confounded response mechanism.

Assuming that $C_{\text{opt}} > 1$, one approach for the analyst working under Assumption III is to choose a computable C likely to satisfy $C > 1$ and then use this C to construct the estimator (2.7). Factors C that will sometimes work in this manner are

$$c_1 = \frac{\bar{x}_o}{\bar{x}_r}, \quad c_2 = \frac{\bar{x}_o}{\bar{x}_s}, \quad c_3 = \frac{\bar{w}_o}{\bar{w}_r}, \quad c_4 = \frac{\bar{w}_o}{\bar{w}_s}. \qquad (2.8)$$

They are based on the logic that if the response mechanism is confounded in such a way that the nonresponse probability is a function of y (for example, $\Theta_k = 1 - e^{-\gamma y_k}$

with $\gamma > 0$), then both $C_{\text{opt}} > 1$, and $\bar{x}_o > \bar{x}_r$ are likely to occur, as Figure 1 illustrates. Conversely, if nonresponse is a decreasing function of y_k, then both $C_{\text{opt}} < 1$, and $\bar{x}_o < \bar{x}_r$ are likely to occur.

One important feature of such correction factors is that they can, but need not, be calculated during the imputation phase. For instance, if the usual ratio imputation $\hat{B}_r x_k$ was carried out at the imputation phase, it is then possible to calculate a suitable correction factor at the estimation phase without changing the originally imputed values.

Note that c_2 implies a somewhat milder correction than c_1: if $c_1 > 1$, we have $1 < c_2 < c_1$. The choices $C = c_3$ and $C = c_4$ are calculated on the ranks of the x-values, rather than on the x-values themselves, to dampen the effect of extreme x-values. More specifically, letting w_k be the rank of x_k in the data set $\{x_k : k \in s\}$, the w-means in c_3 and c_4 are $\bar{w}_s = (1/n)\sum_s w_k$, $\bar{w}_r = (1/m)\sum_r w_k$ and $\bar{w}_o = (1/l)\sum_o w_k$. The four estimators obtained by letting $C = c_i$ in (2.7) according to (2.8) will be denoted as $\bar{y}_{c_i \cdot s}$, $i = 1, \ldots, 4$. In particular, we have

$$\bar{y}_{c1\cdot s} = \bar{y}_r\left[1 + \left(1 - \frac{m}{n}\right)\left\{\left(\frac{\bar{x}_o}{\bar{x}_r}\right)^2 - 1\right\}\right], \qquad (2.9)$$

and

$$\bar{y}_{c2\cdot s} = \bar{y}_r\left[1 + \left(1 - \frac{m}{n}\right)\left\{\frac{\bar{x}_o^2}{\bar{x}_r\bar{x}_s} - 1\right\}\right]. \qquad (2.10)$$

The correction factors given in (2.8) are not ideal when the correlation between x and y is close to 1. In this case, we have $\hat{B}_r \approx \hat{B}_s \approx \hat{B}_o$, provided that the model (2.2) holds. Therefore, the correction factor C should be close to 1. However, the correction factors given in (2.8) could be very different from 1 and using them would bring bias. For this reason, it may be preferable to work with a correction factor C in (2.7) that takes the correlation into account. Correction factors of this kind are

$$k_i = 1 - \{(c_i^2 - 1)(\hat{R}_{xy}^2 - 1)\}, \qquad (2.11)$$

where c_i, $i = 1, \ldots, 4$, are the four correction factors given in (2.8), and $\hat{R}_{xy}$ is the estimated correlation coefficient based on the respondent data. In our Monte Carlo simulation we also included the estimator (2.7) corresponding to the four choices $C = k_i$, $i = 1, \ldots, 4$. These estimators will be denoted as $\bar{y}_{k_i \cdot s}$, $i = 1, \ldots, 4$.

3. VARIANCE ESTIMATION

Since we are interested in variance estimators based on single value imputation, the variance estimation method proposed in Särndal (1990, 1992) is of interest. Assuming unconfounded nonresponse and that the model ξ in (2.3) holds, the variance estimator for the point estimator $\bar{y}_{\text{raimp}}$ in (2.4) obtained by this method is given by

$$\begin{aligned}\hat{V}(\bar{y}_{\text{raimp}}) &= \left(\frac{1}{n} - \frac{1}{N}\right)\frac{\sum_s (y_{\cdot k} - \bar{y}_{\cdot s})^2}{n - 1} \\ &\quad + \left(\frac{1}{n} - \frac{1}{N}\right)A_o\hat{\sigma}^2 + \left(\frac{1}{m} - \frac{1}{n}\right)A_1\hat{\sigma}^2 \\ &= \hat{V}_{\text{ord}} + \hat{V}_{\text{dif}} + \hat{V}_{\text{imp}},\end{aligned} \qquad (3.1)$$

where

$$A_o = \frac{1}{n-1}\left\{\sum_o x_k - \frac{\sum_o x_k^2}{\sum_r x_k} + \frac{\bar{x}_s \sum_o x_k}{\sum_r x_k}\right\},$$

$$A_1 = \frac{\bar{x}_s\bar{x}_o}{\bar{x}_r}$$

and

$$\hat{\sigma}^2 = \frac{\sum_r e_k^2/(m - 1)}{\bar{x}_r\{1 - (\text{cv}_{xr})^2/m\}}, \qquad (3.2)$$

where

$$e_k = y_k - \hat{B}_r x_k, \; \text{cv}_{xr} = \frac{\sqrt{\sum_r (x_k - \bar{x}_r)^2/(m - 1)}}{\bar{x}_r}.$$

The variance of $\bar{y}_{\text{raimp}}$ has two components, namely, the sampling variance and the variance due to imputation. The first term in (3.1) (denoted by $\hat{V}_{\text{ord}}$) is an estimate of the sampling variance calculated using the ordinary variance formula assuming that imputed data are as good as real observations. Since this assumption does not hold, $\hat{V}_{\text{ord}}$ underestimates the true sampling variance. To correct this underestimation, the second term $\hat{V}_{\text{dif}}$ in (3.1) is added. The last term $\hat{V}_{\text{imp}}$ in (3.1) is an estimate of the variance due to imputation.

If we compute the mean of the y-values from the completed data set $\{y_{\cdot k}^c : k \in s\}$ given in (2.6), we get the estimator (2.7). Its variance estimator should take the correction factor C into account. If we can assume that the expectation $E_\xi E_p E_q$ is equal to $E_p E_q E_\xi$ (this is true under unconfounded nonresponse), we can use Särndal's (1990, 1992) method to obtain a variance estimator which takes C into account. However, we are mainly interested in confounded cases. We are therefore proposing a variance estimator based on the following heuristic argument.

The estimator $\hat{\sigma}^2$ in (3.2) uses the respondent data only. It will certainly be biased for confounded mechanisms and some correction is needed in order to use formula (3.1) for the corrected estimator (2.7). We suggest to replace $\hat{\sigma}^2$ in (3.1) by $C^2\hat{\sigma}^2$, to obtain the following variance estimator for the estimator $\bar{y}_{c\cdot s}$ in (2.7):

$$\hat{V}(\bar{y}_{c\cdot s}) = \hat{V}^{c}_{\text{ord}} + C^2(\hat{V}_{\text{dif}} + \hat{V}_{\text{imp}}), \qquad (3.3)$$

where $\hat{V}^{c}_{\text{ord}}$ is computed using the data after imputation with the bias correction factor C. Replacing C^2 by c_i^2 or k_i^2, we obtain the variance estimators corresponding to $\bar{y}_{ci\cdot s}$ or $\bar{y}_{ki\cdot s}$. The resulting variance estimators work quite well in many of the cases covered in the simulation reported in Section 5.

4. SIMULATION STUDY

We are considering eight corrected estimators corresponding to the eight correction factors given in (2.8) and (2.11). A simulation study was conducted to determine whether the corrected estimators succeed in restoring $\bar{y}_s$ under different response mechanisms, in particular, confounded mechanisms. For comparison, we also included the uncorrected estimators $\bar{y}_r$ and $\bar{y}_{\text{raimp}} = \bar{x}_s\bar{y}_r/\bar{x}_r$ given by (2.2). Our primary objective was to examine the corrected estimators when the finite population follows the ratio model ξ given by (2.3). However, we also wanted to see how the corrected estimators behave under relationships other than linear regression through the origin.

We also studied the coverage rates associated with the different estimators when the confidence intervals are computed with the aid of the variance estimators proposed in Section 3.

For the simulation, we generated 12 different finite populations, each of size $N = 100$, by specifying in different ways the constants a, b, c, and d in the regression model:

$$\Xi : y_k = a + bx_k + cx_k^2 + \epsilon_k, \quad E_{\Xi}(\epsilon_k) = 0,$$

$$V_{\Xi}(\epsilon_k) = d^2x_k, \qquad (4.1)$$

where the ϵ_k are assumed to be independent. Four different regression types were created by four different specifications of (a, b, c). These types are called RATIO ($a = c = 0, b > 0$, thus conforming to the ratio model ξ in (2.3)), CONCAVE ($a = 0, b > 0, c < 0$), CONVEX ($a = 0, b > 0, c > 0$) and NONRATIO ($a \neq 0, b > 0$, $c = 0$). For each regression type, three different levels of the model correlation ρ_{xy}, 0.7, 0.8 and 0.9, were obtained by a suitable choice of d. This resulted in 12 specifications of (a, b, c, d) as shown in Table 1.

Table 1

Characteristics of the Populations

POP	TYPE	a	b	c	d	R_{xy}	MEAN of y
1	RATIO	0	1.5	0	6.12	0.69	70.95
2	RATIO	0	1.5	0	4.50	0.81	69.92
3	RATIO	0	1.5	0	2.91	0.90	72.67
4	CONCAVE	0	3	−0.01	6.78	0.71	117.27
5	CONCAVE	0	3	−0.01	4.83	0.81	114.57
6	CONCAVE	0	3	−0.01	2.80	0.90	112.11
7	CONVEX	0	0.25	0.01	5.98	0.71	35.89
8	CONVEX	0	0.25	0.01	4.22	0.81	37.06
9	CONVEX	0	0.25	0.01	2.35	0.90	43.92
10	NON-RATIO	20	1.5	0	6.12	0.71	95.25
11	NON-RATIO	20	1.5	0	4.50	0.81	94.46
12	NON-RATIO	20	1.5	0	2.91	0.90	93.32

For each of the 12 specifications, we generated 100 population values (y_k, x_k), $k = 1, \ldots, 100$, by a two step process. We used the Γ-distribution with parameters α and β. Its density is

$$\frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} \exp(-x/\beta) \quad \text{for} \quad x > 0. \qquad (4.2)$$

First, we generated 100 values x_k, $k = 1, \ldots, 100$, according to the Γ-distribution with parameters $\alpha = 3$, $\beta = 16$, implying that the mean is $\alpha\beta = 48$ and the variance $\alpha\beta^2 = 768$. Then, for each fixed x_k, $k = 1, \ldots, 100$, we generated one value y_k according to the Γ-distribution with parameters

$$\alpha = \frac{\{\mu(x)\}^2}{\sigma^2(x)} = \frac{(a + bx + cx^2)^2}{d^2x}, \qquad (4.3)$$

$$\beta = \frac{\sigma^2(x)}{\mu(x)} = \frac{d^2x}{a + bx + cx^2}, \qquad (4.4)$$

where $x = x_k$ and (a, b, c, d) is one of the 12 vectors fixed in advance. This implies that $E_{\Xi}(y_k \mid x_k) = \alpha\beta = a + bx_k + cx_k^2$ and $V_{\Xi}(y_k \mid x_k) = \alpha\beta^2 = d^2x_k$, as required under the model (4.1). The same x-values were used for all 12 populations. For the populations generated by this process, Table 1 shows the values of the population correlation R_{xy} and the population mean of y. Note that the values of a, b, c, and d were chosen so as to obtain realistic types of populations that can be encountered in practice.

To simulate nonresponse, we used five different nonresponse mechanisms, each defined by independent Bernoulli (Θ_k) trials, where the probability of nonresponse Θ_k for unit k was specified as follows:

(M1) Θ_k is constant and independent for all $k \in U$. This is the uniform response mechanism, therefore unconfounded.

(M2) Θ_k is a decreasing function of x_k specified as $\Theta_k = \exp(-\gamma x_k)$. This is an unconfounded mechanism.

(M3) Θ_k is an increasing function of x_k specified as $\Theta_k = 1 - \exp(-\gamma x_k)$. This is also an unconfounded mechanism.

(M4) Θ_k is a decreasing function of y_k specified as $\Theta_k = \exp(-\gamma y_k)$. This is a confounded mechanism.

(M5) Θ_k is an increasing function of y_k specified as $\Theta_k = 1 - \exp(-\gamma y_k)$. This is also a confounded mechanism.

Note that since we assume x and y to be positively correlated, both (M2) and (M4) are mechanisms such that large units respond more often than small units. The smaller units will be underrepresented in the response set r. Conversely, (M3) and (M5) are mechanisms such that small units respond more often than large units. The larger units will be underrepresented in the response set r.

The first mechanism corresponds to the naive Assumption I discussed in Section 2. (M2) and (M3) correspond to Assumption II while (M4) and (M5) represent fairly simple examples of the confounded mechanisms discussed in connection with Assumption III. For (M2), (M3), (M4) and (M5), the constant γ was determined in such a way that the average nonresponse probability $\bar{\Theta} = (1/N) \sum_U \Theta_k$, is equal to one of the values 10%, 20%, 30% and 40%. Therefore, for each population, there were $5 \times 4 = 20$ different combinations of nonresponse mechanism and nonresponse rate.

For each of the 12 populations, 1,000 samples of size $n = 30$ were drawn. Then for each realized sample, 50 response sets were generated using independent Bernoulli(Θ_k) trials according to one of the 20 combinations of nonresponse mechanism and nonresponse rate. Thus 50,000 response sets were realized for each of the $12 \times 20 = 240$ combinations resulting from cross-classifying the 12 populations with the 20 combinations of nonresponse mechanism and nonresponse rate.

5. RESULTS

We studied the two uncorrected estimators $\bar{y}_r$ (justified under Assumption I) and $\bar{y}_{\text{raimp}} = \bar{x}_s \bar{y}_r / \bar{x}_r$ (justified under Assumption II) and the 8 corrected estimators $\bar{y}_{ci \cdot s}$ and $\bar{y}_{ki \cdot s}$, $i = 1, \ldots, 4$ (justified under Assumption III). (We call both $\bar{y}_r$ and $\bar{y}_{\text{raimp}}$ uncorrected even though (2.5) shows that we can view $\bar{y}_{\text{raimp}}$ as a corrected version of the naive estimator $\bar{y}_r$. Recall that our principal aim is to correct the bias of $\bar{y}_{\text{raimp}}$ when the mechanism is confounded.) The performance of the 10 estimators is judged by the magnitudes of the relative bias (RB), the relative root mean square error (RRMSE), and the coverage rate (CVR). The RB and the RRMSE of a point estimator $\hat{\bar{y}}_U$ for $\bar{y}_U$ are defined respectively as,

$$\text{RB}(\bar{y}) = 100 \times \frac{E_p E_q(\hat{\bar{y}}_U) - \bar{y}_U}{\bar{y}_U},$$

$$\text{RRMSE}(\bar{y}) = 100 \times \frac{\sqrt{E_p E_q(\hat{\bar{y}}_U - \bar{y}_U)^2}}{\bar{y}_U}.$$

The expectations $E_p E_q(\hat{\bar{y}}_U)$ and $E_p E_q(\hat{\bar{y}}_U - \bar{y}_U)^2$ were estimated by Monte Carlo simulation using the 50,000 realized response sets for each of 240 combinations. With this number of replicates, the Monte-Carlo error was less than 0.1%, assuming that the distribution of the $\hat{\bar{y}}_U$'s is approximately normal. We will use the abbreviation ARB to denote the absolute relative bias, $|\text{RB}(\bar{y})|$.

We will also discuss the coverage rate (CVR) of the 95% confidence interval constructed as

$$\hat{\bar{y}}_U \pm 1.96\sqrt{\hat{V}(\hat{\bar{y}}_U)}, \tag{5.1}$$

where $\hat{\bar{y}}_U$ is one of the 10 estimators and $\hat{V}(\hat{\bar{y}}_U)$ the corresponding variance estimator. For $\bar{y}_{\text{raimp}}$ and the 8 corrected estimators, we used the variance estimators described in Section 3. For $\bar{y}_r$, we used the variance estimator

$$\hat{V}(\bar{y}_r) = \left(\frac{1}{m} - \frac{1}{N}\right) \sum_r (y_k - \bar{y}_r)^2/(m - 1).$$

The CVR is calculated as 100 times the proportion of the 50,000 response sets such that the interval computed in the manner of (5.1) includes the true mean $\bar{y}_U$.

For the following discussion, we group the corrected estimators into two groups: s-corrected estimators, which are based on correction factors involving $\bar{x}_s$ or $\bar{w}_s$, that is, c_2, c_4, k_2 and k_4 and r-corrected estimators, which are based on correction factors involving $\bar{x}_r$ or $\bar{w}_r$, that is, c_1, c_3, k_1 and k_3.

The nonresponse mechanism is the key to the performance of the various estimators. Therefore, Tables 2 and 3 show the behavior of the estimators separately for each of the five mechanisms. We noted that the correlation level and the nonresponse rate do not have a very pronounced effect on the ranking of the estimators. Thus the performance measures ARB, RRMSE and CVR were averaged over 12 cases (three correlation levels × four nonresponse rates). These averages are shown in Table 2 for the RATIO type regression and in Table 3 for the CONCAVE, CONVEX and NONRATIO regression types.

Table 2

Average ARB, RRMSE (RM) and CVR of Ten Different Estimators for the RATIO Type Populations

For each mechanism, 12 cases were averaged (four nonresponse rates × three correlation levels)

	M1 (uniform)			M2 (decreasing-x)			M3 (increasing-x)			M4 (decreasing-y)			M5 (increasing-y)		
	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR
$\bar{y}_r$	0.2	13.9	92.5	12.9	19.1	86.0	9.5	16.5	81.1	19.1	23.6	72.3	14.9	19.9	68.2
$\bar{y}_{\text{raimp}}$	0.2	12.3	92.7	0.6	11.8	93.0	0.4	12.9	92.4	5.3	13.0	92.5	6.0	13.9	85.6
$\bar{y}_{c2\cdot s}$	1.0	13.3	92.4	4.4	12.6	88.9	8.9	18.3	93.0	1.8	11.8	92.4	3.6	15.3	92.2
$\bar{y}_{c4\cdot s}$	0.9	13.2	92.3	4.7	12.6	88.6	8.4	17.7	93.0	1.7	11.7	92.3	3.4	14.9	92.2
$\bar{y}_{k2\cdot s}$	1.1	13.2	92.8	2.4	12.0	90.9	8.0	18.5	93.5	1.7	11.7	93.3	2.2	15.3	92.0
$\bar{y}_{k4\cdot s}$	1.0	13.1	92.7	2.6	12.0	90.8	7.3	17.7	93.5	1.6	11.7	93.2	1.8	14.7	91.9
$\bar{y}_{c1\cdot s}$	1.7	14.7	91.4	5.9	13.4	86.4	15.7	26.2	87.6	1.9	12.2	90.9	8.9	21.3	89.8
$\bar{y}_{c3\cdot s}$	1.6	14.4	91.4	6.2	13.5	86.1	14.9	25.1	87.8	2.1	12.2	90.7	8.3	20.4	90.0
$\bar{y}_{k1\cdot s}$	2.0	14.7	92.3	3.1	12.3	90.0	15.9	29.6	88.9	1.1	11.7	92.8	8.3	23.8	90.7
$\bar{y}_{k3\cdot s}$	1.7	14.3	92.3	3.2	12.4	89.8	14.6	27.6	89.3	1.0	11.7	92.7	7.1	21.9	91.0

Table 3

Average ARB, RRMSE (RM) and CVR of Six Different Estimators for CONCAVE, CONVEX, and NONRATIO Populations

(For each mechanism, 12 cases are averaged as in Table 2)

	M1			M2			M3			M4			M5		
	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR
							CONCAVE								
$\bar{y}_r$	0.2	10.4	92.9	10.5	14.8	82.3	7.3	12.7	82.3	12.3	16.0	78.3	8.7	13.4	78.8
$\bar{y}_{\text{raimp}}$	0.2	9.4	94.5	1.4	9.1	93.4	2.6	10.5	94.9	1.9	9.2	94.9	2.1	9.7	92.9
$\bar{y}_{c2\cdot s}$	1.1	11.4	92.4	6.3	11.4	84.7	11.8	18.8	88.4	3.2	10.2	90.0	5.5	14.2	92.3
$\bar{y}_{c4\cdot s}$	1.0	11.1	92.8	6.6	11.5	84.3	11.4	18.0	88.8	3.6	10.3	89.8	5.5	13.7	92.7
$\bar{y}_{k2\cdot s}$	1.0	10.7	93.7	4.5	10.1	89.1	9.5	16.8	91.6	1.7	9.3	93.0	3.7	12.8	93.7
$\bar{y}_{k4\cdot s}$	0.9	10.5	93.8	4.6	10.1	89.0	9.0	16.0	91.8	1.8	9.3	92.8	3.5	12.3	93.9
							CONVEX								
$\bar{y}_r$	0.9	23.7	90.9	19.0	31.6	92.3	15.0	26.5	76.1	33.2	41.7	76.4	37.1	41.4	37.5
$\bar{y}_{\text{raimp}}$	0.6	21.4	90.6	5.8	21.7	92.8	7.0	22.1	85.6	14.0	25.0	90.0	27.6	33.5	52.0
$\bar{y}_{c2\cdot s}$	1.2	21.1	91.8	0.4	19.8	91.8	2.0	22.2	92.4	7.3	20.8	93.4	17.8	28.2	71.7
$\bar{y}_{c4\cdot s}$	1.2	21.3	91.5	0.3	19.9	91.5	1.8	22.3	92.4	6.7	20.6	93.4	18.5	28.5	70.5
$\bar{y}_{k2\cdot s}$	1.6	21.2	91.9	3.0	21.0	92.0	3.0	22.2	92.6	9.8	22.7	91.7	16.2	27.6	74.0
$\bar{y}_{k4\cdot s}$	1.4	21.3	91.6	2.9	21.0	91.8	2.6	22.0	92.3	9.5	22.7	91.7	17.6	27.7	72.6
							NON-RATIO								
$\bar{y}_r$	0.1	10.7	92.9	9.7	14.6	86.5	7.3	12.6	81.3	11.9	16.1	80.8	8.8	13.5	77.8
$\bar{y}_{\text{raimp}}$	0.2	9.6	94.5	2.1	9.5	92.4	2.6	10.5	95.3	2.1	9.6	94.4	1.6	9.9	93.3
$\bar{y}_{c2\cdot s}$	1.1	11.4	92.5	7.0	11.9	83.5	11.9	18.8	89.2	2.6	10.0	90.9	5.3	14.5	92.5
$\bar{y}_{c4\cdot s}$	1.0	11.3	92.4	7.3	12.1	82.8	11.5	18.1	89.4	2.7	10.1	90.6	4.9	13.8	92.7
$\bar{y}_{k2\cdot s}$	1.3	11.2	93.4	5.0	10.9	86.9	11.3	19.0	90.7	1.3	9.6	92.8	4.7	14.3	93.5
$\bar{y}_{k4\cdot s}$	1.1	10.9	93.4	5.2	11.1	86.5	10.6	17.8	91.1	1.3	9.7	92.6	4.1	13.4	93.8

We now comment on the tables. A conclusion of general character is that the respondent mean $\bar{y}_r$ has, as expected, a large bias and a very poor CVR for all of the nonuniform mechanisms. Its performance is satisfactory only for the uniform mechanism (M1). Thus we can focus on the comparisons between the uncorrected $\bar{y}_{\text{raimp}}$ on the one hand and the eight corrected estimators on the other. For both of the criteria ARB and RRMSE, we noted that the *s*-corrected estimators generally gave better results than the *r*-corrected ones. This is clearly seen in Table 2, where *s*-corrected and *r*-corrected estimators are displayed in two separate groups. Given this better behavior of the *s*-corrected group, we deleted the *r*-corrected group in Table 3.

5.1 RATIO Type Regression

From Table 2, we draw the following conclusions.

(i) The mechanism (M1) (uniform nonresponse).

When the mechanism (M1) holds, the uncorrected estimator $\bar{y}_{\text{raimp}}$ is essentially bias free, and there is no need to correct. However, if the analyst, suspecting a confounded mechanism, has nevertheless chosen one of the corrected estimators, the penalty is not severe. The eight corrected estimators show only a small increase in ARB and in RRMSE compared to $\bar{y}_{\text{raimp}}$.

(ii) The mechanisms (M2) and (M3) (unconfounded, nonuniform and *x*-value dependent).

For these mechanisms, the ARB is seen to be very small for the uncorrected estimator $\bar{y}_{\text{raimp}}$, as theory would lead us to expect. Our interest is instead focused on the behavior of the eight corrected estimators, since it is important to know if a penalty is associated with an incorrect decision to use one of these estimators. Such a decision would be brought about by an incorrect assumption that the response mechanism is confounded (when in fact it is unconfounded but nonuniform). Table 2 shows that there is indeed some penalty in the form of both increased ARB and increased RRMSE. The penalty is less severe for the *s*-corrected group. For both groups, the penalty is less severe for the mechanism (M2) than for the mechanism (M3).

(iii) The mechanism (M4) (confounded and *y*-value dependent).

For this mechanism, a striking feature of Table 2 is that all eight corrected estimators give a substantial bias reduction compared to the uncorrected estimator $\bar{y}_{\text{raimp}}$ (and a very large reduction relative to the naive estimator $\bar{y}_r$). The corrected estimators also show some improvement in RRMSE compared to $\bar{y}_{\text{raimp}}$. The *s*-corrected estimators perform better than the *r*-corrected ones. Within the *s*-corrected group of estimators, the differences are minor, as is the case within the *r*-corrected group.

(iv) The mechanism (M5) (confounded and *y*-value dependent).

Table 2 shows that the *s*-corrected estimators have a smaller ARB than the uncorrected $\bar{y}_{\text{raimp}}$; their RRMSE is slightly higher. By contrast, the *r*-corrected estimators "overcorrect" so that both the ARB and the RRMSE exceed the levels observed for $\bar{y}_{\text{raimp}}$. The *r*-corrected group does not perform well for this mechanism.

In summary, Table 2 shows that if the ratio model (2.2) holds and the assumption of a confounded mechanism is correctly made, the decision to use one of the corrected estimators may lead to a reduced bias. The main difficulty facing the analyst is to accurately predict the nature of the response mechanism causing nonresponse. In particular, it may be difficult for the analyst to separate a confounded mechanism (*e.g.*, one with $\Theta_k = e^{-\gamma y_k}$) from a similar nonuniform unconfounded mechanism (*e.g.*, one with $\Theta_k = e^{-\gamma x_k}$). Yet this subtle difference has a marked effect on the bias of $\bar{y}_{\text{raimp}}$ and on the decision whether or not to use a corrected estimator. When the nonuniform unconfounded type applies, we have seen that there is a penalty associated with the corrected estimators, in particular with the *r*-corrected group.

5.2 Other Regression Types

Table 3 shows the performance of six estimators (the two uncorrected and the four *s*-corrected) for the CONCAVE, CONVEX, and NONRATIO regression types. As in Table 2, there is little to choose between the estimators when the uniform mechanism (M1) holds. For the two confounded mechanisms, the results in Table 3 do not send a clear message that *s*-corrected estimation should be attempted even if the assumption of a confounded mechanism is correctly made. Compared to the uncorrected $\bar{y}_{\text{raimp}}$, the *s*-corrected estimators show a clearly improved performance (in terms of smaller ARB and smaller RRMSE) only for the CONVEX population type. Even in this case, a substantial bias remains after the attempt at correction. For the two unconfounded nonuniform mechanisms (M2) and (M3), it is *a priori* clear that one would not expect improved performance on the part of the *s*-corrected estimators when compared to $\bar{y}_{\text{raimp}}$. Oddly enough however, we find that the *s*-corrected estimators work very well for the CONVEX population. These conclusions leave the analyst with a difficult choice if a RATIO type population cannot be assumed. Then it is difficult on the basis of our findings to recommend the use of one of the corrected estimators.

5.3 Coverage Rates

Tables 2 and 3 also show that the variance estimation procedure suggested in Section 3 generally works well. Indeed the coverage rates for the corrected estimators are uniformly good whenever the ARB is small. In particular,

Table 4

Average ARB, RRMSE (RM) and CVR of the Two Uncorrected Estimators and the c_4 - and k_4 - Corrected Estimators
(Averaged Over All Population Types)

	M1			M2			M3			M4			M5			Overall		
	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR
$\bar{y}_r$	0.3	14.7	92.3	13.0	20.0	86.8	9.8	17.1	80.2	19.1	24.4	77.0	17.4	22.1	65.6	11.9	19.6	80.4
$\bar{y}_{\text{raimp}}$	0.3	13.2	93.1	2.5	13.0	92.9	3.1	14.0	92.0	5.8	14.2	93.0	9.3	16.7	81.0	4.2	14.2	90.4
$\bar{y}_{c4 \cdot s}$	1.0	14.2	92.3	4.7	14.0	86.8	8.3	19.0	90.8	3.7	13.2	91.5	8.1	17.7	87.0	5.2	15.6	89.7
$\bar{y}_{k4 \cdot s}$	1.1	14.0	92.9	3.8	13.6	89.5	7.4	18.4	92.2	3.6	13.3	92.6	6.7	17.0	88.0	4.5	15.2	91.0

for the unconfounded mechanisms (M2) and (M3), the coverage rates for the corrected estimators are about equal to or better than those for the uncorrected estimators.

5.4 Overall Comments

From the summary Table 4, we note that, as expected, $\bar{y}_r$ and $\bar{y}_{\text{raimp}}$ show the best performance for the uniform response mechanism (M1). The uncorrected estimator $\bar{y}_{\text{raimp}}$ is the best one for the unconfounded mechanisms (M2) and (M3), while the corrected estimators are the best ones for the confounded mechanism (M4) and (M5).

Finally, on the average over all 240 cases included in our study, we note from the overall column of Table 4 that $\bar{y}_{\text{raimp}}$ and $\bar{y}_{k4 \cdot s}$ perform similarly with the former having a slightly smaller bias and the latter having slightly better coverage rate.

6. CONCLUSIONS

It has long been recognized that nonresponse causes bias in survey estimates, except in rare cases. Imputation is a widely used practice to handle nonresponse, because it is convenient to work with a complete data set. There are many imputation rules as well as some softwares that can be used in large scale surveys. Imputation is sometimes applied without critical questioning, and, although widely used, imputation does not solve the critical problem of bias caused by nonresponse.

In this paper, we have examined ratio imputation. The ordinary ratio imputation $\hat{B}_r x_k$ is justified (that is, it produces no bias) if two conditions hold: (a) the regression model behind the ratio imputation rule holds (that is, a linear regression through the origin); (b) the response mechanism is unconfounded.

The results of our simulation give some idea of the magnitude of the bias of the usual ratio imputation estimator $\bar{y}_{\text{raimp}}$ when one or both of the two conditions break down. We considered several nonuniform response mechanisms, confounded as well as unconfounded mechanisms. We also considered breakdown of the regression model behind ratio imputation.

We argued that a confounded mechanism can sometimes be realistically assumed in a survey. We showed that if an assumption of confounded response mechanism is correctly made, and if the model behind the ratio imputation is valid, one can make some progress toward bias reduction using the s-corrected estimators in this paper. They have substantially less bias than the uncorrected estimator $\bar{y}_{\text{raimp}}$. The s-corrected estimators are generally more effective than the r-corrected estimators for reducing the bias.

Suppose the analyst is working under the assumption that the ratio model (2.2) holds. Our simulation study then leads to suggested estimators according to the following Table 5, depending on the assumed nature of the response mechanism and on the nonresponse rate. The entry "any" means any of the 10 estimators in Table 2.

Table 5

Suggested Estimators for Each Nonresponse Mechanism

Nonresponse Rate	Suggested Estimator		
	Response Mechanism		
	Uniform	Unconfounded	Confounded
($\leq$ 10%)	any	any but $\bar{y}_r$	any but $\bar{y}_r$
($>$ 10%)	any[1]	$\bar{y}_{\text{raimp}}$	s-corrected

Note 1: $\bar{y}_{\text{raimp}}$ as a slight advantage over the others.

If the regression model behind ratio imputation fails, the situation is less clear. Unless the naive assumption of a uniform response mechanism holds (which is unlikely), the uncorrected ratio imputation estimator $\bar{y}_{\text{raimp}}$ can have considerable bias. We found that $\bar{y}_{\text{raimp}}$ is particularly prone to bias for the CONVEX type population where the s-corrected group of estimators usually have smaller bias than $\bar{y}_{\text{raimp}}$. On the other hand, for the CONCAVE and the NONRATIO type populations, $\bar{y}_{\text{raimp}}$ is generally more resistant to bias than the s-corrected estimators.

7. ACKNOWLEDGMENT

The authors wish to thank the referees and the associate editor for their helpful comments. An earlier version of this paper was presented at the Annual Research Conference (ARC) in Arlington, Virginia, March 22-25, 1992.

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.

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

FAY, R.E. (1989). Estimating nonignorable nonresponse in longitudinal surveys through causal modeling. In *Panel Surveys* (Eds. D. Kasprzyk, G.J. Duncan, G. Kalton, and M.P. Singh), 375-399.

GREENLESS, J.S., REECE, W.S., and ZIESCHANG, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, 82, 251-261.

LEE, H., RANCOURT, E., and SÄRNDAL, C.-E. (1994). Experiments with variance estimation from survey data with imputed values. *Journal of Official Statistics*, 10, 231-243.

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

RUBIN, D.B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association*, 72, 538-543.

RUBIN, D.B. (1986). Basic ideas of multiple imputation for nonresponse. *Survey Methodology*, 12, 37-47.

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

SÄRNDAL, C.-E. (1990). Methods for estimating the precision of survey estimates when imputation has been used. *Proceedings: Symposium 90, Measurement and Improvement of Data Quality*, Statistics Canada, 337-347.

SÄRNDAL, C.-E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, 18, 241-252.