Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER Applied Mathematics and Computation 164 (2005) 249–261
APPLIED MATHEMATICS AND COMPUTATION
www.elsevier.com/locate/amc

Indirect methods of imputation of missing data based on available units

M.M. Rueda [a,*], S. González [b], A. Arcos [a]

[a] *Department of Statistics and OR, Faculty of Science, University of Granada, Spain*
[b] *Department of Statistics and OR, University of Jaen, Spain*

Abstract

One of the most difficult problems confronting investigators who analyze data from surveys is how to treat missing data. Many statistical procedures cannot be used immediately if any values are missing. Imputation of missing data before starting statistical analysis is then necessary. This paper proposes imputation methods of the mean based on indirect estimators of available cases. A complete simulation study was performed to test the proposed techniques.

Keywords: Auxiliary information; Missing data; Horvitz–Thompson estimator; Simple random sampling; Stratified random sampling; Imputation

1. Introduction

Surveys are a common method of data collection in economics and social sciences, but they often suffer from the problem of nonresponse. Any research

* Corresponding author. Address: Dept. de Estadística e I.O., Facultad de Ciencias, Avda. Fuentenueva, Universidad de Granada, 18071 Granada, España.
E-mail address: mrueda@ugr.es (M.M. Rueda).

0096-3003/$ - see front matter
doi:10.1016/j.amc.2004.04.102

which makes use of collected data must consider and deal with the question of missing data. In a survey study, data may be missing due to a variety of reasons; the respondent may not be present at the time of survey, or a certain class of respondent may not form part of the survey at all. Some of these factors affect the quality of the data. In such situations, the standard statistical procedures developed for data with no missing values cannot be immediately and straightforwardly applied for deducing inferences. Furthermore, an obvious consequence of nonresponse is that the actual sample size is less than the planned one, which can produce biases in estimations and an increase in sampling variance if missing data follows any pattern.

The problem of missing data can be addressed by various methods during the stages of data collection and processing, the aim of all such methods being to obtain a precise and complete data set. Nevertheless, it is still possible for errors and losses of entries to occur even once the data has been collected and filtered.

An initial option is to carry out a complete case analysis. Methods based on completely recorded units create a rectangular data set by discarding parts of the data. This is the simplest and most common approach to nonresponse. Among the advantages of this kind of analysis are its simplicity and the fact that different univariate statistics can be compared. However, it also has numerous disadvantages. Little and Rubin [4] pointed out the problems of methods that ignore incomplete observations. While these methods may be satisfactory when the percentage of incomplete cases is low, in general terms they lead to biased estimations, since they assume that the loss of data takes place in a completely random way. King et al. [2] illustrate how methods of complete cases are prone to serious errors. Thus, this practice may introduce bias into the estimate and increase sampling variance due to a reduction in sample size, see, e.g., [1,8].

Contending that the deleted observations may contain valuable information, an alternative approach is to try to improve the precision of the estimators by including all cases available for their calculation, see, e.g., [9,7] for an interesting account.

Alternatively, an imputation method to find substitutes for missing observations could be employed. By treating these imputed values as true observations, statistical analysis may be carried out using the standard procedures developed for data without any missing observations. In this study we propose a method for imputation of the mean based on indirect estimators of available cases. The procedure consists of making use of the information available from incomplete observations, and thus improving the precision of the indirect estimators. Therefore, we propose imputation methods on the basis of these indirect estimators present in available cases.

Finally, a simulation study is performed to test the functioning of the proposed techniques.

2. Imputation methods based on indirect estimators of available cases

Consider a population of N units from which a random sample, s, of fixed size, n is drawn according to a sample design $d = (S_d, P_d)$, with first order inclusion probabilities π_i. For this sample we observe the values of two variables, (y_i, x_i), $i = 1, \ldots, n$, for the estimation of some parameters of variable y.

It is assumed that a set of $(n - p - q)$ complete observations of selected units in the sample are available. In addition to these, observations of the x characteristic of p units in the sample are available but the corresponding observations for the y characteristic are missing. Similarly, we have a set of q observations of the y characteristic in the sample but the associated values of the x characteristic are missing. Furthermore, p and q are assumed to be integer numbers verifying $1 \leqslant p, q \leqslant n/2$. For the sake of simplicity, the unit of the sample s is separated into three disjoint sets $s_1 = \{i \in s / x_i, y_i \text{ are available}\}$, $s_2 = \{i \in s / x_i \text{ are available, but } y_i \text{ is not}\}$ and $s_3 = \{i \in s / y_i \text{ are available, but } x_i \text{ is not}\}$.

When an imputation method is applied, the set of complete data is specified by:

$$z_i = \begin{cases} y_i & \text{if } i \in s_1 \cup s_3, \\ \tilde{y}_i & \text{if } i \in s_2, \end{cases}$$

where $\tilde{y}_i$ is the imputed value, and from these data the necessary estimations can be calculated. Thus, the following estimators, among others, are obtained:

Parameter	Estimator
Mean	$\hat{\bar{y}}_{\text{imp}} = \frac{1}{N}\sum_{i\in s}\frac{z_i}{\pi_i}$
Total	$\hat{y}_{\text{imp}} = \sum_{i\in s}\frac{z_i}{\pi_i}$
Distribution function	$\hat{F}_{\text{imp}}(t) = \frac{1}{N}\sum_{i\in s}\frac{\Delta(t-z_i)}{\pi_i}$

Commonly used imputation methods include mean imputation. By using indirect estimation methods, the traditional ratio, difference and regression estimators of the mean can be used as the imputed values. However, if a large proportion of the data is missing, the usual estimators will be based on a relatively small sample and their precision will be reduced correspondingly. We propose, therefore, methods for imputation of the mean in which indirect estimation based on available cases is applied.

The Horvitz–Thompson estimators based on samples s_1, s_2 and s_3 are:

$$\hat{\bar{y}}^1_{\text{HT}} = \frac{1}{N}\sum_{i\in s_1}\frac{y_i}{\pi_i}, \quad \hat{\bar{y}}^3_{\text{HT}} = \frac{1}{N}\sum_{i\in s_3}\frac{y_i}{\pi_i}, \quad \hat{\bar{x}}^1_{\text{HT}} = \frac{1}{N}\sum_{i\in s_1}\frac{x_i}{\pi_i}, \quad \hat{\bar{x}}^2_{\text{HT}} = \frac{1}{N}\sum_{i\in s_2}\frac{x_i}{\pi_i}.$$

The following indirect estimators for the population mean based on complete cases can be formulated:

$$\hat{\bar{y}}_{r1} = \frac{\hat{\bar{y}}^1_{\mathrm{HT}}}{\hat{\bar{x}}^1_{\mathrm{HT}}}\overline{X}, \quad \hat{\bar{y}}_{d1} = \hat{\bar{y}}^1_{\mathrm{HT}} + (\overline{X} - \hat{\bar{x}}^1_{\mathrm{HT}}), \quad \hat{\bar{y}}_{\mathrm{Reg1}} = \hat{\bar{y}}^1_{\mathrm{HT}} + b(\overline{X} - \hat{\bar{x}}^1_{\mathrm{HT}}), \tag{1}$$

where b may be fixed and either known or unknown. In the latter case, if it is minimized, the error obtained, $b = \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)}$, must be estimated.

All these estimators discard the information available from incomplete cases. This practice may introduce bias and errors into the estimation, and so the following classes of estimators, incorporating all the available observations, are proposed:

$$\hat{\bar{y}}_{r2} = \frac{\alpha_r \hat{\bar{y}}^3_{\mathrm{HT}} + (1-\alpha_r)\hat{\bar{y}}^1_{\mathrm{HT}}}{\beta_r \hat{\bar{x}}^2_{\mathrm{HT}} + \beta_r \hat{\bar{x}}^1_{\mathrm{HT}}}\overline{X}, \tag{2}$$

$$\hat{\bar{y}}_{d2} = \alpha_d \hat{\bar{y}}^1_{\mathrm{HT}} + (1-\alpha_d)\hat{\bar{y}}^3_{\mathrm{HT}} + (\overline{X} - (\beta_d \hat{\bar{x}}^1_{\mathrm{HT}} + (1-\beta_d)\hat{\bar{x}}^2_{\mathrm{HT}})), \tag{3}$$

$$\hat{\bar{y}}_{\mathrm{Reg2}} = \alpha_{\mathrm{reg}} \hat{\bar{y}}^1_{\mathrm{HT}} + (1-\alpha_{\mathrm{reg}})\hat{\bar{y}}^3_{\mathrm{HT}} + b[\overline{X} - (\beta_{\mathrm{reg}} \hat{\bar{x}}^1_{\mathrm{HT}} + (1-\beta_{\mathrm{reg}})\hat{\bar{x}}^2_{\mathrm{HT}})]. \tag{4}$$

In the case of the regression estimator, if b is unknown, we can proceed as in the case of no nonresponse. Thus, two possible estimators for b are presented:

$$\hat{b}_1 = \frac{\mathrm{Cov}_{i\in s_1}(x,y)}{\mathrm{Var}_{i\in s_1}(x)} \quad \text{and} \quad \hat{b}_2 = \frac{\mathrm{Cov}_{i\in s_1}(x,y)}{\mathrm{Var}_{i\in s_1 \cup s_2}(x)}, \tag{5}$$

which will generate two regression estimators $\hat{\bar{y}}_{\mathrm{Reg21}}$ and $\hat{\bar{y}}_{\mathrm{Reg22}}$.

The estimators with subindex 1 are the traditional ratio, difference and regression estimators, which are based on complete observations and ignore the incomplete pairs of observations. We propose the estimators with subindex 2 which incorporate all the available observations.

The following step is to look for the estimators with the best behaviour among the proposed classes of estimators. These choices are made in order to minimize the estimation error. The expressions of the mean squared errors of the estimators are easily obtained; by minimizing these errors, the estimator expressions with minimum error are derived. The optimal coefficients $\alpha_{r_{\mathrm{opt}}}$, $\beta_{r_{\mathrm{opt}}}$, $\alpha_{d_{\mathrm{opt}}}$, $\beta_{r_{\mathrm{opt}}}$, $\alpha_{\mathrm{reg}_{\mathrm{opt}}}$ and $\beta_{\mathrm{reg}_{\mathrm{opt}}}$ can be seen in Appendix A.

Unfortunately, these optimum values depend on theoretical variances and covariances among the Horvitz–Thompson estimators, which are generally unknown. However, they can be estimated when the sample is drawn. Furthermore, these values would be estimated by replication methods. The expressions of these variances and covariances for the case of simple random sampling without replacement and stratified sampling can be seen in [7]. These estimations allow us to obtain approximate values, $\tilde{\alpha}_r$, $\tilde{\beta}_r$, $\tilde{\alpha}_d$, $\tilde{\beta}_d$, $\tilde{\alpha}_{\mathrm{reg}}$ and $\tilde{\beta}_{\mathrm{reg}}$, and to build the correspondence estimators $\tilde{\bar{y}}_{r2}$, $\tilde{\bar{y}}_{d2}$ and $\tilde{\bar{y}}_{\mathrm{Reg2}}$.

These estimators do not coincide with the theoretical estimators in expressions Eqs. (2)–(4), but, using the results obtained by Randles [6], who derived the asymptotic distribution of estimators with estimated parameters, Rueda and González [7] prove that asymptotically they have the same distribution, and it is reasonable to assume that the sampling errors will be close to the theoretical ones for large samples.

Finally, note that the usual estimators are included in the proposed classes of estimators, and so the estimators obtained by minimizing the errors in these classes will be better, in terms of precision, than the traditional ones.

Thus, the following imputation procedures are proposed:

- Procedure based on a ratio estimator: in this situation, specify the complete data set using the estimator $\hat{\bar{y}}_{r2}$ for the imputed value.
- Procedure based on a difference estimator: in this situation, specify the complete data set using the estimator $\hat{\bar{y}}_{d2}$ for the imputed value.
- Procedure based on a regression estimator with b unknown: in this case, we propose two regression estimators using the two possible estimators of the regression coefficient. Thus we derive two imputation procedures, with $\hat{\bar{y}}_{\text{reg21}}$ and $\hat{\bar{y}}_{\text{reg22}}$ being the imputed values.

After having used one of these imputation methods and specified the corresponding complete data set, the relevant inferences can be made. The functioning of some of the estimations that could be produced is discussed in Section 3 by means of a simulation study.

3. Simulation study

This section describes estimator properties by applying a simulation study. The populations considered can be divided into two groups: natural and simulated populations.

The FAM1500 population consists of 1500 families in Andalusia (Spain) taken from [3]. The variable of interest, y, denotes family income and the auxiliary x denotes expenditure on food and drink.

The second class includes two simulated populations used by Meeden [5]. For the purposes of the simulation, a superpopulation model is considered in which it is assumed that for each i, $y_i = bx_i + u_i e_i$, where e_i are independent, identically distributed random variables with zero expectations.

In the first population, SIM1, the auxiliary variable is a random sample from a log-normal population with mean and standard deviation 4.9 and 0.586 respectively.

In SIM2 the auxiliary variable is 50 plus a random sample from the standard exponential distribution.

Each of the simulated populations contains 500 units.

The following algorithm is used for the populations with several sample sizes.

Algorithm 1

- Step 1: Take a sample of size n according to the procedure of simple random sampling without replacement.

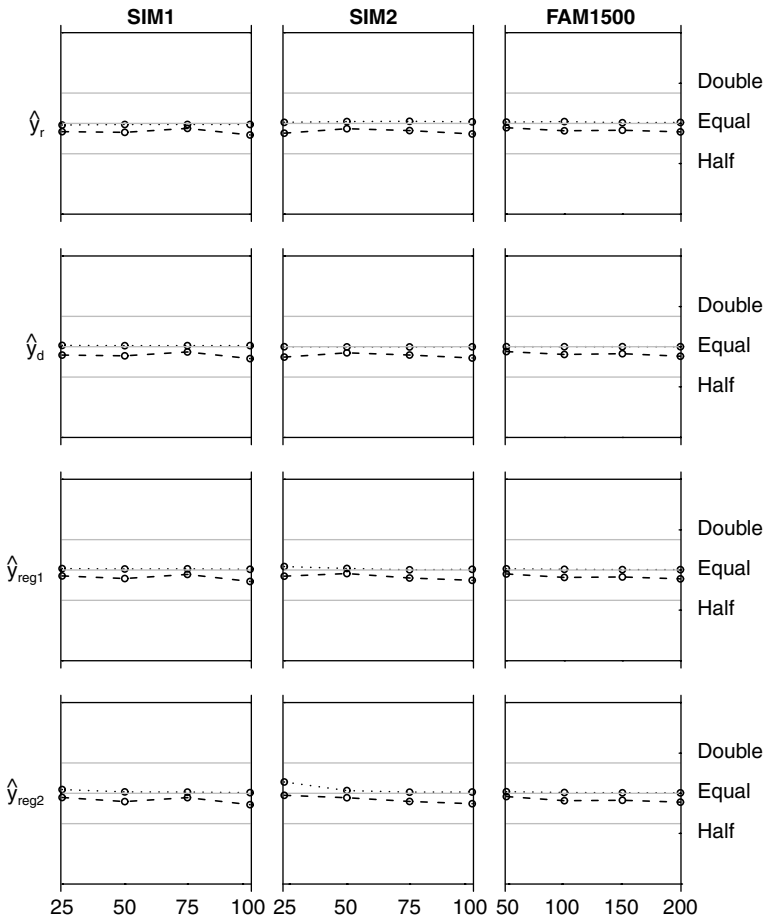

Fig. 1. Log ratios of standard errors comparing the *cases available* imputations and the *complete data* imputations against the *simple* imputation for the estimation of the population mean, $p = 0.32n$, $q = 0.48n$. The dotted curve corresponds to the *complete data* imputation and the dashed curve refers to the *cases available* imputation.

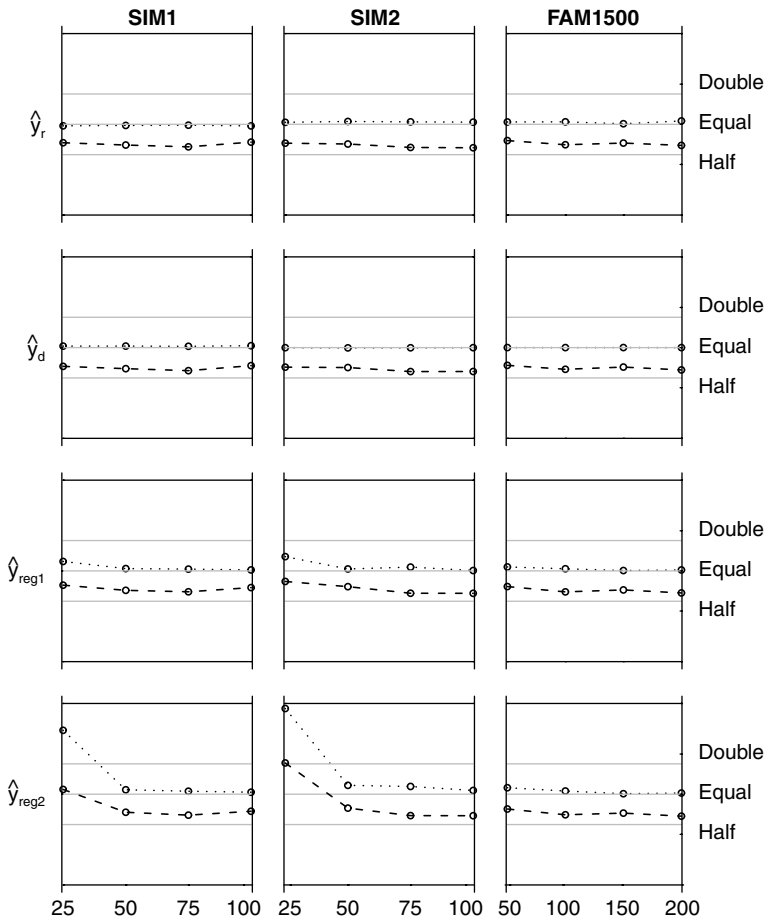

Fig. 2. Log ratios of standard errors comparing the *cases available* imputations and the *complete data* imputations against the *simple* imputation for the estimation of the population mean, $p = 0.4n$, $q = 0.48n$. The dotted curve corresponds to the *complete data* imputation and the dashed curve refers to the *cases available* imputation.

- Step 2: Set the missingness rates, p and q.
- Step 3: Eliminate the sample p elements on the auxiliary characteristic and q elements on the study characteristic, in a random way.
- Step 4: Define the subsamples s_1, s_2 and s_3.
- Step 5: Calculate: $\hat{\bar{y}}_{r1}$, $\hat{\bar{y}}_{r2}$, $\hat{\bar{y}}_{d1}$, $\hat{\bar{y}}_{d2}$, $\hat{\bar{y}}_{\mathrm{Reg11}}$, $\hat{\bar{y}}_{\mathrm{Reg21}}$, $\hat{\bar{y}}_{\mathrm{Reg12}}$, $\hat{\bar{y}}_{\mathrm{Reg22}}$.
- Step 6: Specify the complete data sets using the estimator defined in step 5.
- Step 7: Estimate the population mean and the population median based on these complete data sets.

- Step 8: Use the values obtained in 1000 items for the calculation of the mean squared errors of the estimators.

Specifically, sample sizes of 25, 50, 75 and 100 were taken for the simulated populations and 50, 100, 150 and 200 for the FAM1500 population, due to the larger size of the latter. In addition to this, it is interesting to note that the missingness rates were taken such that integer values were generated for all sample sizes.

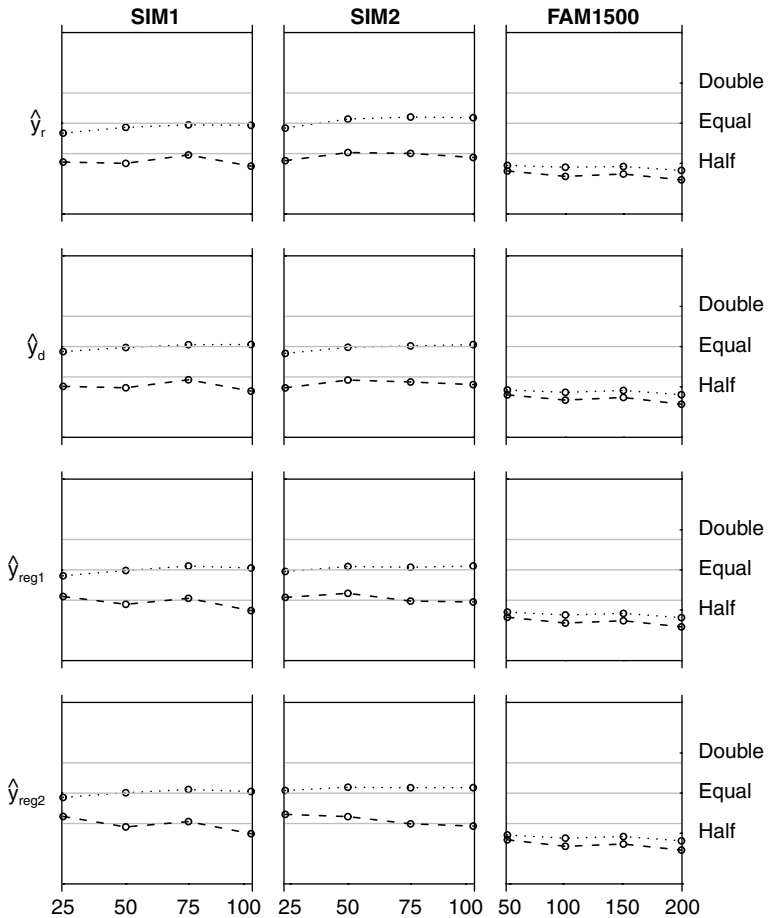

Fig. 3. Log ratios of standard errors comparing the *cases available* imputations and the *complete data* imputations against the *simple* imputation for the estimation of the population median, $p = 0.32n$, $q = 0.48n$. The dotted curve corresponds to the *complete data* imputation and the dashed curve refers to the *cases available* imputation.

Results of the application of this algorithm for some values of p and q can be seen in Figs. 1–4. Each figure plots the log ratios of the normalized standard errors of estimators (dividing their mean squared error by the mean squared error of the corresponding estimator based on the simple mean imputation) for the imputation methods considered. The dashed curves correspond to the proposed imputation (based on available observations) and the dotted curves refer to the imputation based on complete observations. The central horizontal lines correspond to the imputation based on the simple mean.

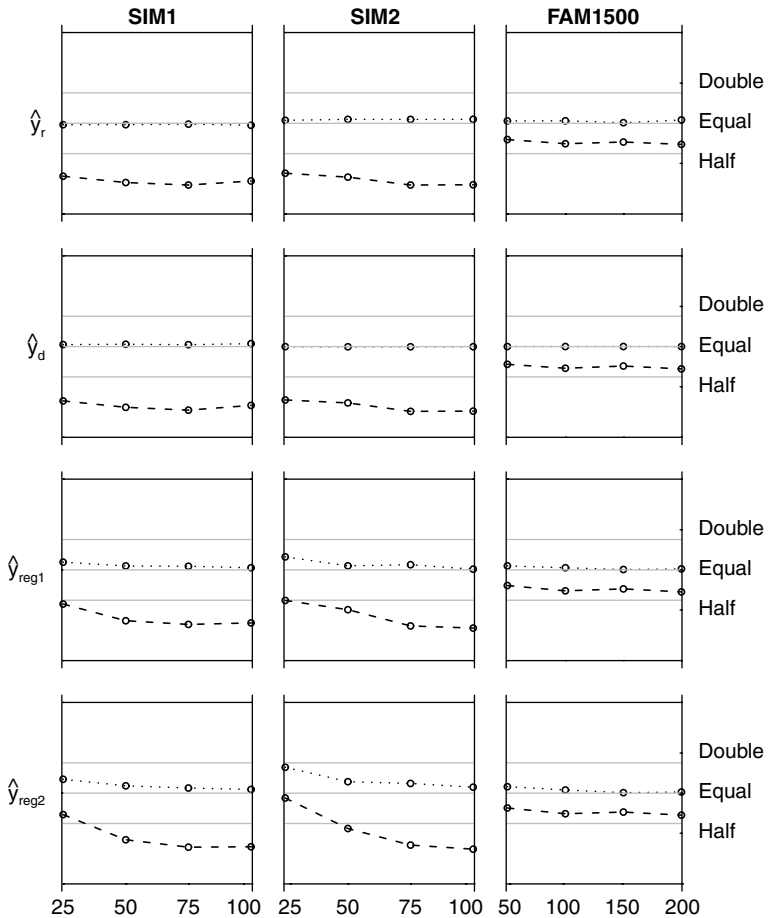

Fig. 4. Log ratios of standard errors comparing the *cases available* imputations and the *complete data* imputations against the *simple* imputation for the estimation of the population median, $p = 0.4n$, $q = 0.48n$. The dotted curve corresponds to the *complete data* imputation and the dashed curve refers to the *cases available* imputation.

Tables 1–3 show the error reduction values achieved, as a percentage:

Table 1
Error reduction as percentage for FAM1500 population

	$\tilde{y}_i =$	Mean				Median			
		$\hat{\bar{y}}_{r2}$	$\hat{\bar{y}}_{d2}$	$\hat{\bar{y}}_{\mathrm{Reg21}}$	$\hat{\bar{y}}_{\mathrm{Reg22}}$	$\hat{\bar{y}}_{r2}$	$\hat{\bar{y}}_{d2}$	$\hat{\bar{y}}_{\mathrm{Reg21}}$	$\hat{\bar{y}}_{\mathrm{Reg22}}$
	50	11.96	10.57	89.44	10.42	47.51	48.02	44.47	42.73
$p = 0.32n$	100	18.92	16.25	83.64	17.02	44.06	41.42	40.72	40.77
$q = 0.48n$	150	15.80	14.78	85.32	15.74	38.51	36.55	35.45	35.44
	200	19.41	19.57	80.52	19.12	40.45	40.01	38.65	38.75
	50	34.64	33.37	66.77	38.48	66.44	66.04	65.19	64.70
$p = 0.4n$	100	40.74	39.08	60.99	41.88	61.87	61.29	62.63	62.99
$q = 0.48n$	150	35.69	35.97	64.26	35.80	53.87	54.89	54.84	54.69
	200	42.76	39.94	60.04	41.02	54.84	54.13	54.23	54.35

Table 2
Error reduction as percentage for SIM1 population

	$\tilde{y}_i =$	Mean				Median			
		$\hat{\bar{y}}_{r2}$	$\hat{\bar{y}}_{d2}$	$\hat{\bar{y}}_{\mathrm{Reg21}}$	$\hat{\bar{y}}_{\mathrm{Reg22}}$	$\hat{\bar{y}}_{r2}$	$\hat{\bar{y}}_{d2}$	$\hat{\bar{y}}_{\mathrm{Reg21}}$	$\hat{\bar{y}}_{\mathrm{Reg22}}$
	25	14.04	19.88	85.55	16.57	48.36	54.79	37.72	35.19
$p = 0.32n$	50	16.41	20.68	83.24	19.85	56.08	60.13	53.78	54.20
$q = 0.48n$	75	8.60	13.47	90.98	11.87	49.71	55.12	52.21	51.91
	100	21.30	25.70	78.47	24.05	60.66	65.50	62.22	61.90
	25	32.02	37.04	67.21	74.09	69.15	72.38	61.57	55.39
$p = 0.4n$	50	36.08	40.29	63.69	40.14	73.33	76.31	71.41	70.84
$q = 0.48n$	75	39.03	42.64	60.65	42.08	75.13	77.58	73.50	74.10
	100	30.96	36.58	68.94	35.30	71.99	75.59	71.64	72.98

Table 3
Error reduction as percentage for SIM2 population

	$\tilde{y}_i =$	Mean				Median			
		$\hat{\bar{y}}_{r2}$	$\hat{\bar{y}}_{d2}$	$\hat{\bar{y}}_{\mathrm{Reg21}}$	$\hat{\bar{y}}_{\mathrm{Reg22}}$	$\hat{\bar{y}}_{r2}$	$\hat{\bar{y}}_{d2}$	$\hat{\bar{y}}_{\mathrm{Reg21}}$	$\hat{\bar{y}}_{\mathrm{Reg22}}$
	25	21.89	21.05	79.00	26.25	52.48	54.46	44.75	42.03
$p = 0.32n$	50	15.05	12.67	87.09	15.05	53.44	52.69	45.78	48.87
$q = 0.48n$	75	19.08	16.90	82.46	19.13	56.42	56.14	53.88	56.19
	100	24.11	22.53	77.46	23.43	59.68	59.92	55.99	58.43
	25	37.90	35.89	64.10	71.16	70.14	70.40	63.04	50.46
$p = 0.4n$	50	40.26	36.20	64.41	40.54	73.27	72.24	63.25	65.71
$q = 0.48n$	75	44.26	41.90	58.04	48.63	77.65	77.12	75.31	75.54
	100	44.30	41.98	58.40	43.86	77.58	77.01	73.96	75.73

$$\text{e.r.}(\hat{\theta}_2) = \left(1 - \frac{\text{MSE}(\hat{\theta}_2)}{\text{MSE}(\hat{\theta}_1)}\right)\%,$$

where $\hat{\theta}_2$ are the estimators of the parameter θ obtained when estimators based on available cases are used for the imputation, and $\hat{\theta}_1$ are the estimators of the parameter θ obtained when estimators based on complete cases are used.

In the three populations considered, all the estimates of the mean and the median based on the cases available imputation present a smaller error than the respective estimators based on the complete case imputation. In specific terms, the error reductions range from 8.5% to 95.5% in the estimation of the mean, and from 35.4% to 77.6% in the estimation of the median.

The estimators obtained with imputation based on the available cases always (except $n = 25$, for median estimation) improved considerably on the results provided by those based on mean imputation, this error reduction being greater in the case of the median estimation.

As expected, when the total missingness rate $\frac{p+q}{n}$ increases, the gain in the precision of the proposed estimators is greater.

In conclusion, we found evidence that greater efficiency can be obtained by using the proposed method of imputation.

4. Conclusions

For many years, studies concerning the sampling of finite populations have been aimed at determining optimal strategies, on the one hand, to select a sampling design adapted to the population, and on the other, to obtain a suitable estimator to be used with such a sampling design. All these studies have been based on the assumption that, for all the units selected in the sample, information is available concerning all the variables considered and that this information is free of errors.

However, time has revealed the impracticability of these studies, as they do not consider the possibility that information might not be obtained for some of the individuals selected in the sample.

This is why, in the last few years, a change has occurred concerning the focus of studies carried out in the field of sampling among finite populations. Although several ways of dealing with nonresponse have been attempted, the most fully developed has been the automatic imputation of data. Various authors have concentrated on defining efficient imputation techniques in order to obtain a matrix of the complete data and thus apply all the results of strategy optimality. One such imputation method is the imputation of the mean, which is frequently used due to its simplicity. In the present study, we propose using the estimations obtained by various indirect methods, such as imputed

values, which are subsequently modified in order to use all the information provided by the sample.

The positive qualities of these estimators lead us to believe that the proposed imputation techniques will produce increased efficiency with respect to traditional methods.

After carrying out a complete simulation study, we conclude that although the methods for imputation of the mean that apply ratio, difference and regression estimators of complete cases present no obvious advantages over the classical method for imputation of the mean, the pattern changes considerably if we take into account the cases when part of the data is missing from the indirect estimations of the mean. The proposed imputation methods are a little more complex to apply, but they produce an increase in efficiency that is considerable in the estimation of such important parameters as the mean and the median.

Acknowledgement

This research was partially supported by MEC contract number MTM2004-04038.

Appendix A

Optimal coefficients $\alpha_{r_{\text{opt}}}$, $\beta_{r_{\text{opt}}}$, $\alpha_{d_{\text{opt}}}$, $\beta_{r_{\text{opt}}}$, $\alpha_{\text{reg}_{\text{opt}}}$ and $\beta_{\text{reg}_{\text{opt}}}$ of proposed estimators are:

$$\alpha_{r_{\text{opt}}} = \frac{-C_r + \left(E_r B_r - \frac{C_r}{A_r} B_r^2\right)/(D_r - B_r^2/A_r)}{A_r}, \tag{A.1}$$

$$\beta_{r_{\text{opt}}} = \frac{-E_r + \frac{C_r}{A_r} B_r}{D_r - B_r^2/A_r}, \tag{A.2}$$

$$\alpha_{d_{\text{opt}}} = \frac{A_d - \frac{C_d D_d - A_d B_d}{E_d C_d - B_d^2} B_d}{C_d}, \tag{A.3}$$

$$\beta_{d_{\text{opt}}} = \frac{C_d D_d - A_d B_d}{E_d C_d - B_d^2}, \tag{A.4}$$

$$\alpha_{\text{reg}_{\text{opt}}} = \frac{-C_{\text{reg}}}{A_{\text{reg}}} - \frac{B_{\text{reg}}}{A_{\text{reg}}} \frac{B_{\text{reg}} C_{\text{reg}} - A_{\text{reg}} E_{\text{reg}}}{A_{\text{reg}} D_{\text{reg}} - B_{\text{reg}}^2}, \tag{A.5}$$

$$\beta_{\mathrm{reg}_{\mathrm{opt}}} = \frac{B_{\mathrm{reg}}C_{\mathrm{reg}} - A_{\mathrm{reg}}E_{\mathrm{reg}}}{A_{\mathrm{reg}}D_{\mathrm{reg}} - B_{\mathrm{reg}}^2}, \tag{A.6}$$

where

$$\begin{aligned}
A_r &= 2R^2\,\mathrm{Var}(\hat{\bar{x}}^2_{\mathrm{HT}}) + 2R^2\,\mathrm{Var}(\hat{\bar{x}}^1_{\mathrm{HT}}) - 4R^2\,\mathrm{Cov}(\hat{\bar{x}}^2_{\mathrm{HT}}, \hat{\bar{x}}^1_{\mathrm{HT}}),\\
B_r &= -2R\,\mathrm{Cov}(\hat{\bar{y}}^3_{\mathrm{HT}}, \hat{\bar{x}}^2_{\mathrm{HT}}) + 2R\,\mathrm{Cov}(\hat{\bar{y}}^3_{\mathrm{HT}}, \hat{\bar{x}}^1_{\mathrm{HT}}) + 2R\,\mathrm{Cov}(\hat{\bar{y}}^1_{\mathrm{HT}}, \hat{\bar{x}}^2_{\mathrm{HT}}) - 2R\,\mathrm{Cov}(\hat{\bar{y}}^1_{\mathrm{HT}}, \hat{\bar{x}}^1_{\mathrm{HT}}),\\
C_r &= -2R^2\,\mathrm{Var}(\hat{\bar{x}}^1_{\mathrm{HT}}) + 2R^2\,\mathrm{Cov}(\hat{\bar{x}}^2_{\mathrm{HT}}, \hat{\bar{x}}^1_{\mathrm{HT}}) - 2R\,\mathrm{Cov}(\hat{\bar{y}}^1_{\mathrm{HT}}, \hat{\bar{x}}^2_{\mathrm{HT}}) + 2R\,\mathrm{Cov}(\hat{\bar{y}}^1_{\mathrm{HT}}, \hat{\bar{x}}^1_{\mathrm{HT}}),\\
D_r &= 2\,\mathrm{Var}(\hat{\bar{y}}^3_{\mathrm{HT}}) + 2\,\mathrm{Var}(\hat{\bar{y}}^1_{\mathrm{HT}}) - 4\,\mathrm{Cov}(\hat{\bar{y}}^3_{\mathrm{HT}}, \hat{\bar{y}}^1_{\mathrm{HT}}),\\
E_r &= -2\,\mathrm{Var}(\hat{\bar{y}}^1_{\mathrm{HT}})2\,\mathrm{Cov}(\hat{\bar{y}}^3_{\mathrm{HT}}, \hat{\bar{y}}^1_{\mathrm{HT}}) - 2R\,\mathrm{Cov}(\hat{\bar{y}}^3_{\mathrm{HT}}, \hat{\bar{x}}^1_{\mathrm{HT}}) + 2R\,\mathrm{Cov}(\hat{\bar{y}}^1_{\mathrm{HT}}, \hat{\bar{x}}^1_{\mathrm{HT}}),\\
A_d &= \mathrm{Var}(\hat{\bar{y}}^3_{\mathrm{HT}}) - \mathrm{Cov}(\hat{\bar{y}}^1_{\mathrm{HT}}, \hat{\bar{y}}^3_{\mathrm{HT}}) + \mathrm{Cov}(\hat{\bar{y}}^1_{\mathrm{HT}}, \hat{\bar{x}}^2_{\mathrm{HT}}) - \mathrm{Cov}(\hat{\bar{y}}^3_{\mathrm{HT}}, \hat{\bar{x}}^2_{\mathrm{HT}}),\\
B_d &= -\mathrm{Cov}(\hat{\bar{y}}^1_{\mathrm{HT}}, \hat{\bar{x}}^1_{\mathrm{HT}}) + \mathrm{Cov}(\hat{\bar{y}}^3_{\mathrm{HT}}, \hat{\bar{x}}^1_{\mathrm{HT}}) - \mathrm{Cov}(\hat{\bar{y}}^3_{\mathrm{HT}}, \hat{\bar{x}}^2_{\mathrm{HT}}) + \mathrm{Cov}(\hat{\bar{y}}^1_{\mathrm{HT}}, \hat{\bar{x}}^2_{\mathrm{HT}}),\\
C_d &= \mathrm{Var}(\hat{\bar{y}}^1_{\mathrm{HT}}) + \mathrm{Var}(\hat{\bar{y}}^3_{\mathrm{HT}}) - 2\,\mathrm{Cov}(\hat{\bar{y}}^1_{\mathrm{HT}}, \hat{\bar{y}}^3_{\mathrm{HT}}),\\
D_d &= \mathrm{Var}(\hat{\bar{x}}^2_{\mathrm{HT}}) - \mathrm{Cov}(\hat{\bar{x}}^1_{\mathrm{HT}}, \hat{\bar{x}}^2_{\mathrm{HT}}) + \mathrm{Cov}(\hat{\bar{y}}^3_{\mathrm{HT}}, \hat{\bar{x}}^1_{\mathrm{HT}}) - \mathrm{Cov}(\hat{\bar{y}}^3_{\mathrm{HT}}, \hat{\bar{x}}^2_{\mathrm{HT}}),\\
E_d &= \mathrm{Var}(\hat{\bar{x}}^2_{\mathrm{HT}}) + \mathrm{Var}(\hat{\bar{x}}^1_{\mathrm{HT}}) - 2\,\mathrm{Cov}(\hat{\bar{x}}^1_{\mathrm{HT}}, \hat{\bar{x}}^2_{\mathrm{HT}}),\\
A_{\mathrm{reg}} &= 2\,\mathrm{Var}(\hat{\bar{y}}^1_{\mathrm{HT}}) + 2\,\mathrm{Var}(\hat{\bar{y}}^3_{\mathrm{HT}}) - 4\,\mathrm{Cov}(\hat{\bar{y}}^1_{\mathrm{HT}}, \hat{\bar{y}}^3_{\mathrm{HT}}),\\
B_{\mathrm{reg}} &= 2b[-\mathrm{Cov}(\hat{\bar{y}}^1_{\mathrm{HT}}, \hat{\bar{x}}^1_{\mathrm{HT}}) + \mathrm{Cov}(\hat{\bar{y}}^1_{\mathrm{HT}}, \hat{\bar{x}}^2_{\mathrm{HT}}) + \mathrm{Cov}(\hat{\bar{y}}^3_{\mathrm{HT}}, \hat{\bar{x}}^1_{\mathrm{HT}}) - \mathrm{Cov}(\hat{\bar{y}}^3_{\mathrm{HT}}, \hat{\bar{x}}^2_{\mathrm{HT}})],\\
C_{\mathrm{reg}} &= -2\,\mathrm{Var}(\hat{\bar{y}}^3_{\mathrm{HT}}) + 2\,\mathrm{Cov}(\hat{\bar{y}}^1_{\mathrm{HT}}, \hat{\bar{y}}^3_{\mathrm{HT}}) + 2b[-\mathrm{Cov}(\hat{\bar{y}}^1_{\mathrm{HT}}, \hat{\bar{x}}^2_{\mathrm{HT}}) + \mathrm{Cov}(\hat{\bar{y}}^3_{\mathrm{HT}}, \hat{\bar{x}}^2_{\mathrm{HT}})],\\
D_{\mathrm{reg}} &= b^2[2\,\mathrm{Var}(\hat{\bar{x}}^1_{\mathrm{HT}}) + 2\,\mathrm{Var}(\hat{\bar{x}}^2_{\mathrm{HT}}) - 4\,\mathrm{Cov}(\hat{\bar{x}}^1_{\mathrm{HT}}, \hat{\bar{x}}^2_{\mathrm{HT}})],\\
E_{\mathrm{reg}} &= -2b^2\,\mathrm{Var}(\hat{\bar{x}}^2_{\mathrm{HT}}) + 2b^2\,\mathrm{Cov}(\hat{\bar{x}}^1_{\mathrm{HT}}, \hat{\bar{x}}^2_{\mathrm{HT}}) - 2b\,\mathrm{Cov}(\hat{\bar{y}}^3_{\mathrm{HT}}, \hat{\bar{x}}^1_{\mathrm{HT}}) + 2b\,\mathrm{Cov}(\hat{\bar{y}}^3_{\mathrm{HT}}, \hat{\bar{x}}^2_{\mathrm{HT}}).
\end{aligned} \tag{A.7}$$

References

[1] J.M. Brick, G. Kalton, Handling missing data in survey research, Statistical Methods in Medical Research 5 (1996) 215–238.

[2] G. King, J. Honaker, A. Joseph, K. Scheve, Listwise deletion is evil: what to do about missing data in Political Science, 78 (1996), Unpublished document.

[3] F.R. Fernández, J.A. Mayor, Muestreo en poblaciones finitas: curso básico, Ed. PPU, 1994.

[4] R.J.A. Little, D.B. Rubin, Statistical Analysis with Missing Data, John Wiley, New York, 1987.

[5] G. Meeden, Median estimation using auxiliary information, Survey Methodology 21 (1995) 71–77.

[6] R.H. Randles, On the asymptotic normality of statistics with estimated parameters, The Annals of Statistics 10 (1982) 462–474.

[7] M. Rueda, S. González, Missing data and auxiliary information in surveys, Computational Statistic 4 (2004).

[8] J.L. Schafer, Analysis of Incomplete Multivariate Data, Chapman and Hall, London, 1997.

[9] H. Toutenburg, V.K. Srivastava, Estimation of ratio of population means in survey sampling when some observations are missing, Metrika 48 (1998) 177–187.