Computational Statistics (2004) 19:551–567

Computational Statistics

Missing data and auxiliary information in surveys

M. Rueda[1], S. González[2]

[1] Department of Statistics & OR, University of Granada, Spain
[2] Department of Statistics, University of Jaén, Spain

Summary

This paper proposes estimation methods with auxiliary information when some observations are missing from the sample. These ratio, difference and regression methods are proposed for any sampling design and are compared with other complete case estimators.

Keywords: Auxiliary information, missing data, Horvitz-Thompson estimator.

1 Introduction

The infeasibility of having all the observations in a sample is not an uncommon aspect of data collection in many instances of sample surveys. Missing data occur in survey research because an element in the target population is not included in the survey sampling frame (noncoverage), because a sample element has not participated in the survey (total nonresponse) or because a responding sample element fails to provide an acceptable response to one or more of the survey items (item nonresponse). This latter type of nonresponse is a common occurrence and may arise for different reasons (a respondent refuses to answer an item, does not know the answer to the item, gives an answer that is inconsistent with answers to other items, the interviewer fails to ask the question or record the answer, etc.) Unfortunately, the problem of missing data arises frequently in practice.

One obvious consequence of nonresponse is that the actual sample size is less than the planned one. This can produce biases in estimations if nonrespondents differ from respondents on the characteristic of interest, and also lead to greater sampling variance.

There exist different methods to handle missing data during the stages of data collection and processing. The aim of these methods is to obtain a precise and complete data set. Nevertheless, it is still possible to find errors and losses of some entries even after the data has been collected and filtered.

When some observations in the sample are missing (item nonresponse), a first option would be to carry out a complete case analysis. Methods based on completely recorded units create a rectangular data set by discarding all observations with any missing variable. Thus, when parameters are estimated, only the observations for which all the variables of interest have a valid value are used. Little and Rubin (1987) pointed out the statistical shortcoming of all the methods that ignore incomplete observations. While these methods can provide satisfactory results 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. (1998) illustrate how methods of complete cases are prone to serious errors. To sum up, this practice can be said to introduce a bias into the estimate and an increase in sampling variance due to a reduction in sample size, see, e.g., Brick and Kalton (1969), Schafer (1997).

Alternatively, an imputation method may be used to find substitutes for missing observations, see, e.g., Little and Rubin (1987), Särndal (1992) and Rubin (1987) for an interesting account. Certain commonly used imputation methods take the imputed values as true observations, and the statistical analysis may be carried out using the standard procedures developed for data without any missing observations. Such a practice, it is well recognized, may tend to invalidate the inferences and may often have serious consequences. Some statistics specialists are reluctant to apply this method because it manipulates the original information, although there are also reasons to justify its use. Other procedures such as the multiple imputation and the model-assisted approaches account for the fact that imputed values are not true observations, as they reflect the additional variance due to imputation error.

As a third option, we could try to improve the precision of the estimators by including all the cases available for their calculation.

Indirect estimation methods are easily comprehensible techniques for the estimation of total population in survey sampling when an auxiliary characteristic correlated with the study characteristic is available; see, e.g., Sukhatme, Sukhatme, Sukhatme and Asok (1984). These techniques provide generally biased but more efficient estimators in comparison with the traditional unbiased estimator. These methods of estimation assume that the sample data contain no missing observations. This specification may not be tenable in many practical applications, see, e.g., Rubin (1977). Some authors have defined indirect estimators when the sample is drawn according to the procedure of simple random sampling without replacement when some observations are missing, see, e.g., Tracy and Osahan (1994) and Toutenburg and Srivastava (1998, 1999, 2000). However, there appears to be no investigation reported in the literature when another sample design is used, and this is the main concern of the present paper. In this article, therefore, we consider the indirect estimation of total population on the basis of a random sample drawn according to any sample design. Using the methods of ratio, difference and regression estimation, we propose estimators for the population total of study characteristics besides the conventional estimators which amputate incomplete observations.

This article is structured as follows: in section 2 we present estimators for the total population which are better, in the sense of precision, than traditional estimators. Section 3 considers estimator properties through a simulation study in the case of simple random sampling without replacement.

Lastly, in the Appendix, the problem is developed for the case of simple random sampling without replacement and for the case of stratified sampling.

2 Proposed estimators

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 the values of two variables, (y_i, x_i), $i = 1, \ldots, n$, are observed for the estimation of the total population, Y.

It is assumed that a set of $(n - p - q)$ complete observations on selected units in the sample are available. In addition to these, observations on the x characteristic on p units in the sample are available but the corresponding observations on the y characteristic are missing. Similarly, we have a set of q observations on the y characteristic in the sample but the associated values on the x characteristic are missing. Further, p and q are assumed to be integer numbers verifying $0 < p, q < n/2$.

This population has the following structure:

y_1	$\dots$	y_{n-p-q}	Missing	$\dots$	Missing	y_{n-q+1}	$\dots$	y_n
x_1	$\dots$	x_{n-p-q}	$x_{n-p-q+1}$	$\dots$	x_{n-q}	Missing	$\dots$	Missing

For the sake of simplicity, we separate the unit of the sample s into three disjoint sets:

$$\begin{aligned} s_1 &= \{i \in s/x_i, y_i \text{ are available}\} \\ s_2 &= \{i \in s/x_i \text{ are available, but } y_i \text{ is not}\} \\ s_3 &= \{i \in s/y_i \text{ are available, but } x_i \text{ is not}\} \end{aligned}$$

If we write:

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

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

$$\hat{y}_{r1} = \frac{\widehat{y}^1_{HT}}{\widehat{x}^1_{HT}} = \frac{\sum\limits_{i \in s_1} \frac{y_i}{\pi_i}}{\sum\limits_{i \in s_1} \frac{x_i}{\pi_i}} * X \tag{1}$$

$$\widehat{y}_{d1} = \widehat{y}^1_{HT} + (X - \widehat{x}^1_{HT}) \tag{2}$$

$$\hat{y}_{Reg1} = \widehat{y}^1_{HT} + b(X - \widehat{x}^1_{HT})) \tag{3}$$

where b can be fixed and known or unknown. In this latter case, if the error is minimized we obtain that:

$$b = \frac{\text{Cov}(x, y)}{\text{Var}(x)}$$

which is what must be estimated.

All these estimators discard the information available on incomplete cases. This practice can introduce biases and errors into the estimation. For this reason, we propose the following classes of estimators, which incorporate all the available observations:

$$\hat{y}^*_{r2} = \frac{\alpha_r \widehat{y}^3_{HT} + (1 - \alpha_r)\widehat{y}^1_{HT}}{\beta_r \widehat{x}^2_{HT} + \beta_r \widehat{x}^1_{HT}} * X \tag{4}$$

$$\widehat{y}^*_{d2} = \alpha_d \widehat{y}^1_{HT} + (1-\alpha_d)\widehat{y}^3_{HT} + (X - (\beta_d \widehat{x}^1_{HT} + (1-\beta_d)\widehat{x}^2_{HT})) \qquad (5)$$

$$\hat{y}^*_{Reg2} = \alpha_{reg}\hat{y}^1_{HT} + (1-\alpha_{reg})\hat{y}^3_{HT} + b\left[X - (\beta_{reg}\hat{x}^1_{HT} + (1-\beta_{reg})\hat{x}^2_{HT})\right] \quad (6)$$

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

$$\hat{b}_1 = \frac{\widehat{\text{Cov}}_{i\in s_1}(x,y)}{\widehat{\text{Var}}_{i\in s_1}(x)} \qquad (7)$$

$$\hat{b}_2 = \frac{\widehat{\text{Cov}}_{i\in s_1}(x,y)}{\widehat{\text{Var}}_{i\in s_1 \bigcup s_2}(x)} \qquad (8)$$

where $\widehat{\text{Cov}}_{i\in s_1}(x,y)$, $\widehat{\text{Var}}_{i\in s_1}$ and $\widehat{\text{Var}}_{i\in s_1 \bigcup s_2}$ represent the variances and covariances based on the corresponding subsamples. Using these estimations of b, we can define the classes of regression estimators $\hat{y}^*_{Reg21}$ and $\hat{y}^*_{Reg22}$ by replacing the value of b with that of its respective estimation.

Note that 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. This choice is made seeking to minimize the estimator error. The expressions of the mean squared errors of the estimators are easily obtained, and by minimizing these errors, we obtain the estimator expressions with minimum errors.

Thus we have:

$$
\begin{aligned}
\alpha_{r_{opt}} &= \frac{-C_r + (E_r B_r - \frac{C_r}{A_r} B_r^2)/(D_r - B_r^2/A_r)}{A_r} \\
\beta_{r_{opt}} &= \frac{-E_r + \frac{C_r}{A_r} B_r}{D_r - B_r^2/A_r} \\
\alpha_{d_{opt}} &= \frac{A_d - \frac{C_d D_d - A_d B_d}{E_d C_d - B_d^2} B_d}{C_d} \\
\beta_{d_{opt}} &= \frac{C_d D_d - A_d B_d}{E_d C_d - B_d^2} \\
\alpha_{reg_{opt}} &= \frac{-C_{reg}}{A_{reg}} - \frac{B_{reg}}{A_{reg}} \frac{B_{reg} C_{reg} - A_{reg} E_{reg}}{A_{reg} D_{reg} - B_{reg}^2} \\
\beta_{reg_{opt}} &= \frac{B_{reg} C_{reg} - A_{reg} E_{reg}}{A_{reg} D_{reg} - B_{reg}^2}
\end{aligned}
$$

where:

$$
\begin{aligned}
A_r &= 2R^2 \operatorname{Var}(\widehat{x}^2_{HT}) + 2R^2 \operatorname{Var}(\widehat{x}^1_{HT}) - 4R^2 \operatorname{Cov}(\widehat{x}^2_{HT}, \widehat{x}^1_{HT}) \\
B_r &= -2R \operatorname{Cov}(\widehat{y}^3_{HT}, \widehat{x}^2_{HT}) + 2R \operatorname{Cov}(\widehat{y}^3_{HT}, \widehat{x}^1_{HT}) + \\
&\quad 2R \operatorname{Cov}(\widehat{y}^1_{HT}, \widehat{x}^2_{HT}) - 2R \operatorname{Cov}(\widehat{y}^1_{HT}, \widehat{x}^1_{HT}) \\
C_r &= -2R^2 \operatorname{Var}(\widehat{x}^1_{HT}) + 2R^2 \operatorname{Cov}(\widehat{x}^2_{HT}, \widehat{x}^1_{HT}) - \\
&\quad 2R \operatorname{Cov}(\widehat{y}^1_{HT}, \widehat{x}^2_{HT}) + 2R \operatorname{Cov}(\widehat{y}^1_{HT}, \widehat{x}^1_{HT}) \\
D_r &= 2 \operatorname{Var}(\widehat{y}^3_{HT}) + 2 \operatorname{Var}(\widehat{y}^1_{HT}) - 4 \operatorname{Cov}(\widehat{y}^3_{HT}, \widehat{y}^1_{HT}) \\
E_r &= -2 \operatorname{Var}(\widehat{y}^1_{HT}) + 2 \operatorname{Cov}(\widehat{y}^3_{HT}, \widehat{y}^1_{HT}) - \\
&\quad 2R \operatorname{Cov}(\widehat{y}^3_{HT}, \widehat{x}^1_{HT}) + 2R \operatorname{Cov}(\widehat{y}^1_{HT}, \widehat{x}^1_{HT}) \\
A_d &= \operatorname{Var}(\widehat{y}^3_{HT}) - \operatorname{Cov}(\widehat{y}^1_{HT}, \widehat{y}^3_{HT}) + \operatorname{Cov}(\widehat{y}^1_{HT}, \widehat{x}^2_{HT}) - \operatorname{Cov}(\widehat{y}^3_{HT}, \widehat{x}^2_{HT}) \\
B_d &= -\operatorname{Cov}(\widehat{y}^1_{HT}, \widehat{x}^1_{HT}) + \operatorname{Cov}(\widehat{y}^1_{HT}, \widehat{x}^2_{HT}) + \\
&\quad \operatorname{Cov}(\widehat{y}^3_{HT}, \widehat{x}^1_{HT}) - \operatorname{Cov}(\widehat{y}^3_{HT}, \widehat{x}^2_{HT}) \\
C_d &= \operatorname{Var}(\widehat{y}^1_{HT}) + \operatorname{Var}(\widehat{y}^3_{HT}) - 2 \operatorname{Cov}(\widehat{y}^1_{HT}, \widehat{y}^3_{HT}) \\
D_d &= \operatorname{Var}(\widehat{x}^2_{HT}) - \operatorname{Cov}(\widehat{x}^1_{HT}, \widehat{x}^2_{HT}) + \operatorname{Cov}(\widehat{y}^3_{HT}, \widehat{x}^1_{HT}) - \operatorname{Cov}(\widehat{y}^3_{HT}, \widehat{x}^2_{HT}) \\
E_d &= \operatorname{Var}(\widehat{x}^2_{HT}) + \operatorname{Var}(\widehat{x}^1_{HT}) - 2 \operatorname{Cov}(\widehat{x}^1_{HT}, \widehat{x}^2_{HT}) \\
A_{reg} &= 2 \operatorname{Var}(\hat{y}^1_{HT}) + 2 \operatorname{Var}(\hat{y}^3_{HT}) - 4 \operatorname{Cov}(\hat{y}^1_{HT}, \hat{y}^3_{HT}) \\
B_{reg} &= 2b \left[-\operatorname{Cov}(\hat{y}^1_{HT}, \hat{x}^1_{HT}) + \operatorname{Cov}(\hat{y}^1_{HT}, \hat{x}^2_{HT}) + \right. \\
&\quad \left. \operatorname{Cov}(\hat{y}^3_{HT}, \hat{x}^1_{HT}) - \operatorname{Cov}(\hat{y}^3_{HT}, \hat{x}^2_{HT}) \right] \\
C_{reg} &= -2 \operatorname{Var}(\hat{y}^3_{HT}) + 2 \operatorname{Cov}(\hat{y}^1_{HT}, \hat{y}^3_{HT}) + \\
&\quad 2b \left[-\operatorname{Cov}(\hat{y}^1_{HT}, \hat{x}^2_{HT}) + \operatorname{Cov}(\hat{y}^3_{HT}, \hat{x}^2_{HT}) \right]
\end{aligned}
$$

$$\begin{aligned} D_{reg} &= b^2 \left[2\,\mathrm{Var}(\hat{x}^1_{HT}) + 2\,\mathrm{Var}(\hat{x}^2_{HT}) - 4\,\mathrm{Cov}(\hat{x}^1_{HT}, \hat{x}^2_{HT})\right] \\ E_{reg} &= -2b^2\,\mathrm{Var}(\hat{x}^2_{HT}) + 2b^2\,\mathrm{Cov}(\hat{x}^1_{HT}, \hat{x}^2_{HT}) - \\ &\quad 2b\,\mathrm{Cov}(\hat{y}^3_{HT}, \hat{x}^1_{HT}) + 2b\,\mathrm{Cov}(\hat{y}^3_{HT}, \hat{x}^2_{HT}) \end{aligned}$$

The expressions of these variances and covariances for the case of simple random sampling without replacement and for the case of stratified sampling can be seen in the Appendix.

Unfortunately these optimum values depend on theoretical variances and covariances among the Horvitz-Thompson estimators, which are generally unknown, so the optimal estimator cannot be used. However, they can be estimated when the sample is drawn. Furthermore, these values would be estimated by replication methods, see, e.g., Wolter (1985).

In the absence of good a priori knowledge of these characteristics, we replace the optimal α and β-values by sample based estimates in 4, 5 and 6 thus obtaining the following estimators, which can be evaluated from the sample obtained:

$$\hat{y}_{r2} = \frac{\widehat{\alpha}_r \widehat{y}^3_{HT} + (1 - \widehat{\alpha}_r)\widehat{y}^1_{HT}}{\widehat{\beta}_r \widehat{x}^2_{HT} + \widehat{\beta}_r \widehat{x}^1_{HT}} * X \tag{9}$$

$$\widehat{y}_{d2} = \widehat{\alpha}_d \widehat{y}^1_{HT} + (1 - \widehat{\alpha}_d)\widehat{y}^3_{HT} + (X - (\widehat{\beta}_d \widehat{x}^1_{HT} + (1 - \widehat{\beta}_d)\widehat{x}^2_{HT})) \tag{10}$$

$$\hat{y}_{Reg2} = \widehat{\alpha}_{reg}\hat{y}^1_{HT} + (1 - \widehat{\alpha}_{reg})\hat{y}^3_{HT} + b\left[X - (\widehat{\beta}_{reg}\hat{x}^1_{HT} + (1 - \widehat{\beta}_{reg})\hat{x}^2_{HT})\right] \tag{11}$$

$$\hat{y}_{Reg21} = \widehat{\alpha}_{reg}\hat{y}^1_{HT} + (1 - \widehat{\alpha}_{reg})\hat{y}^3_{HT} + \hat{b}_1\left[X - (\widehat{\beta}_{reg}\hat{x}^1_{HT} + (1 - \widehat{\beta}_{reg})\hat{x}^2_{HT})\right] \tag{12}$$

$$\hat{y}_{Reg22} = \widehat{\alpha}_{reg}\hat{y}^1_{HT} + (1 - \widehat{\alpha}_{reg})\hat{y}^3_{HT} + \hat{b}_2\left[X - (\widehat{\beta}_{reg}\hat{x}^1_{HT} + (1 - \widehat{\beta}_{reg})\hat{x}^2_{HT})\right] \tag{13}$$

These estimators do not coincide with the theoretical estimators in expressions 4, 5 and 6 and involve the estimated parameters. Randles (1982) derived the limit distribution for such statistics. Following his notation, we denote the estimator $\widehat{y}_{d2}$ as $T_n(\widehat{\lambda})$ with $\widehat{\lambda} = (\widehat{\alpha}_d, \widehat{\beta}_d)$. We replace $\widehat{\lambda}$ in $T_n(\cdot)$ with a variable ς. Now we calculate the limit of the expectation of the statistic $T_n(\lambda)$ when the current value of the parameter is $\lambda = (\alpha_d, \beta_d)$

$$\mu(\lambda) = \lim E_\lambda(T_n(\varsigma)) = Y$$

where E_λ denotes the expectation with respect to the design.

Since $\mu(\cdot)$ has partial derivates on $\varsigma = \lambda$ equal to zero, it now follows from Randless (1982) that $T_n(\widehat{\lambda})$ and $T_n(\lambda)$ have the same limit distribution, i.e., $\widehat{y}_{d2}$ has the same limit distribution as $\widehat{y}^*_{d2}$ with $\alpha_{d_{opt}}$ and $\beta_{d_{opt}}$ 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 the sense of mean square error, than the traditional ones.

3 Simulation study

This section examines estimator properties by means of a simulation study.

The populations considered can be divided into two groups: natural populations and simulated populations.

The FAM1500 population consists of 1500 families in Andalusia (Spain) taken from Fernández and Mayor (1994). The variable of interest, y, denotes family income and the auxiliary x denotes expenditure on food and drink.

The second class includes three simulated populations used by Meeden (1995). For the simulation, a superpopulation model is considered in which it is assumed that for each i, $y_i = bx_i + u_i e_i$, in which e_i are independent identically distributed random variables with zero expectations.

In the first population, SIM1, the x_i's form a random sample from a gamma distribution with a shape parameter of twenty and a scale parameter of one.

In the second population, SIM2, the auxiliary variable is a random sample from a log-normal population with mean and standard deviation 4.9 and 0.586 respectively.

In SIM3 the auxiliary variable is fifty plus a random sample from the standard exponential distribution.

All the simulated populations contain 500 units.

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

Algorithm

- STEP 1: Take a sample of size n according to the procedure of simple random sampling without replacement.
- 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: $\hat{y}_{r1}$, $\hat{y}_{r2}$, $\hat{y}_{Reg1}$, $\hat{y}_{Reg2}$, $\hat{y}_{Reg11}$, $\hat{y}_{Reg21}$, $\hat{y}_{Reg12}$, $\hat{y}_{Reg22}$, $\hat{y}_{d1}$, $\hat{y}_{d2}$
- STEP 6: Use the values obtained in 1000 items for the calculation of the mean squared errors of the estimators.
- STEP 7: Normalize these mean squared errors, dividing them by the mean squared error of the simple estimator and latter on take the log ratios of these mean squared errors.

Results of the application of this algorithm for some values of p and q can be seen in figures 1, 2 and 3.

In each figure are being plotted the log ratios of standard errors of considered estimators. The dashed curves correspond to the proposed estimator and the dotted curves refer to the estimator based on complete observations. The central horizontal lines correspond to the simple estimators.

It is interesting to note that the missingness rates were taken such that integer values were generated for all sampling sizes.

In the FAM1500 population, all the estimates based on the cases available present a smaller error than the respective estimators based on the complete data. The results obtained from the latter, in general, are no better than those based on the simple estimator, which does not make use of auxiliary data, while those proposed in this paper all present a smaller error than when the simple error is used as the basis for comparison.

A similar pattern was observed in the artificial populations SIM2 and SIM3. The estimators based on the available cases always improved considerably on the results provided by those based on the complete data, and were nearly always better than those based on the simple estimator.

A noteworthy feature is that in the SIM1 population the results obtained with the difference estimator, based on the complete cases, were very bad (the error was more than twice that obtained with the baseline estimator). Nevertheless, the error of the proposed estimator $\hat{y}_{d2}$ was only a fifth of that provided by the $\hat{y}_{d1}$ estimator and less than half that of the direct estimator $\hat{y}$, for any of the sample sizes considered. In this population, the ratio and

Figure 1: Log ratios of standar errors comparing the *cases available* estimators and the *complete data* estimators against the *simple* estimator, p=0.32n q=0.4n. The dotted curve corresponds to the *complete data* estimator and the dashed curve refers to the *cases available* estimator.

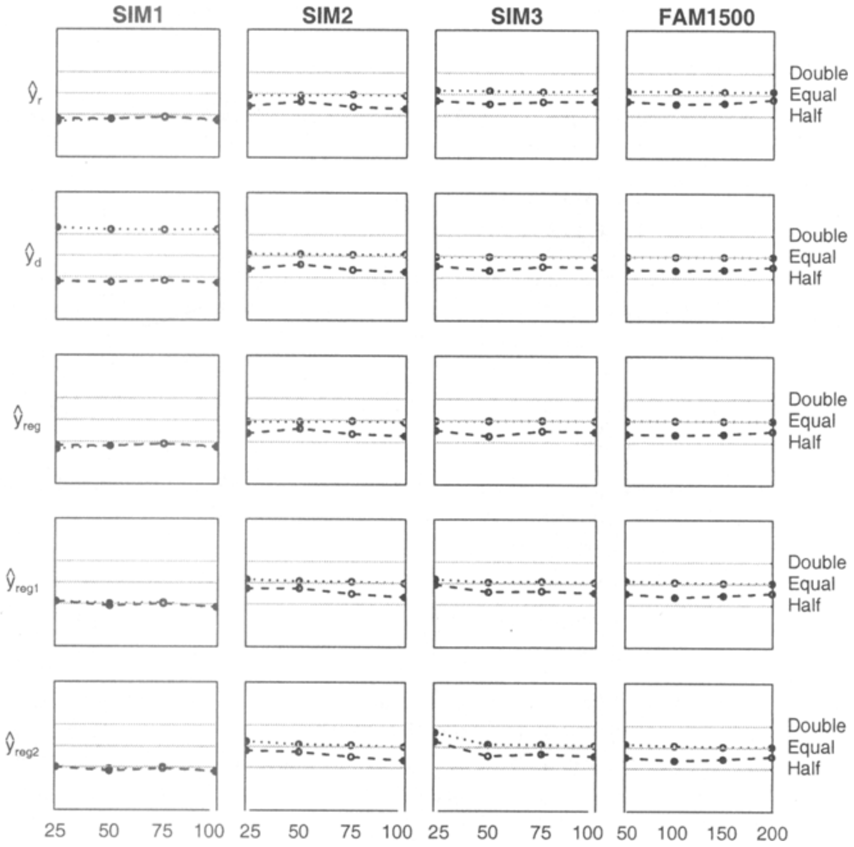


Figure 2: Log ratios of standar errors comparing the *cases available* estimators and the *complete data* estimators against the *simple* estimator, p=0.32n, q=0.48n. The dotted curve corresponds to the *complete data* estimator and the dashed curve refers to the *cases available* estimator.

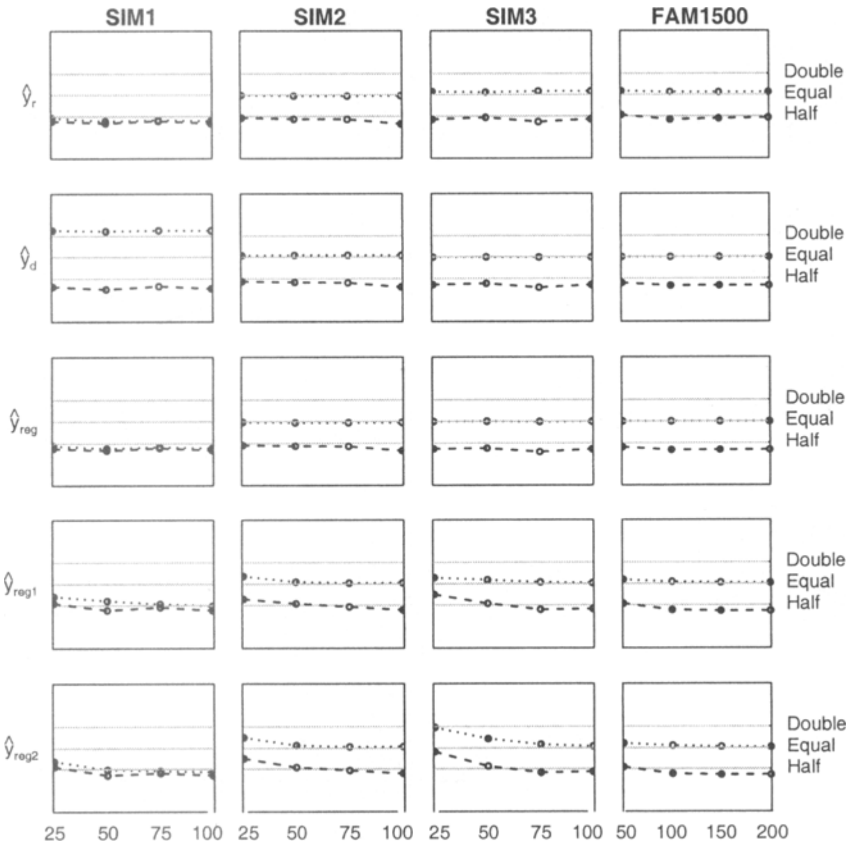
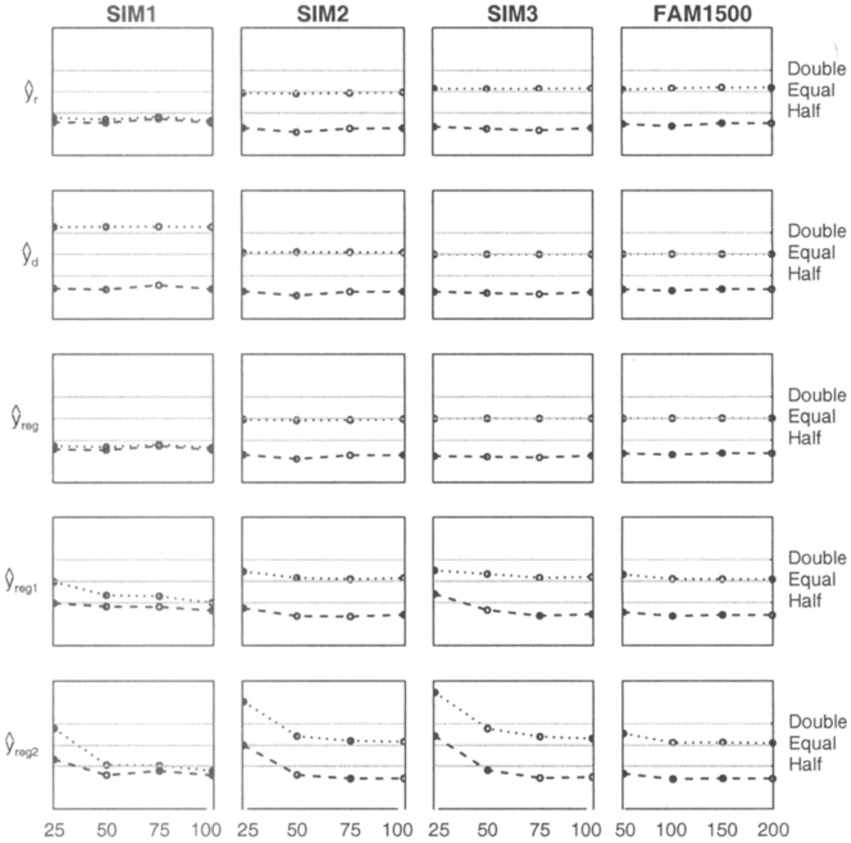


Figure 3: Log ratios of standar errors comparing the *cases available* estimators and the *complete data* estimators against the *simple* estimator, p=0.4n, q=0.48n. The dotted curve corresponds to the *complete data* estimator and the dashed curve refers to the *cases available* estimator.



regression estimators based both on the complete data and on available cases considerably improved on the precision of the direct estimator, while between the two estimators there was a less evident reduction in the error than among the other populations.

The behaviour pattern of the estimators $\hat{y}_{Reg21}$ and $\hat{y}_{Reg22}$, in relation to each other, is unclear. Depending on the population and on the sample size considered, one has a smaller error than the other. Evidently, the best behaviour is presented by the regression estimator based on the true value of b.

It has also been observed that, as expected, when the total missingness rate $\frac{p+q}{n}$ increased, the gain in the precision of the proposed estimators is greater.

The simulations were repeated, interchanging the values of p and q and the results obtained were very similar.

To sum up, these simulations show how the use of all the available data by the proposed estimators leads to a considerable error reduction in the estimation of totals, with respect to the respective estimators usually applied. This error reduction can be very great in certain cases, such as estimation by differences, which often functions unsatisfactorily. Moreover, it should be noted that there is a direct relation between error reduction and the missingness rate.

Acknowledgments

The authors would like to thank the referees for their many helpful comments and suggestions.

Research partially supported by MCYT (Spain) contract number BFM2001-3190.

References

[1] Brick, J.M., Kalton, G. (1996), Handling missing data in survey research, Statistical methods in medical research, 5, 215-238.

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

[3] King, G., Honaker, J., Joseph, A., Scheve, K. (1998), Listwise deletion is evil: what to do about missing data in Political Science, Unpublished document.

[4] Little, R.J.A., Rubin, D.B. (1987), Statistical analysis with missing data, John Wiley, New York.

[5] Meeden, G. (1995), Median estimation using auxiliary information, Survey Methodology, 21(1), 71-77.

[6] Morales, L. (2000), El efecto de la no respuesta parcial en el análisis de datos de una encuesta: una comparación entre la eliminación de observaciones y la imputación múltiple, Metodología de Encuestas, 2, 2, 217-218.

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

[8] Rubin, D.B. (1976), Inference and missing data, Biometrika, 63, 581-592.

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

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

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

[12] Singh, S., Joarder, A.M. (1998), Estimation of finite population variance using nonresponse in survey sampling, Metrika 47, 241-249.

[13] Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S., Asok, C. (1984), Sampling Theory of Surveys with Applications. Iowa State University Press. Iowa.

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

[15] Toutenburg, H., Srivastava, V.K. (1999), Amputation versus imputation of missing values through ratio method in sample surveys, Unpublished document.

[16] Toutenburg, H., Srivastava, V.K. (2000), Efficient estimation of population mean using incomplete survey data on study and auxiliary characteristic, Unpublished document.

[17] Tracy, D.S., Osahan, S.S. (1994), Random nonresponse on study variable versus on study as well as auxiliary variables, Statistica, 54, 163-168.

[18] Wolter, K.M. (1985), Introduction to variance estimation Springer-Verlag.

A Appendix: Values of the variances and covariances of the Horvitz-Thompson estimators in the case of simple random sampling and stratified sampling

If the sample design considered is simple random sampling, then $\pi_i = n/N$ and $\pi_{ij} = n(n-1)/N(N-1), \forall i, j$, and the variances and covariances of the estimators are given by:

$$Var(\widehat{y}^1_{HT}) = N^2 \frac{S^2_y}{n-p-q}\left(1 - \frac{n-p-q}{N}\right)$$

$$Var(\widehat{y}^3_{HT}) = N^2 \frac{S^2_y}{q}\left(1 - \frac{q}{N}\right)$$

$$\mathrm{Var}(\widehat{x}^2_{HT}) = N^2 \frac{S^2_x}{p}\left(1 - \frac{p}{N}\right)$$

$$Cov(\widehat{y}^1_{HT}, \widehat{x}^1_{HT}) = N^2 \left[\frac{1}{n-p-q} - \frac{1}{N}\right] S_{xy}$$

$$\mathrm{Cov}(\widehat{y}^3_{HT}, \widehat{x}^2_{HT}) = \begin{cases} N^2 \left(\frac{1}{p} - \frac{1}{N}\right) S_{xy} & \text{if } p \geq q \\ N^2 \left(\frac{1}{q} - \frac{1}{N}\right) S_{xy} & \text{if } p < q \end{cases}$$

$$\mathrm{Cov}(\widehat{y}^1_{HT}, \widehat{y}^3_{HT}) = \begin{cases} N^2 \left(\frac{1}{n-p-q} - \frac{1}{N}\right) S^2_y & \text{if } \frac{n-p}{2} \geq q \\ N^2 \left(\frac{1}{q} - \frac{1}{N}\right) S^2_y & \text{if } \frac{n-p}{2} < q \end{cases}$$

$$\mathrm{Cov}(\widehat{y}^1_{HT}, \widehat{x}^2_{HT}) = \begin{cases} N^2 \left(\frac{1}{n-p-q} - \frac{1}{N}\right) S_{xy} & \text{if } \frac{n-q}{2} \geq p \\ N^2 \left(\frac{1}{p} - \frac{1}{N}\right) S_{xy} & \text{if } \frac{n-q}{2} < p \end{cases}$$

$$\mathrm{Cov}(\widehat{x}^1_{HT}, \widehat{y}^3_{HT}) = \begin{cases} N^2 \left(\dfrac{1}{n-p-q} - \dfrac{1}{N} \right) S_{xy} & \text{if } \dfrac{n-p}{2} \geq q \\[2ex] N^2 \left(\dfrac{1}{q} - \dfrac{1}{N} \right) S_{xy} & \text{if } \dfrac{n-p}{2} < q \end{cases}$$

$$\mathrm{Cov}(\widehat{x}^1_{HT}, \widehat{x}^2_{HT}) = \begin{cases} N^2 \left(\dfrac{1}{n-p-q} - \dfrac{1}{N} \right) S^2_x & \text{if } \dfrac{n-q}{2} \geq p \\[2ex] N^2 \left(\dfrac{1}{p} - \dfrac{1}{N} \right) S^2_x & \text{if } \dfrac{n-q}{2} < p \end{cases}$$

where S_y, S_x and S_{xy} are the population variances and covariances of the variables.

If a random stratified sample design is used, then $\pi_i = \frac{n_h}{N_h}$ if the unit i is found within the stratum h of size N_h from which we obtained the sample size n_h and $\pi_{ij} = \frac{n_h}{N_h}\frac{n_h-1}{N_h-1}$ if the units i, j are in the same stratum h and $\pi_{ij} = \frac{n_h}{N_h}\frac{n_{h'}}{N_{h'}}$ if the units i, j are in different stratum, $h \neq h'$. In this case, the values of the variances and covariances are given by:

$$\mathrm{Var}(\widehat{y}^1_{HT}) = \sum_{1 \leq h \leq L} W^2_h \frac{S^2_{hy}}{n_h - p_h - q_h} \left(1 - \frac{n_h - p_h - q_h}{N_h} \right) N^2$$

$$\mathrm{Var}(\widehat{y}^3_{HT}) = \sum_{1 \leq h \leq L} W^2_h \frac{S^2_{hy}}{q_h} \left(1 - \frac{q_h}{N_h} \right) N^2$$

$$\mathrm{Var}(\widehat{x}^2_{HT}) = \sum_{1 \leq h \leq L} W^2_h \frac{S^2_{hx}}{p_h} \left(1 - \frac{p_h}{N_h} \right) N^2$$

$$\mathrm{Cov}(\widehat{y}^1_{HT}, \widehat{x}^1_{HT}) = N^2 \sum_{1 \leq h \leq L} W^2_h \, \mathrm{Cov}(\widehat{y}^1_h, \widehat{x}^1_h)$$

$$\mathrm{Cov}(\widehat{y}^3_{HT}, \widehat{x}^2_{HT}) = N^2 \sum_{1 \leq h \leq L} W^2_h \, \mathrm{Cov}(\overline{y}^3_h, \overline{x}^2_h)$$

$$\mathrm{Cov}(\widehat{y}^1_{HT}, \widehat{y}^3_{HT}) = N^2 \sum_{1 \leq h \leq L} W^2_h \, \mathrm{Cov}(\overline{y}^1_h, \overline{y}^3_h)$$

$$\mathrm{Cov}(\widehat{y}^1_{HT}, \widehat{x}^2_{HT}) = N^2 \sum_{1 \leq h \leq L} W^2_h \, \mathrm{Cov}(\overline{y}^1_h, \overline{x}^2_h)$$

$$\mathrm{Cov}(\widehat{x}^3_{HT}, \widehat{y}^1_{HT}) = N^2 \sum_{1 \leq h \leq L} W^2_h \, \mathrm{Cov}(\overline{x}^3_h, \overline{y}^1_h)$$

$$\mathrm{Cov}(\widehat{x}^1_{HT}, \widehat{x}^2_{HT}) = N^2 \sum_{1 \leq h \leq L} W_h^2 \,\mathrm{Cov}(\overline{y}^1_h, \overline{x}^2_h)$$

where $\overline{y}^i_h$ and $\overline{x}^i_h$ are the sample means of the variables y and x in the stratum h based on sample s_i.