Estimation of the population mean using auxiliary information when some observations are missing

M del Mar Rueda[1], Silvia González[2], Antonio Arcos[1], Yolanda Román[1], M Dolores Martínez[1], and Juan Fco. Muñoz

[1] Department of Statistics and Operational Research
Faculty of Sciences. University of Granada,
18071 Granada, Spain
(e-mail: mrueda@ugr.es, arcos@ugr.es, yroman@ugr.es, mmiranda@ugr.es)
[2] Department of Statistics and Operational Research
University of Jaén
Jaén, Spain
(e-mail: sgonza@ujaen.es)

Abstract. In this paper we use tools of classical statistical estimation theory in finding a suitable estimator of the population mean using auxiliary information when some observations in the sample are missing. We study model and design properties of the proposed estimator. We also report the results of a wide simulation study on the efficience of the estimator which reveals very promising results.
Keywords: Auxiliary information, missing data, superpopulation model.

1 Introduction

Missing data is a common problem in virtually all surveys. Frequently, survey sampling is conducted to gather complete information on all sampling units but, due to a variety of reasons, for a fraction of the subjects, either no data at all is available or information on one or more variables is missing. Missing data can contribute to bias in the estimates and make the analyses harder to conduct and results harder to present.

The most frequently used method to compensate for item non response is imputation (see [Little and Rubin, 1987]). Some statistics specialists are reluctant to apply this method because it manipulates the original information. Many empirical studies do not follow this approach. They simply discard all the sampling units with missing values and employ the usual inference procedures, which can produce that the actual sample size was less than the planned one, biases in estimations and increases in sampling variance if missing data follows any pattern.

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.

In this paper we propose a prediction approach to deal with the presence of missing data. Specifically, we address the case where only the value of the variable of interest is missing for some subjects, and the value of the auxiliary variable is missing for other distinct subjects. We propose a new estimator for the mean of the variable of interest, using all known data for principal and auxilary variables.

2 Estimation with auxiliary information and missing data

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., [Singh, 2003], [Sampath and Chandra, 1990], [Srivastava and Jhajj, 1981]. These methods of estimation assume that the sample data contains 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 by a simple random sampling without replacement when some observations are missing and the population mean of auxiliary characteristic is available (see [Tracy and Osahan, 1994], [Toutenburg and Srivastava, 1998] and [Rueda and González, 2004]).

There appears to be no effort reported in the literature when both the asumptions are violated simultaneously (some observations are missing in both variables and the population mean of the auxiliary variable is not known). We will consider this situation under a general sampling design.

Let be a population,U, of N units from which a random sample, s, of fixed size, n is drawn according to a noninformative 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.

We assume that only 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 $0 < p, q < n/2$.

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

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

Prediction theory for sampling surveys can be considered as a general framework for statistical inference on the characteristics of finite populations. The prediction approach is based on this idea: for any given $s \in S_d$ of size n we can write:

$$\overline{Y} = f_s \overline{y}_s + (1 - f_s)\overline{y}_{\tilde{s}} \tag{1}$$

where $f_s = n/N$ is the sampling rate, $\overline{y}_s = \sum_s Y_k/n$ is the mean for units in the sample, and $\overline{y}_{\tilde{s}} = \sum_{\tilde{s}} Y_k/(N-n)$ is the mean for the nonsample units.

In this representation of the mean, the sample mean $\overline{y}_s$ is known, and then we attempt a post survey prediction of the mean $\overline{y}_{\tilde{s}}$ of the nonsurveyed units.

Now for any given $s \in S_d$ we can write:

$$\overline{Y} = f_{s1}\overline{y}_{s1} + f_{s2}\overline{y}_{s2} + f_{s3}\overline{y}_{s3} + (1 - f_s)\overline{y}_{\tilde{s}} \tag{2}$$

where $f_{s1} = \dfrac{n-p-q}{N}$, $f_{s2} = \dfrac{p}{N}$, $f_{s3} = \dfrac{q}{N}$ and $f_{s4} = 1 - \dfrac{n}{N}$,

In this representation of the mean, the sample means $\overline{y}_{s1}$ and $\overline{y}_{s3}$ are known, thus the problem of predicting $\overline{Y}$ is equivalent to the problem of predicting the means $\overline{y}_{s2}$ and $\overline{y}_{\tilde{s}}$.

We denote by E_ξ the expected value under the model ξ and E_d the expected value under the design d. The minimum $E_\xi MSE_d$ criterium will be considered. We only consider the linear and unbiased under model predictors.

Consider any predictor T of $\overline{Y}$; it can be represented, for any given sample s as:

$$T = f_{s1}\overline{y}_{s1} + f_{s2}U_2 + f_{s3}\overline{y}_{s3} + (1 - f_s)U_4 \tag{3}$$

where U_2 and U_4 are considered as predictors of $\overline{y}_{s2}$ and $\overline{y}_{\tilde{s}}$ respectively. Tools of classical statistical estimation theory will be useful in finding the suitable predictors U_2 and U_4.

Firstly we study the problem of estimation of $\overline{y}_{s2}$. If the predictor T is of the form 3 and it verify: $E_\xi(T) = \mu = \frac{1}{N}\sum_{i\in U} E_\xi(Y_i)$, it is logical to consider the class of linear estimators U_2 with the condition: $E_\xi(U_2) = \mu_{s2} = \frac{1}{p}\sum_{i\in s2} E_\xi(Y_i)$. In the sample s_2 we do not have the values of the study characteristic but we have all the values of the auxiliary charastericitic, x. We now consider the frequently used regression model, where $\eta_i = \beta x_i$, $i = 1, ..., N$, where β is a unknown quantity. By generalized least squares theory, the minimum variance linear unbiased under the model estimator of β is, for a given sample, given by $\widehat{\beta}$ the sample regression coefficient. Then we consider the predictor $U_2^* = \widehat{\beta}\overline{x}_{s2}$ that is linear and unbiased under the model of $\overline{y}_{s2}$.

Regarding the estimation of $\overline{y}_{\tilde{s}}$, there is not any information available in s_4, neither from the study characteristic neither from the auxiliary characteristic, so it is logical to consider the sample mean $U_4^* = \overline{y}_{s1\bigcup s3}$.

We consider the predictor of $\overline{Y}$:

$$T^* = f_{s1}\overline{y}_{s1} + f_{s2}U_2^* + f_{s3}\overline{y}_{s3} + (1 - f_s)U_4^* \tag{4}$$

As $E_d(\overline{y}_{s1}) = E_d(\overline{y}_{s3}) = 0$, T^* is a linear ξ-unbiased predictor of $\overline{Y}$ for any design d, and therefore the random variable obtained from T^* if Y_k is fixed at y_k is ξ-unbiased estimator of population mean $\overline{y}$. The estimator T^* is also asymptotically normal. The proof use the asymptotical normality of U_4^* and U_2^* (see, e.g., Valliant et al., 2000).

Writting

$k_1 = \dfrac{n - p - q(N - p)}{N(n - p)}$, $k_2 = \dfrac{q(N - p)}{N(n - p)}$ and $k_3 = \dfrac{p}{N}$

the proposed estimator can be expressed as follows:

$$T^* = k_1\overline{y}_{s1} + k_2\overline{y}_{s3} + k_3\widehat{\beta}\overline{x}_{s2} \tag{5}$$

2.1 Simple random sampling

Next, we are going to consider a simple random sampling without replacement. We are interested in finding the statistical properties of the estimator with respect to this sampling design.

First, the estimator is unbiased under this design the approximate variance of T^* is

$$AV(T^*) = S_y^2\left[k_1^2 a + k_2^2 b + 2k_1k_2c\right] + \beta^2 k_3^2 S_x^2 d + 2k_3\beta S_{xy}\left[k_1 e + k_2 f\right] \tag{6}$$

where

$a = \frac{1}{n-p-q} - \frac{1}{N}$, $b = \dfrac{1}{q} - \dfrac{1}{N}$, $d = \dfrac{1}{p} - \dfrac{1}{N}$

$$c = \begin{cases} \dfrac{1}{n-p-q} - \dfrac{1}{N} & \text{if} \quad \dfrac{n-p}{2} \geq q \\ \dfrac{1}{q} - \dfrac{1}{N} & \text{if} \quad \dfrac{n-p}{2} < q \end{cases}$$

$$e = \begin{cases} \dfrac{1}{n-p-q} - \dfrac{1}{N} & \text{if} \quad \dfrac{n-q}{2} \geq p \\ \dfrac{1}{p} - \dfrac{1}{N} & \text{if} \quad \dfrac{n-q}{2} < p \end{cases}$$

$$f = \begin{cases} \dfrac{1}{p} - \dfrac{1}{N} & \text{if} \quad p \geq q \\ \dfrac{1}{q} - \dfrac{1}{N} & \text{if} \quad p < q \end{cases}$$

A consistent estimator of $AV(T^*)$ can be simply obtained by substituting S_y^2, S_x^2 and S_{yx} with their sample values s_y^2, s_x^2 and s_{yx}.

Table 1. Relation between lines type and nonresponse rates

Type of line	CASE 1	CASE 2	CASE 3
dotted	p = 0.32n	p = 0.32n	p = 0.4n
	q = 0.4n	q = 0.48n	q = 0.48n
dashed	p = 0.4n	p = 0.48n	p = 0.48n
	q = 0.32n	q = 0.32n	q = 0.4n

3 Simulation study

The next step in our study consists of carrying out a simulation study to reveal the behaviour of the proposed estimator. For this purpose, we examined four populations: CANCER, CO60, CO70 and HOSPITAL (see [Valliant *et al.*, 2000]).

In order to study the properties of the proposed estimator, the following process was repeated 1000 times: a simple random sample was selected, for which in a completely random way the selected proportion of cases for both variables was removed. The values of the proposed estimator T^* and of the estimator of the simple mean were then calculated. The results of this simulation are shown in Figure 1, and Table 1 describes the correspondence between the types of line and the nonresponse rates.

The above Figure represents the log-ratios of the mean squared errors of both estimators. The simulation results shown that for all the populations, sampling sizes and nonresponse rates considerated, the behaviour of the proposed estimator is better than that of the standard one (the sample mean). Moreover, there is an absence of variation in the error of estimation, produced by exchanging the proportion of nonresponders between the main variable and the auxiliary variable. Another interesting feature is that the precision improves in proportion to the increase in the sample size.

After comparing the T^* estimator and the standard estimator of the mean, we considered it useful to study the relation between the efficiency of the proposed estimator and that of the estimator defined by Toutenburg and Srivastava (1988), under the same conditions. We conclude that the behaviour of the T^* estimator is considerably better than that the Toutemburg estimator $\hat{y}_{T4}$.

References

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

[Rubin, 1977]D. B. Rubin. Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association*, pages 538–543, 1977.

[Rueda and González, 2004]M. Rueda and S. González. Missing data and auxiliary information in surveys. *Computational Statistics*, pages 555–567, 2004.

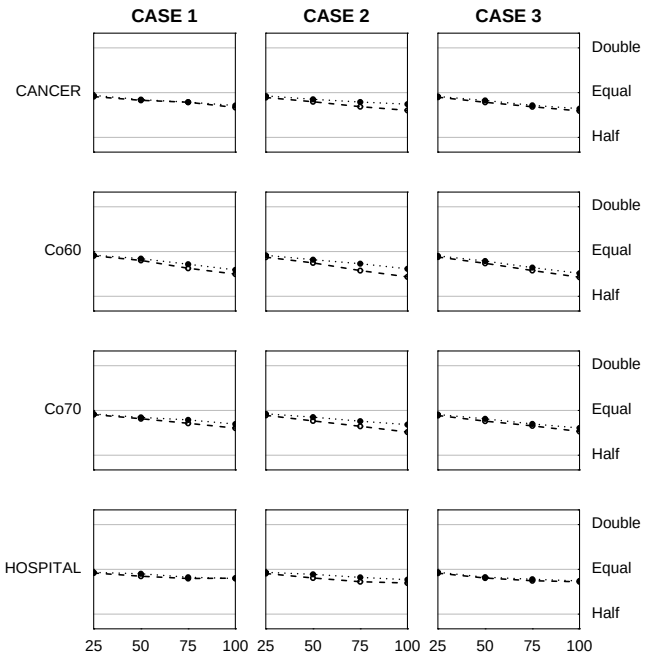

Fig. 1. Log ratios of standar error of the *predictive* estimators over the *simple* estimator.

[Sampath and Chandra, 1990]S. Sampath and S. K. Chandra. General class of estimators for the population total under unequal probability sampling schemes. *Metron*, pages 409–419, 1990.

[Singh, 2003]S. Singh. *Advanced sampling theory with applications. How Michael selected Amy*. Kluwer Academic Press, London, 2003.

[Srivastava and Jhajj, 1981]S. K. Srivastava and H. S. Jhajj. A class of estimators of the population mean in survey sampling using auxiliary information. *Biometrika*, pages 341–343, 1981.

[Toutenburg and Srivastava, 1998]H. Toutenburg and V. K. Srivastava. Estimation of ratio of population means in survey sampling when some observations are missing. *Metrika*, pages 177–187, 1998.

[Tracy and Osahan, 1994]D. S. Tracy and S. S. Osahan. Random non-response on study variable versus on study as well as auxiliary variables. *Statistica*, pages 163–168, 1994.

[Valliant *et al.*, 2000]R. Valliant, A. H. Dorfman, and R. M. Royall. *Finite population sampling and inference*. John Wiley, London, 2000.