

Using auxiliary information for missing data¹

Giancarlo Diana

Department of Statistical Science, University of Padova (Italy)

e-mail: giancarlo.diana@unipd.it

Pier Francesco Perri

Department of Economics and Statistics, University of Calabria (Italy)

e-mail: pierfrancesco.perri@unical.it

Abstract: We propose three regression-type estimators of a population mean when nonresponse occurs. The estimators make optimal use of the mean and variance of an auxiliary variable and outperform some estimators recently investigated by Diana and Perri (2009). A numerical illustration, carried out on two different populations, highlights the efficiency gain from using our new proposal.

Keywords: nonresponse, regression estimator, two-phase sampling, variance

1. Introduction

It is well known that surveys covering human populations are largely affected by the problem of nonresponse. Estimates from incomplete data may be misleading especially when the group of respondents differs from that of nonrespondents.

In the course of time, many imputation methods have been proposed to overcome the problem of missing data in the estimation of unknown population parameters. In sampling theory, imputation methods generally make use of an auxiliary variable to improve the efficiency of the estimates. In this context, the *ratio*, *product* and *regression* methods are traditional tools widely employed in practice. Basically, these methods are employed to estimate the unknown mean (total) of a study variable when the mean of an auxiliary variable is assumed to be known on the entire population. Nevertheless, extensions and developments have been proposed when missing data on the study variable occur, whether the auxiliary mean is known or not.

Recently, there has been a growing interest in the problem of missing data and many works have appeared in the sampling literature. Among others, Diana and Perri (2009), taking advantage from the information on the mean of an auxiliary variable, proposed three estimators of the population mean which work better than a recent proposal of Kadilar and Cingi (2008). In this short paper, we intend to show how the performance of our previous estimators can be further improved by using additional information. In particular, we propose three new estimators which involve the variance of an auxiliary variable.

¹Work supported by the Italian Ministry of University and Research, MIUR-PRIN 2007: “Efficient use of auxiliary information at the design and at the estimation stage of complex surveys: methodological aspects and applications for producing official statistics”

2. Suggested estimators

Let $\mathcal{Y}$ be the study variable surveyed on a finite population $\mathcal{U} = \{1, \dots, N\}$ of N identifiable units with values $y_i, i \in \mathcal{U}$. Let $\mathcal{X}$ be an auxiliary variable, correlated with $\mathcal{Y}$, which is used to estimate the unknown mean of $\mathcal{Y}$, namely $\bar{Y} = \sum_{i \in \mathcal{U}} y_i / N$. Let S_y^2 be the population variance of $\mathcal{Y}$. Hereafter, $\sum_A$ is our shorthand for $\sum_{i \in A}$.

Let s be a simple random sample without replacement of size n ($n < N$) drawn from $\mathcal{U}$ to estimate $\bar{Y}$. Suppose that the y_i 's can be observed only on a subset of s , $\mathcal{R} \subset s$, of size r ($r < n$). Accordingly, let $\bar{\mathcal{R}}$ be the subset of s of size $n - r$ consisting of nonresponding units ($s = \mathcal{R} \cup \bar{\mathcal{R}}$). For each unit $i \in \bar{\mathcal{R}}$, the y_i values are missing and are to be imputed. In this paper, following Singh and Deo (2003), we assume a missing completely at random (MCAR) mechanism. Imputation is performed by employing the auxiliary variable $\mathcal{X}$, whose values are assumed to be known for every sampled unit $i \in s$. We assume that the auxiliary mean, $\bar{X}$, may or may not be known. Let $\bar{x}_n = \sum_s x_i / n$, $\bar{y}_r = \sum_{\mathcal{R}} y_i / r$ and $\bar{x}_r = \sum_{\mathcal{R}} x_i / r$.

Without loss of generality, for given n and r , the estimation of $\bar{Y}$ can be framed in terms of two-phase sampling (Singh and Deo, 2003). Using this approach, Diana and Perri (2009), implementing three different regression-type methods of imputation for missing y_i , suggested the following estimators for $\bar{Y}$:

$$\bar{y}_1 = \bar{y}_r + b(\bar{X} - \bar{x}_n), \quad \bar{y}_2 = \bar{y}_r + b(\bar{X} - \bar{x}_r), \quad \bar{y}_3 = \bar{y}_r + b(\bar{x}_n - \bar{x}_r)$$

being b the least squares estimate of the regression coefficient of $\mathcal{Y}$ on $\mathcal{X}$ computed in the subsample $\mathcal{R}$. The approximate first order (i.e., up to terms of order n^{-1}) MSE is:

$$MSE(\bar{y}_1) \cong S_y^2 [\gamma_n(1 - \rho^2) + \gamma_{rn}]$$

$$MSE(\bar{y}_2) \cong \gamma_r S_y^2 (1 - \rho^2)$$

$$MSE(\bar{y}_3) \cong S_y^2 [\gamma_n + \gamma_{rn}(1 - \rho^2)]$$

where $\gamma_n = 1/n - 1/N$, $\gamma_r = 1/r - 1/N$, $\gamma_{rn} = 1/r - 1/n$ and ρ is the population correlation coefficient between $\mathcal{Y}$ and $\mathcal{X}$.

To the first order of approximation, the estimators are *unbiased* and *optimum* in the sense that they attain the minimum variance bound that can be reached, given the amount of auxiliary information employed. Their optimality derives from the theoretical results given in Diana and Tommasi (2003). This is why our proposals are more efficient than Kadilar and Cingi (2008)'s: $\tilde{y}_1 = [\bar{y}_r + b(\bar{X} - \bar{x}_n)]/(\bar{X}/\bar{x}_n)$, $\tilde{y}_2 = [\bar{y}_r + b(\bar{X} - \bar{x}_r)]/(\bar{X}/\bar{x}_r)$, $\tilde{y}_3 = [\bar{y}_r + b(\bar{x}_n - \bar{x}_r)]/(\bar{x}_n/\bar{x}_r)$.

We remark that other first order unbiased estimators, which utilize the same information on the auxiliary variable in the form of its population or sample mean, can be at most efficient as our estimators. Improvement upon them can be reached, for instance, by involving the mean of more than one auxiliary variable or taking advantage from different characteristics of the same auxiliary variable. With this in mind, and aimed at increasing the efficiency of $\bar{y}_i, i = 1, 2, 3$, we opt to integrate the knowledge of the mean of $\mathcal{X}$ with the auxiliary information on its variance. Therefore, let $S_x^2, s_{x,n}^2$ and $s_{x,r}^2$ be the population and the sample (in s and $\mathcal{R}$) variances of $\mathcal{X}$. Thus, paralleling estimators $\bar{y}_i, i = 1, 2, 3$, and following Diana and Tommasi (2003), we propose the new estimators of $\bar{Y}$:

$$\bar{y}_1^* = \bar{y}_r + b_{yx|\delta_x}(\bar{X} - \bar{x}_n) + b_{y\delta_x|x}(S_x^2 - s_{x,n}^2) \quad (1)$$

$$\bar{y}_2^* = \bar{y}_r + b_{y|x|\delta_x}(\bar{X} - \bar{x}_r) + b_{y\delta_x|x}(S_x^2 - s_{x,r}^2) \quad (2)$$

$$\bar{y}_3^* = \bar{y}_r + b_{y|x|\delta_x}(\bar{x}_n - \bar{x}_r) + b_{y\delta_x|x}(s_{x,n}^2 - s_{x,r}^2) \quad (3)$$

where $b_{y|x|\delta_x}$ and $b_{y\delta_x|x}$ denote, respectively, the sample estimates on $\mathcal{R}$ of the *partial regression coefficients* of $\mathcal{Y}$ on $\mathcal{X}$ and δ_x , $\delta_x = (\mathcal{X} - \bar{X})^2$. The estimators rely on three different imputation methods which can be obtained moving from the methods described in Diana and Perri (2009).

To the first order of approximation, the MSE of the new estimators is:

$$MSE(\bar{y}_1^*) \cong S_y^2 [\gamma_n(1 - \rho_{y,\tilde{x}}^2) + \gamma_{rn}] \quad (4)$$

$$MSE(\bar{y}_2^*) \cong \gamma_r S_y^2 (1 - \rho_{y,\tilde{x}}^2) \quad (5)$$

$$MSE(\bar{y}_3^*) \cong S_y^2 [\gamma_n + \gamma_{rn}(1 - \rho_{y,\tilde{x}}^2)] \quad (6)$$

where $\rho_{y,\tilde{x}}^2 = \mathbf{S}_{y\tilde{x}}^t \mathbf{S}_{\tilde{x}\tilde{x}}^{-1} \mathbf{S}_{y\tilde{x}} / S_y^2$ is the squared *multiple correlation coefficient* of $\mathcal{Y}$ on $\tilde{\mathbf{X}} = (\mathcal{X}, \delta_x)^t$, $\mathbf{S}_{y\tilde{x}} = [Cov(\mathcal{Y}, \mathcal{X}), Cov(\mathcal{Y}, \delta_x)]^t$ and variance-covariance matrix $\mathbf{S}_{\tilde{x}\tilde{x}} = [S_x^2, Cov(\mathcal{X}, \delta_x); Cov(\mathcal{X}, \delta_x), Var(\delta_x)]$. Moreover $Cov(\mathcal{X}, \delta_x) = \mu_{03}$, $Cov(\mathcal{Y}, \delta_x) = \mu_{12}$ and $Var(\delta_x) = \mu_{02}(\mu_{04}\mu_{02}^{-2} - 1)$, where $\mu_{rs} = \sum_{\mathcal{U}} (y_i - \bar{Y})^r (x_i - \bar{X})^s / (N - 1)$.

Expressions (4)-(6) depend on population parameters which are generally unknown. In practical situations, the efficiency of the estimators can be evaluated by adopting consistent estimates of the unknown parameters. However, expressions (4)-(6) are useful for theoretical considerations on efficiency comparisons. In fact, we note that the MSE of the new estimators depends on the multiple correlation coefficient that can never be reduced by increasing the number of variables upon which the dependence of $\mathcal{Y}$ is to be measured (Stuart et al., 2004). In our context, this means that $\rho_{y,\tilde{x}} \geq \rho$, implying that $MSE(\bar{y}_i) \geq MSE(\bar{y}_i^*)$, $i = 1, 2, 3$. Consequently, the new estimators, based on the auxiliary information on both mean and variance, outperform our previous proposal which allowed only for the auxiliary mean. Moreover, among the new estimators, it is easy to verify that $MSE(\bar{y}_i^*) \geq MSE(\bar{y}_2^*)$, $i = 1, 3$.

3. Numerical illustration

To evaluate the efficiency gain from the proposed estimators, we carry out a numerical illustration based on two different real populations. The performance of $\bar{y}_i^*$ upon $\bar{y}_i$ is discussed with respect to the sample mean $\bar{y}_r$ by means of the *Percent Relative Efficiency*, $PRE(\tilde{\bar{y}}) = 100 \times Var(\bar{y}_r) / MSE(\tilde{\bar{y}})$, $\tilde{\bar{y}} = \bar{y}_i^*, \bar{y}_i$ with $i = 1, 2, 3$.

Population 1. The data come from the *Survey of Household Income and Wealth* carried out by *The Bank of Italy* for the year 2002. In the analysis, we assume the 8011 Italian households as the target population on which the *household net disposable income* ($\mathcal{Y}$) and the *number of household income-earners* ($\mathcal{X}$) are investigated. We assume $n = 400$ and $r = 250$. Data are available at www.bancaditalia.it/statistiche.

Population 2. The data have been collected by a market research company. The population consists of $N = 2376$ points of sale for which the *sale area* ($\mathcal{Y}$) (in square meters) and the *number of employees* ($\mathcal{X}$) are surveyed. We assume $n = 300$ and $r = 100$.

Table 1: *PRE of the proposed estimators under two different populations.*

	Population 1	Population 2
	$S_y = 22225, \rho = 0.458, \rho_{y.\tilde{x}} = 0.462$	$S_y = 1930, \rho = 0.898, \rho_{y.\tilde{x}} = 0.913$
$\bar{y}_r$	100	100
$\bar{y}_1$	114.73	132.44
$\bar{y}_2$	126.5	514.33
$\bar{y}_3$	108.82	227.61
$\bar{y}_1^*$	115.07	133.93
$\bar{y}_2^*$	127.17	599.85
$\bar{y}_3^*$	109.02	238.06

Table 1 reports the *PRE* for the estimators involved in the empirical study. Pairwise comparison $\bar{y}_i$ vs $\bar{y}_i^*$, $i = 1, 2, 3$, confirms the theoretical results and provides an idea on the magnitude of the benefits of using the proposed estimators in the case of nonresponse. The efficiency gain is poor for Population 1, but it appears quite important in Population 2, especially with regard to $\bar{y}_2^*$. In every way, the use of the auxiliary variance certainly improves the performance of the estimators without any relevant theoretical and computational efforts. To conclude, we observe that we have also considered a model-free calibration estimator which turned out to be equivalent to $\bar{y}_2$.

The message we give is that it is worthwhile using the population and/or sample variance of the auxiliary variable in the case of missing data. Obviously, given the model that generates missing data, the magnitude of the improvement depends on the surveyed population and on the characteristics of the auxiliary variable. However, in the choice of the auxiliary variable lies the art of the researcher who has the hard task of finding suitable auxiliary variables which can procure substantial benefits in the inferential process. Along the same lines, we conclude by observing that further improvements can be produced by including additional information in the estimators, such as the auxiliary moments (central or not) of order greater than the second.

References

- Diana G. and Perri P. (2009) Improved estimators of the population mean for missing data, *Accepted on Communications in Statistics - Theory and Methods*.
- Diana G. and Tommasi C. (2003) Optimal estimation for finite population mean in two-phase sampling, *Statistical Methods and Applications*, 12, 41–48.
- Kadilar C. and Cingi H. (2008) Estimators for the population mean in the case of missing data, *Communications in Statistics - Theory and Methods*, 37, 2226–2236.
- Singh S. and Deo B. (2003) Imputation by power transformation, *Statistical Papers*, 44, 555–579.
- Stuart A., Ord K. and Arnold S. (2004) *Kendall's Advanced Theory of Statistics: Classical Inference & the Linear Model*, Vol. 2A, Sixth Ed., Arnold, London.