Design-based exploitation of big data by a doubly calibrated estimator

Uno stimatore a calibrazione doppia per l'uso dei big data in un'ottica basata sul disegno

Maria Michela Dickson [1], Giuseppe Espa[2] and Lorenzo Fattorini[3]

Abstract Big data typically constitute masses of unstructured data, not always available for a whole population. When sampling only the sub-population where big data are available, but neglecting the remaining portion, this can be viewed as a fixed component of nonresponses, which sums the natural component of nonresponses present in each survey. In this paper, big data information is exploited to handle nonresponse, while a size variable available for the whole population is exploited to handle the neglected part of the population by means of a doubly calibrated estimation. Design-based expectation and variance are derived up to the first order approximation. A variance estimator is proposed. A Monte Carlo simulation exploring various scenarios demonstrates the efficiency of the strategy.

Abstract *I big data costituiscono una mole di dati non strutturata, non sempre disponibile per tutte le unità di una popolazione. Quando si campiona solo dalla sotto-popolazione per cui i big data sono disponibili, trascurando la restante parte, questo può essere visto come una ulteriore fonte di mancate risposte che si aggiunge a quella naturalmente presente in ogni indagine campionaria. Nel presente lavoro, viene proposto uno stimatore a doppia calibrazione, nel quale i big data vengono utilizzati per gestire le mancate risposte, mentre, per gestire la parte di popolazione trascurata nella selezione, viene utilizzata una variabile dimensionale disponibile per*

[1] Maria Michela Dickson, Department of Statistical Sciences, University of Padova; email: mariamichela.dickson@unipd.it

[2] Giuseppe Espa, Department of Economics and Management, University of Trento; email: giuseppe.espa@unitn.it

[3] Lorenzo Fattorini, Department of Economics and Statistics, University of Siena; email: lorenzo.fattorini@unisi.it

l'intera popolazione. Valore atteso e varianza approssimati sino al primo ordine sono derivati in un'ottica completamente basata sul disegno. Inoltre si propone uno stimatore della varianza. Infine, mediante simulazioni Monte Carlo, vengono studiati scenari differenti per dimostrare l'efficienza della strategia proposta.

Key words: Auxiliary variables; Big data; Calibration estimators; Simulation study.

1. Introduction

In the last twenty years the availability of data has hugely increased, making possible developments in many fields of research, primarily in statistical sciences. More recently, this increase in data availability is also characterized by an increase in size of the amount of information collected, opening an extended debate around the term *big data*.

This new and potentially infinite source of data is connoted, on one side, by a not definite frame and, on the other side, by a real-time updating. Clearly the mentioned features represent pros and cons that researchers and practitioners must consider when using big data for statistical analysis (Tam, 2015). While the large amount of information is an unquestionable positive issue, the lack of a frame makes difficult the definition of a population of interest. This matter became relevant in sampling theory and in the consequent inference that can be done under a design-based perspective.

Nevertheless, in many practical circumstances, the opportunity of exploiting big data information may be an advantage in surveys. At the same time, considering only units provided by these additional information, could lead to a missed observation of the other units, which remain excluded from the study. For instance, it happens in socio-economic surveys on people living conditions, which tend to exclude units that cannot be contacted from the population, or in environmental studies, in which remote sensing information are not available for some areas. In this framework, these units never enter the sample and can be viewed as a fixed component of nonresponses, which sums the natural component of nonresponses present in each survey.

The aim of the present paper is to reach the desirable chance to take advantage from big data, when available, but make inference on a whole population in the above mentioned practical circumstances. So that, we propose an estimator that may permit to achieve this goal, by means of calibration estimators (Deville and Särndal, 1992). First, a calibration is used to correct for nonresponses from the sample to the population provided by big data, and, then, a second calibration is implemented on the whole population, leading to a *doubly calibrated estimator*. For the proposed estimator, expectation and variance are derived and a variance estimator is proposed. The efficiency and applicability of the strategy is showed by a Monte Carlo simulation study on a population that mirrors real characteristics.

2. Preliminaries, notation and methods

Let $U = \{u_1, \dots, u_N\}$ a population of N units. We denote with y_j, $j \in U$ the value for unit j of a survey variable Y. The aim is to estimate the population total

$$T_Y = \sum_{j \in U} y_j.$$

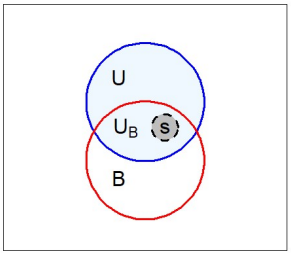
An auxiliary size variable Z is also available for each unit of U, such that the total $T_Z = \sum_{z \in U} z_j$ is known for any z_j, $j \in U$.

Moreover, a big data population B of M units intersects population U, so that we can define U_B as a sub-population of U composed by $N_B < N$ units. The total $T_{Y(B)} = \sum_{j \in U_B} y_j$ in U_B is unknown and it is the first quantity to be estimated in the procedure.

A sample S of size $n < N_B$ may be selected from the sub-population U_B by means of a fixed-size design having first and second order inclusion probabilities π_j, π_{jh} for any $h > j \in U_B$. As always happen in practice, the sample may be affected by nonresponses, so we define $R \subset S$ as the respondent sample. Note that the sampling scheme adopted to select S generates a sampling design on U_B but not on U.

To perform the first step of calibration, auxiliary information for U_B units are necessary. Let $\boldsymbol{x}_i = [x_{i1}, \dots, x_{iK}]$, with $i \in B$, the $\boldsymbol{X}$-vector of K auxiliary variables available in the population B of big data. The totals $\boldsymbol{T}_{X(B)} = \sum_{j \in U_B} \boldsymbol{x}_j$ and $T_{Z(B)} = \sum_{z \in U_B} z_j$ are known for all units in U_B. To better clarify the population setup, see Figure 1.

Figure 1: A stylized configuration of the population U_B, as the intersection between U and B, and sample S.



Because sample S is drawn from U_B, the H-T estimator (Horvitz and Thompson, 1953) of the total is $\hat{T}_{Y(B)} = \sum_{j \in S} \frac{y_j}{\pi_j}$ which is an unbiased estimator of $T_{Y(B)}$. Therefore, it would be a biased estimator of T_Y. However, considering nonresponses, the estimator

$$\hat{T}_{Y(B)/R} = \sum_{j \in R} \frac{y_j}{\pi_j} \neq \hat{T}_{Y(B)}$$

is a biased estimator even of $T_{Y(B)}$. In order to reduce the bias, following results proposed in Fattorini et al. (2013), it is possible to exploit auxiliary information $\boldsymbol{X}$ under a design-based point of view, obtaining the calibration estimator

$$\hat{T}_{Y(B)(cal)} = \widehat{\boldsymbol{b}}_R^t \boldsymbol{T}_{X(B)}$$

where $\widehat{\boldsymbol{b}}_R = \widehat{\boldsymbol{A}}_R^{-1}\widehat{\boldsymbol{a}}_R$ is the least-square coefficient vector of the regression of Y variable on $\boldsymbol{X}$ variables, performed on the respondent sample R, i.e. $\widehat{\boldsymbol{A}}_R = \sum_{j\in R}\frac{\boldsymbol{x}_j\boldsymbol{x}_j^t}{\pi_j}$ and $\widehat{\boldsymbol{a}}_R = \sum_{j\in R}\frac{y_j\boldsymbol{x}_j}{\pi_j}$. The design-based properties of $\hat{T}_{Y(B)(cal)}$ were derived in Fattorini et al. (2013). In that paper, it has been demonstrated that the estimator is approximately unbiased and consistent if the relationship between Y and $\boldsymbol{X}$ is similar in both respondent and non-respondent sub-groups, and it has been derived variance estimation. So that, the authors provide a design-based solution to the problem of nonresponses.

The additional problem here is that we must estimate T_Y rather than $T_{Y(B)}$. However, because $\hat{T}_{Y(B)(cal)}$ is, at his best, an approximately unbiased and consistent estimator of $T_{Y(B)}$, it is a biased estimator of T_Y. Since the selection of S is only on U_B, the elements of $U - U_B$ cannot enter the sample. In this case, it is necessary to correct the estimator to reduce the bias due to the sample under-coverage. A choice may be to use the size variable Z, which is available for the whole population U. So that, the resulting *double calibration* estimator turns out to be

$$\hat{T}_{Y(\mathrm{dcal})} = \frac{\hat{T}_{Y(B)(cal)}}{\hat{T}_{Z(B)}}T_Z = \frac{\widehat{\boldsymbol{b}}_R^t\boldsymbol{T}_X}{\hat{T}_{Z(B)}}T_Z$$

where $\hat{T}_{Z(B)} = \sum_{j\in S}\frac{z_j}{\pi_j}$ is the H-T estimator of T_Z.

The double calibration estimator benefits of some desirable properties deriving from calibration but, for the sake of brevity, we do not report all proofs. However, following results of Fattorini et al. (2013), $\hat{T}_{Y(dcal)}$ is approximately unbiased if (i) the linear relationship between Y and $\boldsymbol{X}$ is approximately the same in the respondent and non-respondent sub-groups of U_B, and if (ii) the proportional relationship between Y and Z is approximately the same in the two sub-populations U_B and $U - U_B$. Under these conditions, following the consistency of the H-T estimator (see Isaki and Fuller, 1982) it is possible to derive that $\hat{T}_{Y(dcal)}$ converges in probability to T_Y.

Regarding variance and its estimation, following Särndal et al. (1992, p.175) the estimator $\hat{T}_{Y(dcal)}$ has an approximate variance equal to

$$V\left(\hat{T}_{Y(dcal)}\right) \approx \left(\frac{T_Z}{T_{Z(B)}}\right)^2 \sum_{h>j\in U_B}\left(\pi_j\pi_h - \pi_{jh}\right)\left(\frac{u_j}{\pi_j} - \frac{u_h}{\pi_h}\right)^2$$

Given that, the Sen-Yates-Grundy variance estimator (Sen, 1953; Yates and Grundy, 1953) is given by

$$\hat{V}_{SYG}\left(\hat{T}_{CAL}\right) = \left(\frac{T_Z}{\hat{T}_{Z(B)}}\right)^2 \sum_{h>j\in S}\frac{\left(\pi_j\pi_h - \pi_{jh}\right)}{\pi_{jh}}\left(\frac{\hat{u}_j}{\pi_j} - \frac{\hat{u}_h}{\pi_h}\right)^2$$

where $\hat{u}_j = \left(r_j y_j \boldsymbol{x}_j^t - r_j\widehat{\boldsymbol{a}}_R^t\widehat{\boldsymbol{A}}_R^{-1}\boldsymbol{x}_j\boldsymbol{x}_j^t - z_j\hat{T}_{Z(B)}^{-1}\widehat{\boldsymbol{a}}_R^t\right)\widehat{\boldsymbol{A}}_R^{-1}\boldsymbol{T}_x,\ j \in S$ are the empirical *influence values* (Davison and Hinkley, 1997) computed for each $j \in S$.

In order to expose the validity of the proposed methodology, in the next section a simulation study is presented.

3. Simulation study

A Monte-Carlo simulation is discussed in this section to investigate the performances of the proposed estimator. A population of $N = 10000$ units has been considered. No random model was adopted for generating nonresponses. The population covered by big data U_B constituted by 7500 units has been partitioned in respondent and non-respondent sub-groups. Therefore, the response pattern is a fixed characteristic of the units, just like the value of the survey variable. The size of respondent sub-group is equal to $N_R = 2250, 4500, 6750$ units, corresponding to 30%, 60% and 90% of the population units. We suppose available two auxiliary variables X_1 and X_2 for units included in U_B. These variables have been generated according to a normal distribution with mean equal to 1, variance equal to 1 and correlation coefficient equal to 0.20. The variable under estimation Y has been generated as

$$y_j = 1 + 0.5x_{1j} + 0.5x_{2j} + \varepsilon_j, \quad \forall j \in U$$

where ε_j is an error component with mean equal to 0 and variance constant, such that the model explains in one case, the 60% and, in another, the 90% of the variance of the y_js. In addition, a size variable Z is available for the whole population and it has been generated as $z_j = 2y_j + \gamma_j$, where γ_j is an error component with mean equal to 0 and variance proportional to $k|Y|$, with k set to assure correlation between Y and Z equal to 0.90. From the described populations, 10000 Monte-Carlo random samples of size $n = 100, 150, 250, 375, 500$ units have been selected by means of simple random sampling without replacement (SRSWOR). For each sample, relative Root Mean Square Error (rRMSE) and relative Bias (rB) have been computed. Table 1 and Table 2 report obtained results.

Table 1: Relative bias and relative RMSE for the population with model $R^2 = 0.60$ and correlation between Y and Z equal to 0.90

	$N_R = 2250$		$N_R = 4500$		$N_R = 6750$	
n	***rB***	***rRMSE***	***rB***	***rRMSE***	***rB***	***rRMSE***
100	0.0349	0.2831	0.0310	0.2252	0.0265	0.2026
150	0.0232	0.2186	0.0171	0.1717	0.0142	0.1509
250	0.0110	0.1579	0.0076	0.1273	0.0096	0.1136
375	0.0088	0.1267	0.0029	0.0998	0.0054	0.0908
500	0.0081	0.1086	0.0021	0.0855	0.0011	0.0764

Table 2: Relative bias and relative RMSE for the population with model $R^2 = 0.90$ and correlation between Y and Z equal to 0.90

	$N_R = 2250$		$N_R = 4500$		$N_R = 6750$	
n	***rB***	***rRMSE***	***rB***	***rRMSE***	***rB***	***rRMSE***
100	0.0266	0.1915	0.0254	0.1805	0.0231	0.1768
150	0.0185	0.1517	0.0153	0.1422	0.0128	0.1351
250	0.0709	0.1095	0.0077	0.1065	0.0097	0.1031
375	0.0058	0.0895	0.0033	0.0835	0.0061	0.0829
500	0.0057	0.0764	0.0030	0.0718	0.0021	0.0701

Conclusions

Results show that in both populations, both relative Bias and relative RMSE decrease as sample size response portion increases. We have explored the case in which the size variable and the target variable are strongly correlated, confirming results gathered by Fattorini et al. (2013). Clearly, when the value of R^2 of the model used to generate Y is higher, performances of the estimator are better. However, we have considered very low sampling fractions and, nevertheless, the relative Bias rapidly decreases, becoming very close to zero with a sampling fraction equal to 5%. The behaviour of rRMSE complies this result, decreasing when the sample size and the size of respondents sub-group increase.

References

1. Davison, A.C., Hinkley, D.V.: Bootstrap methods and their application. Vol. 1. Cambridge university press (1997).
2. Deville. J.-C., Särndal C.-E.: Calibration estimators in survey sampling. J. Am. Stat, Assoc. 87. 376–382 (1992).
3. Fattorini, L, Franceschi, S., Maffei, D.: Design-based treatment of unit nonresponse in environmental surveys using calibration weighting. Biom. J., 55, 925-943 (2013).
4. Horvitz, D. G., Thompson, D.J.: A generalization of sampling without replacement from a finite universe. J. Am. Stat, Assoc. 47. 663-685 (1952).
5. Isaki, C.T., Fuller, W.A.: Survey design under the regression superpopulation model. J. Am. Stat. Assoc. 77. 89-96 (1982).
6. Särndal, C.-E., Swensson, B., Wretman, J.: Model Assisted Survey Sampling. Springer, New York (1992).
7. Sen, A.R.: On the estimate of variance in sampling with varying probabilities. J. Indian Soc. Agric. Statist., 5, 119-127 (1953).
8. Tam, S.M.: A statistical framework for analysing big data. The Survey Statistician. 72. 36-51 (2015).
9. Yates, F., Grundy, P.M.: Selection without replacement from within strata with probability proportional to size. J. R. Statist. Soc. B, 15, 235-261 (1953).