

Shrinkage estimation with preliminary data

Stima shrinkage in presenza di dati preliminari

Fabio Bacchini, Piero Demetrio Falorsi, Roberto Iannaccone

National Institute of Statistics (Istat), Rome - Italy

e-mail: bacchini@istat.it

Riassunto: In un recente lavoro è stato proposto uno stimatore che permette l'utilizzo di tutta l'informazione preliminare disponibile al momento del calcolo dell'indicatore, sia quella riferita ai comuni appartenenti al campione selezionato per la stima preliminare che quella dei comuni non appartenenti al campione. Questo approccio è utilizzato dall'Istat per la stima dei permessi di costruzione. Le stime ottenute hanno evidenziato una sistematica sovrastima quando è piccolo il numero dei comuni non campionari rispondenti. Partendo da queste premesse il lavoro si è concentrato sulla ricerca di una sintesi migliore dell'informazione disponibile definendo ed applicando una nuova procedura di tipo shrinkage che, almeno nelle prime applicazioni, risolve il problema della sovrastima.

Keywords: Preliminary estimates, two-phase sampling, response probabilities, shrinkage.

1. Introduction

Over recent years the need for increased timeliness of short term indicator has become more pressing at international level as confirmed by the studies of Eurostat (2002) and Oecd (2004). The two projects provides an extensive survey of the methodologies implemented to reducing the time taken to deliver indicator. One of this is the selection of a specific subsample for preliminary estimation. However, especially in the presence of administrative data, it is possible that although the collection of data on subsample units is still not complete, information is available for the units that does not belong to the sample used for preliminary estimation.

In a recent article Falorsi et al. (2005) propose an estimation method based on two-phase sampling theory and on the hypothesis that the response mechanism in data is the result of random process whose parameters can be suitable estimated. The methods allows the use of all data available at a certain time for the units included in a second phase sample and respondent in the first phase sample, taking into account in the weights the non response mechanism.

From an application to the monthly Italian building permits survey has been evident a trade off between the accuracy of estimation and the response rate in first phase sample when the indicator is calculated at 90 days from the reference period as required by Short-Term Statistics (STS) regulation.

In this work we propose a different way to put together all the information available at the estimation time, through a combination of two estimators, one for the sample units and the other for non sample units, in such way that the variability is reduced. The resulting shrinkage estimator reinforces the previous one and improves the estimation accuracy.

Moreover inside the new framework it is possible to obtain both the standard deviation for the single component of the new estimator and for the final value.

2. Methodology

As described in Falorsi et al. (2005, hereafter FABI) an unbiased estimator of the total of interest, that use all the preliminary information available at time t, is:

$$\hat{Y}_1 = \sum_{i \in r} y_i \omega_i \quad (1)$$

where the respondent r are represent by the union of two exclusive subset, belonging (s) or not ($\bar{s}$) to the sample ($r = r_s \cup r_{\bar{s}}$). Suppose, for semplicity, that the first phase sample is equal to the whole population and that all the equation members are referred to time t, the weights ω_i is equal to:

$$\omega_i = 1/(\pi_i \phi_{is} + (1 - \pi_i) \phi_{i\bar{s}}) \quad (2)$$

where π_i is the inclusion probabilities of the second phase design, ϕ_{is} is the probabilities to be respondent (within the given deadline) under the condition that unit i belong to the sample s and $\phi_{i\bar{s}}$ is the probabilities to be respondent (within the given deadline) under the condition that unit i does not belong to the sample subset $\bar{s}$. As illustrated by Eltinge and Yansaneh (1997), the predicted probabilities are used only in order to obtain an appropriate definition of the weighting cells. Application to the building permits has shown that this estimator does not perform properly when the number $r_{\bar{s}}$ is low.

In this work we propose a different approach that summarize in a different way all the information available at a given deadline. This new framework combines two different estimations, the first one related to the sample units and the second one from units that does not belong to the sample.

First, we define with $\hat{X}_1$ the population estimation obtained using only sample units:

$$\hat{X}_1 = \sum_{i=1}^{n_{r_s}} y_i / (\pi_i \eta_{is}) \quad (3)$$

where η_{is} is the probability to be respondent calculated following Eltinge and Yansaneh (1997).

Second, we define with $\hat{X}_2$ the population estimation calculated using only units not included in the sample:

$$\hat{X}_2 = \sum_{i=1}^{n_{r_{\bar{s}}}} y_i / ((1 - \pi_i) \eta_{i\bar{s}}) \quad (4)$$

where $\eta_{i\bar{s}}$ is the probability to be respondent.

Combining the two unbiased estimation of the population value, we are interested in the value of α that minimize the variance:

$$\hat{Z} = \alpha \hat{X}_1 + (1 - \alpha) \hat{X}_2 \quad (5)$$

this value is equal to:

$$\alpha = [\hat{V}(\hat{X}_2) - cov(\hat{X}_1, \hat{X}_2)] / [\hat{V}(\hat{X}_1) + \hat{V}(\hat{X}_2) - 2cov(\hat{X}_1, \hat{X}_2)] \quad (6)$$

To find α we calculate both the variance of $\hat{X}_1$ and $\hat{X}_2$ than their covariance. Moreover we derive the variance estimation of $\hat{Z}$

$$\hat{V}(\hat{Z}) = \alpha^2 \hat{V}(\hat{X}_1) + (1 - \alpha)^2 \hat{V}(\hat{X}_2) + 2\alpha(1 - \alpha)cov(\hat{X}_1, \hat{X}_2) \quad (7)$$

3. Data

Until the end of 2002, the survey on building permits was organized as a monthly census of all Italian municipalities. Each municipality had to communicate information either on building permits or null activity, meaning that no building permits have been released for that month.

In order to improve the timeliness (STS regulation requires data on building permits at 90 days from the reference period), and starting in 2003, Istat has flanked a sample survey to the traditional census on all Italian municipalities. In particular the sample has been selected using a stratified strategy. The municipalities with more than 50,000 citizens are included in the sample with certainty and form a take all stratum. The other municipalities are subdivided into strata defined by the cross-classification of two auxiliary variables: region and class of resident population. A fixed number of municipalities is selected in each stratum without replacement and with equal probabilities.

Following this strategy, for the month t at time $t + 3$ (90 days) two different sources of information are available: respondents included in the sample and respondents not included in the sample. Data used for the analysis refers to the period January-June 2004 that was the last period associate to a consolidate set of final data.

4. Results

For each month of the semester January - June has been elaborated both the shrinkage and the FABI estimation. For the first estimation are presented also the different results for sample and not sample subset with their relative standard errors (S.d.). Moreover for each month results are compared for different release time according to the new information available. Due to extension limit of the paper we present only evidence for March 2004 in table 1.

Shrinkage estimation performs better than FABI at 90 (June 2004) and 120 (September 2004) days from March 2004. This results is associated with a value of $\alpha = 0.69$ (90 days) and $\alpha = 0.79$ (120 days). This results derived directly by equations 6 considering that variance of $\hat{X}_2$ is greater than the variance of $\hat{X}_1$. From 150 days to the end of the period estimations results are quite similar for the two methodologies with a value of α that slowly decrease from $\alpha = 0.54$ to $\alpha = 0.42$.

Table 1: Comparison between shrinkage estimation and FABI for different time release - March 2004

Time release (days)	March 2004						
	$\hat{X}_1$	S.d.	$\hat{X}_2$	S.d.	Z	S.d.	FABI
90	20,957	1,611	24,683	2,542	22,117	1,266	27,619
120	21,719	1,747	23,196	3,655	22,023	1,519	22,127
150	21,337	1,657	22,349	1,825	21,802	1,111	21,801
180	21,328	1,629	21,108	1,491	21,210	979	21,406
210	21,117	1,596	22,163	1,654	21,624	1,029	21,709
240	20,909	1,577	23,530	1,827	22,058	1,075	22,655
270	20,909	1,577	24,093	1,516	22,552	964	23,327
300	20,884	1,565	23,737	1,451	22,398	934	23,047
330	20,884	1,565	22,922	1,284	22,063	861	22,449
360	20,884	1,565	22,881	1,267	22,049	853	22,413

References

- Eurostat (2002) *Final report of the expert group on sampling for timely European indicators*, Luxembourg.
- Eltinge JE, Yansaneh IS. (1997) Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. consumer expenditure survey, *Survey Methodology*, 23(1), 33–40.
- Falorsi P.D., Alleva G., Bacchini F., Iannaccone R. (2005) Estimated based on preliminary data from a specific subsample and from respondents not included in the subsample, *Statistical Methods and Applications*, 14, 83–99.
- OECD (2004) *Short-Term Economic Statistics (STES) Timeliness Framework*, Short-term Economic Statistics Expert Group (STESEG): Task Force on Timeliness and Benchmarking.
- Särndal CE., Swensson B., Wretman J. (1992) *Model Assisted survey sampling*, Springer-Verlag, New York.