

Estimation of variance for Consumer price index

Stima della varianza per l'indice dei prezzi al consumo

D'Alò M., Di Consiglio L., Falorsi S., Solari F.

Istat, Via Magenta 4, 00185-Roma, email: solari@istat.it

Riassunto: L'Istat ha avviato un processo di studio per una profonda revisione dell'indagine dei prezzi al consumo. In particolare, è in via di definizione un disegno campionario per la rilevazione che prevede una selezione probabilistica ai differenti stadi di campionamento. Questo lavoro presenta un metodo per la stima delle componenti di varianza associate ai diversi stadi di campionamento. Tale analisi rappresenta la base indispensabile per la definizione di un più efficiente disegno di campionamento. La determinazione della dimensione relativa delle componenti di varianza è, infatti, necessaria per determinare la numerosità campionaria e l'allocazione delle unità nei diversi stadi di selezione.

Keywords: price index estimation, linear mixed models, variance component, multistage sampling

1. Introduction

The Italian National Institute (Istat) has lately undertaken an extensive review of the Italian Consumer Price Survey (ICPS) in order to redesign different aspects of the current data collection and estimation methods. One of the main aspects of the redesign of the ICPS is the introduction of a fully probabilistic sampling design. In fact, the consumer price index (CPI) is, currently, estimated on the basis of data collected through a non probabilistic sample design with three stages of selection: municipalities (administrative centres of the provinces), outlets, *representative item*, (definition of a fixed basket of products, set up by experts independently from the first two stages. For each representative item, the most sold *variety* is chosen in the outlet at the beginning of each year); see Istat (2005a) for a detailed description of the ICPS.

The multistage structure of the sampling has been preserved in order to control the whole cost of sampling collection even if a probability selection scheme has been proposed at each sampling stage (see Istat, 2005b). In order to determine an efficient allocation of resources at each stage of sampling, a study on the variance components has been undertaken. To this purpose linear mixed models have been used and this paper reports the results for the estimation of the random components for the Food major group and for Clothing.

In paragraph 2 the main features of the probabilistic sampling redesign are described, paragraph 3 introduces the mixed models for the elementary index and paragraph 4 contains the formulation of the model variance for the CPI estimator. Finally, in paragraph 5 the results of the fitted models are reported.

2. The sampling design

Here we briefly describe the features of the ICPS sampling redesign which are useful in order to understand the following analysis. The municipalities are the primary sampling units, the outlets the secondary sampling units and the *items* the ultimate sampling units. The municipalities are stratified by region and, then, selected by means of balanced sampling. Then, for each selected municipality, a sample of outlets is selected for each type of product; these samples are positively coordinated in such a way that maximum overlap of the selected outlets is guaranteed. Outlets are stratified by type

of outlet. Instead of setting up a fixed basket, items are selected in a class of aggregate products (AP), which are subset of homogeneous enough items, AP themselves are independently selected in each sampled outlet. The inclusion probabilities at each sampling stage are proportional to turnover values (known by the register or suitably imputed). A detailed description of the sampling strategy of the redesign of ICPS is contained in Istat, (2005b).

Let $r_{cvdj} = p_{t,cvdj} / p_{0,cvdj}$ be the elementary price index in t for item j of aggregate product d in the v outlet of the municipality c . Note that sampling redesign is such that the probability of inclusion π_{cvdj} is proportional to its turnover w_{cvdj} , so that if the target variable is a Laspeyres index, $I^{0,t}$, then, the Horvitz-Thompson estimate reduces to the sample mean of the elementary indexes. Let the *anticipated variance* (AV) of the estimator $\tilde{I}^{0,t}$ of $I^{0,t}$ be defined as $E_p[V_m(\tilde{I}^{0,t})|s]$, being s the given sample of items, $V_m(\tilde{I}^{0,t})$ the model variance (MV) and $E_p(\cdot)$ the expected value under the sampling design; then, it is known that the sampling designs with $\pi_{cvdj} \propto w_{cvdj}$ allows to minimize the AV (Särndal *et al.*, 1992).

3. Mixed models

According to the sampling redesign, the elementary index can be modelled using three nested components of variance each of them corresponding to the three stages of sampling described in §2. Then, the elementary index r_{cvdj} can be represented by means of the following linear mixed model

$$r_{cvdj} = \beta' X_{cvdj} + \alpha_c + \delta_{cv} + \gamma_{cd} + \varepsilon_{cvdj}, \quad (1)$$

with β a fixed effect relative to the vector of auxiliary variables X , while α_c , δ_{cv} and γ_{cd} are the mutually independent random effects with mean 0 and ε_{cvdj} are the errors independently distributed from the random effects. This type of model has been applied, for example, in Baskin (1993) for the US consumer price index.

Plausible choices of the set of the auxiliary variables X are stratification variables such as regions, dimension of the municipalities or type of outlet, type of product. Separate models have to be specified for large groups of goods and services such as Food, Clothing, Transportation, Housing. In particular, this work reports the results of the analysis of population models for Food and for Clothing.

Since the models have been fitted on data collected with the current ICPS design (data of year 2002), model (1) has to be simplified as the available data do not allow to fit the complete nested model (1). In fact, the number of observations in each outlet does not permit to estimate all the variance components. Anyway, also the simpler formulation

$$r_{cvdj} = \beta' X_{cvdj} + \alpha_c + \delta_{cv} + \gamma_{cd} + \varepsilon_{cvdj} \quad (2)$$

could not be estimated on our sets of data, since for these two groups we have encountered computational restrictions that do not allow models of such complexity (but note that model has been fitted on a Health Services, whose number of observation is much more limited).

Recalling that the aggregate product in the current survey are chosen *a priori* from the municipality and outlet selection (representative items), then, the following simplified version:

$$r_{cvdj} = \beta' X_{cvdj} + \alpha_c + \delta_{cv} + \gamma_d + \varepsilon_{cvdj} \quad (3)$$

has been applied for different definitions of the auxiliary variables.

4. The model variance of the CPI

Recall that the for the probabilistic sampling redesign the estimated CPI obtained by means of Horvitz Thompson estimate reduces to the sample mean $\tilde{I} = \frac{1}{n} \sum_c \sum_v \sum_d \sum_j r_{cvdj}$, where n is the whole sample size. Then after having estimated the variance component of model specification (1) the MV of $\tilde{I}$ conditional to the sample is given by:

$$V_{m1}(\tilde{I}^{0,t}) = \frac{1}{n^2} \left(\sigma_\alpha^2 \sum_c n_c^2 + \sigma_{\alpha\delta}^2 \sum_c \sum_v n_{cv}^2 + \sigma_{\alpha\delta\gamma}^2 \sum_c \sum_v \sum_d n_{cvd}^2 \right) + \sigma_\varepsilon^2, \quad (4)$$

where n_c , n_{cv} and n_{cvd} denote sample size at municipality, outlet and AP level.

As described in §3 only model (3) can be estimated with available data, related MV is given by

$$V_{m3}(\tilde{I}^{0,t}) = \frac{1}{n^2} \left(\sigma_\alpha^2 \sum_c n_c^2 + \sigma_{\alpha\delta}^2 \sum_c \sum_v n_{cv}^2 + \sigma_\gamma^2 \sum_d n_d^2 \right) + \sigma_\varepsilon^2 \quad (5)$$

where n_d represent sample size for the representative item d .

Moreover, since current IPCS design is non-probabilistic, MV is a useful instrument to give an assessment of the variability of the estimates which are currently disseminated. Recall that the current estimator for the CPI is, in an extremely simplified formulation, given by

$$\hat{I} = \sum_c \sum_v \sum_d \sum_j r_{cvdj} w_{cvdj}. \quad (6)$$

Given the current ICPS, model (3) is more appropriate for the definition of the MV of (6)

$$V_{m2}(\hat{I}^{0,t}) = \sigma_\alpha^2 \sum_c w_c^2 + \sigma_{\alpha\delta}^2 \sum_c \sum_v w_{cv}^2 + \sigma_\gamma^2 \sum_d w_d^2 + \sigma_\varepsilon^2 \sum_c \sum_v \sum_d \sum_j w_{cvdj}^2, \quad (7)$$

where w_c , w_{cv} , w_d , w_{cvdj} are the turnovers relative to municipality, outlet, AP and single item level.

5. Results

The following results relate to the models for Food and Clothing with auxiliary variables given by the stratification variables proposed in the sampling redesign; i.e. region; type of product; type of outlet. Note that the classification by region result not significant for the model for Food, while it is for Clothing; anyway this stratification variable has been kept in the sampling plan to provide estimates of the CPI at regional level. Furthermore, fixed effects related to the dimension of municipalities in terms of population (in 3 or 2 classes of population determined on the basis of quantiles of the sample municipalities) and the location of outlet inside the municipalities (centre, outskirts, etc) have been taken into account but both resulted not significant.

In order to fit the models, SAS procedure *proc mixed* has been applied. SAS uses REML estimates of the random components so that non-negative estimates are guaranteed.

The following tables display the results obtained for the model for Food and for the model for Clothing. Tables 1 and 2 report the estimation of the variance components together with their standard error P-values, while tables 3 and 4 show the significance levels for the fixed effects.

Table 1. *Variance components estimates, standard error and P-values for food*

Source of variation	Estimate	Standard error	P-value
Municipality	0.05476	0.02065	<.0040
Outlet	0.6162	0.06001	<.0001
Product	0.1739	0.03488	<.0001
Residual	22.3699	0.1110	<.0001

Table 2. *Variance components estimates, standard error and P-values for clothing.*

Source of variation	Estimate	Standard error	P-value
Municipality	0.05785	0.02948	<.0248
Outlet	1.8683	0.1282	<.0001
Product	0.03914	0.01437	<.0032
Residual	8.3259	0.09589	<.0001

Table 3. *Significance levels for the fixed effects for food.*

Fixed effect	Degree of freedom	P-value
Type of outlet	1	<.0001
Type of product	46	<.0001
Region	19	<.7599

Table 4. *Significance levels for the fixed effects clothing.*

Fixed effect	Degree of freedom	P-value
Type of outlet	1	<.0014
Type of product	18	<.0001
region	19	<.0007

This preliminary analyses on the two groups of goods have to be extended to all the twelve groups that represent the classification of Goods and services.

Note that the study can produce negatively biased estimates of the variability due to the type of data collected by the current IPCS. To prevent this occurrence, a more accurate study should be performed on an *ad hoc* set of data or at least sample allocation has to be updated after the first occurrence of the redesigned survey.

Finally, since fitting of model (2) can be carried out on smaller amount of data, a reduction of the available sample can overcome computational complexity.

References

- Baskin R. (1993) Estimation of Variance Components for the U.S. Consumer Price Index via Gibbs Sampling
- Istat (2005a) www.istat.it/prezzi/precon/aproposito/metodologia.pdf *Come si rilevano i prezzi al consumo*, Marzo 2005.
- Istat (2005b) *Proposta di un piano di campionamento probabilistico per la rilevazione territoriale dei prezzi al consumo*, Technical Report.
- Särndal C. E., Swensson B., Wretman J. (1992). *Model Assisted Survey Sampling*. New York, Springer.
- Valliant R. (1999). Uses of Models in the Estimation of Price Indexes: a Review. In *Proceedings of the Survey Research Methods Section, American Statistical Association*.