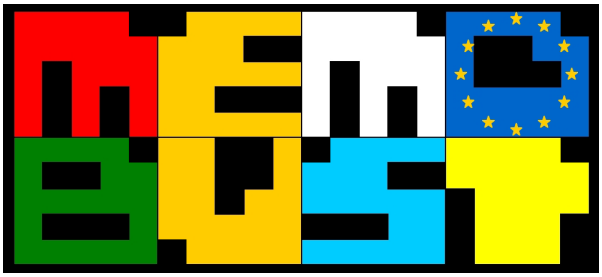

This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Preliminary Estimates with Design-Based Methods

Contents

General section

1. Summary

Timeliness is a particularly critical component of quality for producing short-term business statistics at the National Statistical Institutes (NSIs) of the European Community, as each Member State has to meet the standard quality requirements of the Regulation No 1165/98 – amended by the Regulation No 1158/2005 - about terms for transmission of the results and details of the information provided on statistical indicators, particularly on short-term statistics. The Amendment EU Regulation on Short Term Statistics requests all the statistical institutes of the EU Member States to transmit *preliminary short term* indicators to EUROSTAT with a reduced delay comparing to the timeliness set in the original 1998 Regulation (Eurostat, 2000, 2001, 2005). In OECD context, also, research projects were settled and useful documentation produced (Di Fonzo, 2005).

Frequently, in the NSIs short term statistics are based on fixed panel surveys of enterprises or rotating panels with a partial overlap from one year to another. More precisely, the amended regulation provides for a substantial improvement of timeliness for the production of the most important short-term indicators.

A common approach for dealing with *preliminary estimates* focuses essentially on the study and the definition of efficient estimators, exploiting almost exclusively auxiliary information in the estimation phase. Often preliminary estimation merely involves the use of the quick respondent units. In fact, in order to obtain "good" preliminary estimates, standard survey strategy often aims to achieve high quick response rate by means of a well-structured plan of follow-up. In some surveys the "largest" units are carefully supervised.

The main theoretical problem to be faced in a short-term preliminary estimation context concerns the possible self-selection of quick respondents that can lead to biased estimates of the unknown population mean and variances.

A useful documentation on preliminary estimation problems (even though not comprehensive) can be downloaded from the OECD web site[1].

This module focuses on estimation methods referring to the design-based approach. In particular describes a method proposed in Rao et al. (1989) which uses, to produce estimate referred to time t, data pertaining both to time t and t-1, with the aim to minimise the mean square error of the estimate.

Apart from this particular method, design-based (or model-assisted) estimation methods for preliminary estimates using quick respondents refer to the class of non-response weighting adjustment procedures, which are used in general when the Theoretical Sample (TS) is not achieved in practice in the Observed Sample (OS). In the case of preliminary estimates the observed sample coincides with the quick respondent set of units, available at the point of time when preliminary estimation has to be performed.

It is worth noting that in the context of preliminary estimate production the two most frequent situations at NSIs are: (i) using for the preliminary estimates the same estimator used for the final

[1] For the issue of the preliminary subsample the link is:
http://www.oecd.org/document/17/0,3746,en_2649_33715_30386193_1_1_1_1,00.html.

estimates computed on quick respondents, or (ii) referring to model-based estimators and ignoring the sampling design which generated data.

2. General description of the method

In general, the standard process going from data collection to elaboration of survey data needs to be accomplished within a fixed period of time, especially if the final estimates must be disseminated at a prefixed time point τ_f. In this context, direct estimators of the target parameters – based on the sampling units included in the TS and selected through a probabilistic sampling design – are design unbiased and consistent; the sampling error depends on the variability of the phenomenon under study, on the planned sample size and on the effectiveness of the selection procedure. Direct estimates based on the OS – that is a subset of quick respondents of TS with size depending on the nonresponse rate – can be biased in function of the random response process generating the OS.

We assign the term "preliminary" to the estimates computed using the statistical information available at a point of time τ_p preceding the time τ_f, on the basis of the OS denoted in this case as Preliminary Sample (PS), i.e., the sub-sample of *quick respondent units* that is available to be processed to produce the estimate at τ_p. The corresponding *final estimate* is based on a final sample, including both *quick* and *late respondents*, observed from τ_p and τ_f. The most straightforward practice in this situation is to apply the same estimation techniques utilised to produce the final estimates. Alternative estimation techniques (De Sandro and Gismondi, 2004; D'Alò et al., 2007) should take under control the *bias* and the *revision error*, given by the difference between final and preliminary estimates. In order to test the quality of the preliminary estimator, the revision error should be evaluated for different survey occasions.

Some indicators of the revision error can be defined and compared on the basis of the time series of provisional estimates and final ones. Among them, the following indicators can be evaluated.

- *Average total revision*, that is the average of the difference between the latest available value and the first release for each observation period. This measure indicates a possible bias of the first release.
- *Average absolute revision*, that is the average of the absolute difference between the latest available value and the first release for each observation period, regardless of their sign. This measure indicates the stability of the first release.
- *Range of total revisions*. Highest and lowest total revisions to the first release for all observation periods. This range indicates the volatility of the first release. The total range covers all the revisions and may include outliers.

Preliminary estimation methods may be classified in function of the stage on which specific preliminary methods are applied. In fact, it is possible to identify methods which act:

- at the sampling design stage, by selecting a preliminary subsample of TS (cf. the module "Sample Selection – Subsampling for Preliminary Estimates");
- at the estimation stage, in the following ways:
    1. by means of imputation techniques of missing data, that are applied to the non-respondent units in TS but not in PS (cf. the topic "Imputation");

2. by means of weighting adjustment, i.e., calculating nonresponse correction factor when early respondents are used in the standard estimator, the same adopted for the final estimate, modifying the sampling weights assigned to the units in PS in order to take into account non respondents in TS;
3. by applying direct and indirect estimators, using known population totals of auxiliary variables and/or time series of preliminary and final estimates of the variable of interest.

The estimation of individual response probabilities – useful to modify sampling weights of the ordinary Horvitz-Thompson estimator – is quite difficult because of randomness of some nonresponse and the lack of enough reliable auxiliary variables (Rizzo et al., 1996). Imputation techniques render easier the estimation process, but normally do not reduce bias because they are founded on data concerning respondent units only. These evidences stressed a wider recourse to a model-based approach, as remarked in Särndal et al. (1993), Valiant et al. (2000), Särndal and Lundström (2005).

In the model-assisted approach, weighting may be based on a *calibration approach.* A *calibrated weight* is obtained by the multiplication of the *direct or design weight* – defined as the reciprocal of the inclusion probability – with a *correction factor*. The correction factor is a nonresponse adjustment weight that attempt to compensate for unit nonresponse. A commonly used procedure for obtaining these weights is to divide the total sample into a set of weighting classes based on information known for both respondents and non-respondents and then to increase the base weights for the respondents in a weighting class to represent the non-respondents in that class (Kalton, 1983; Särndal and Lundström, 2005). Several methods to define adjustment cells are presented in literature (Rizzo et al., 1996; Eltinge and Yansaneh, 1997; Breiman et al., 1984; Little, 1986).

Depending on the informative context, the totals used for calibration may: (*i*) be known at population level; (*ii*) be estimated - using expansion weights – by the TS units or (*iii*) represent the final estimates of previous survey occasions. In order to reduce the bias, the auxiliary variables should explain both the main study variables and the inverse response probability.

In some survey there is an extensive amount of information available for the non-respondents. This information may derive from the sampling frame or by matching sampled elements with administrative records. Besides, in panel surveys and other surveys involving more than one wave of data collection, extensive information of non-respondents at later waves is available from their responses at early waves.

It is useful to underline, finally, that when the target variables are dependent on the provisional response mechanism, the preliminary estimates may be affected by some bias.

2.1 A design-based estimation method based on composite estimator

For this method the approach is based on a probabilistic design as both the theoretical sample and the observed quick respondent sample are considered as generated by a random design. The expected value $E(.)$ and the variance $V(.)$ of the estimators are considered with respect to these sampling designs. Furthermore, a random mechanism of nonresponse is supposed to generate the anticipated sample.

In this context, Rao et al. (1989) proposed the *composite estimators* that may represent an improvement of *the standard estimator.*

Generally speaking, the basic composite estimator is obtained as weighted average of the *preliminary estimate* for time *t* and the *final estimate* of time *t*-1 adjusted for the difference between preliminary estimates referred to *t* and *t*-1.

For the estimate of a population total y_t, let Y_t^P, Y_t and $Y_t^* = Y_t - Y_t^P$ be respectively the preliminary estimate, the final estimates and the measurement errors in preliminary estimates at time *t*, $t = 1,\ldots,T$. The proposed composite estimator is:

$$Y_{t,\alpha} = \alpha Y_t^p + (1-\alpha)\left[Y_{t-1} + (Y_t^p - Y_{t-1}^p)\right],$$

being α a weight varying between 0 and 1.

To determine the "optimal" α, i.e., that assuring minimum variance, some reasonable assumptions are made:

a1: $E\left(Y_t^p - Y_{t-1}^p\right) = E\left(Y_t - Y_{t-1}\right)$, $E(.)$ denoting the expected value,

a2: $\left|B\left(Y_i^p\right)\right| \geq \left|B\left(Y_i\right)\right|$, i=t, t-1 and $B(.)$ denoting the bias; furthermore, it is assumed for simplicity that $B\left(Y_i\right) = 0$ and $B\left(Y_i^p\right) = \delta$.

Then we get

$$\alpha = \left[V(Y_{t-1} - Y_{t-1}^p) + Cov\left(Y_t^p, Y_{t-1}\right) - Cov\left(Y_t^p, Y_{t-1}^p\right)\right]\left[V(Y_{t-1} - Y_{t-1}^p) + \delta^2\right]^{-1}.$$

In a similar way the optimal α for the composite estimator for change $\left(y_t - y_{t-1}\right)$ is shown in Rao et al. (1989), where the impact of the size of δ on the equivalence of using the optimal α for estimate level or change is discussed as well.

Variance and covariance terms can be estimated on survey data using usual formulas. The paper by Rao et al. (1989) introduces a further assumption about covariances which allows to simplify the expression for α; this assumption, anyway, is valid when bias in preliminary estimates is due to undercoverage but not when it is due to nonresponse.

3. Preparatory phase

4. Examples – not tool specific

5. Examples – tool specific

6. Glossary

For definitions of terms used in this module, please refer to the separate "Glossary" provided as part of the handbook.

7. References

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984), *Classification and Regression Trees*. Chapman and Hall, New York.

D'Alò, M., De Vitiis, C., Falorsi, S., Righi, P., and Gismondi, R. (2007), Sampling Strategies for Preliminary Estimates Production in Short-Term Business Surveys. *Proceedings of the 2007 intermediate conference Risk and prediction*, Società Italiana di Statistica.

De Sandro, L. and Gismondi, R. (2004), Provisional Estimation of the Italian Monthly Retail Trade Index. *Contributi-Istat*, 24/2004.

Deville, J.-C. and Tillé, Y. (2004), Efficient Balanced Sampling: the Cube Method. *Biometrika* **91**, 893–912.

Di Fonzo, T. (2005). The OECD project on revisions analysis: First elements for discussion. Paper presented at OECD STESEG meeting, Paris, 27-28 June 2005. http://www.oecd.org/dataoecd/55/17/35010765.pdf

Eltinge, L. and Yansaneh, I. S. (1997), Diagnostics for Formation of Nonresponse Adjustment Cells, With an Application to Income Nonresponse in the U.S. Consumer Expenditure Survey. *Survey Methodology* **23**, 33–40.

EUROSTAT (2000), *Short-term Statistics Manual*. Eurostat, Luxembourg.

EUROSTAT (2001), Conclusion of the First Meeting of the Expert Group Contro-Stratified European Sample for Retail Trade, Final Report, July 2001. Eurostat, Luxembourg.

EUROSTAT (2005), *Council Regulation No 1165/98 Amended by the Regulation No 1158/2005 of the European Parliament and of the Council – Unofficial Consolidated Version*. Eurostat, Luxembourg.

Little, R. J. A. (1986), Survey Nonresponse Adjustments for Estimates of Means. *International Statistical Review* **54**, 139–157.

Kalton, G. (1983), Compensating for Missing Survey Data. Survey Research Center, University of Michigan, Ann Arbor, MI.

OECD, Short-Term Economic Statistics (STES) Timeliness Framework. http://www.oecd.org/std/short-termeconomicstatisticsstestimelinessframework.htm

Rao, J. N. K., Srinath, K. P., and Quenneville, B. (1989), Estimation of Level and Change using Current Preliminary Data. In: Kasprzyk, Duncan, Kalton, and Singh (eds.), *Panel Surveys*, John Wiley & Sons, New York, 457–485.

Rizzo, L., Kalton, G., and Brick, M. (1996). A Comparison of Some Weighting Adjustment Methods for Panel Nonresponse. *Survey Methodology* **22**, 43–53.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992), *Model Assisted Survey Sampling*. Springer Verlag.

Särndal, C.-E. and Lundström, S. (2005), *Estimation in Surveys with Nonresponse*. John Wiley & Sons, New York.

Specific section

8. Purpose of the method

The method is used for the preliminary estimation of the target variable, with the aim to obtain the estimates relying on statistical information available at time preceding the time *t*, i.e., on the basis of only a set of quick respondents which define the so-called preliminary sample.

9. Recommended use of the method

1. When model-based method cannot be used because auxiliary variables are not available or the time series is not long enough.

10. Possible disadvantages of the method

1. The improvement of the revision error can be weak.

11. Variants of the method

1.

12. Input data

1. Final estimates of preceding time *t*-1 and standard preliminary estimates at time *t*.

13. Logical preconditions

1. Missing values
    1. Not applicable.
2. Erroneous values
    1. Not applicable.
3. Other quality related preconditions
    1. Not applicable.
4. Other types of preconditions
    1. Not applicable.

14. Tuning parameters

1. Alpha (α) to be evaluated on the basis of variances and covariances.

15. Recommended use of the individual variants of the method

1.

16. Output data

1. Ds-output1 = composite preliminary estimates of the target parameter.

17. Properties of the output data

1. The composite preliminary estimates should guarantee a lower revision error than the direct estimates.

18. Unit of input data suitable for the method

Preliminary and Final Estimates at previous time and preliminary estimates at time *t*.

19. User interaction - not tool specific

1.

20. Logging indicators

1.

21. Quality indicators of the output data

1. Revision errors.
2. Quality assessment of the result.

22. Actual use of the method

1. None.

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Imputation – Main Module

24. Related methods described in other modules

1. Sample Selection – Subsampling for Preliminary Estimates
2. Weighting and Estimation – Preliminary Estimates with Model-Based Methods

25. Mathematical techniques used by the method described in this module

1. Variance-Covariance estimation

26. GSBPM phases where the method described in this module is used

1. 5.6 Calculate aggregates

27. Tools that implement the method described in this module

1. No software tools are available

28. Process step performed by the method

Estimation of target parameters on the basis of information collected on quick respondents.

Administrative section

29. Module code

Weighting and Estimation-M-Preliminary Estimates Design-Based

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	31-12-2012	first version	Claudia De Vitiis	ISTAT
0.2	11-02-2013	first revisions	Claudia De Vitiis	ISTAT
0.3	30-09-2013	revised according to review by Norway	Claudia De Vitiis	ISTAT
0.3.1	07-11-2013	revised according to review by Editorial Board	Claudia De Vitiis	ISTAT
0.3.2	13-11-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	26-3-2014 13:33