

Weighting Procedure for Longitudinal Data of the Italian Labour Force Survey

Procedura di Riporto per i Dati Longitudinali della Rilevazione Trimestrale sulle Forze di Lavoro

Antonio R. Discenza

ISTAT – Servizio Formazione e Lavoro, Roma, discenza@istat.it

Riassunto: La rilevazione trimestrale sulle forze di lavoro contiene una componente longitudinale che permette la costruzione di archivi longitudinali e la produzione di matrici di transizione. La particolare natura del disegno dell'indagine impone la specificazione di una popolazione di riferimento “longitudinale” e condiziona la conseguente scelta della procedura di riporto all'universo. Per la stima di dati longitudinali a 12 mesi è stata implementata una procedura di riporto a più passi che, mediante post-stratificazione, sia in grado di vincolare le stime alla popolazione longitudinale, riesca ad attenuare la distorsione dovuta alla mancata risposta e garantisca la congruenza delle stime longitudinali con quelle “ufficiali” diffuse trimestralmente dall'Istituto Nazionale di Statistica.

Keywords: longitudinal data, wave non-response, weighting, raking-ratio

1. Introduction

The Labour Force Survey (LFS) is the most important source of information about the Italian labour market; each quarter it provides official estimates for a number of relevant figures. It is carried out every year in January, April, July and October by using a rotating sample design. The survey was designed to produce cross-sectional data but, as each sample household is interviewed for 4 quarters, the records can be linked together to produce a rich source of longitudinal data at 3, 9, 12 and 15 months (Ceccarelli et al., 2002). Two-quarter longitudinal data have been produced linking existing LFS data. A number of methodological problems could distort the data resulting from this linking. Specific solutions and a new weighting procedure compensating for the biasing effects of non-response have been created. A full series of 12 months longitudinal data has been produced starting from April 1993. It can be used to produce estimates of gross flows over time between different economic activity status.

2. Methodological issues

Given the LFS sample design, several methodological problems have been addressed to obtain longitudinal data.

From the sample design 50% of the cross-sectional household sample for each quarter should be re-interviewed 12 months later. Producing linked records seems to be

relatively straightforward. Unfortunately, a unique identification variable for each individual in every quarter is not available, then record linkage is carried out combining several key variables (household code, sex, date of birth, educational level, civil status, relation to head). Actually, these variables are affected by errors of several types (response, coding, editing, etc.) reducing the performance of a deterministic record linkage. For this reason a probabilistic procedure has been used to match also those records whose key variables were affected by errors (Paggiaro, Torelli 1999).

Also, the longitudinal component of the LFS is affected by unit non-response such as:

- Municipality non-response: some municipalities are substituted in April at the beginning of a new annual survey cycle and some others may, for different reasons, fail to provide the interviews in subsequent waves;
- Household and individual non-response: some people do not fill in the questionnaire because they refuse to respond or the interviewers are unable to contact one or more individuals in a household. This kind of non-response can be assimilated to the “false negative” which occurs when the record linkage fails to match two records because of errors in the key variables.

Unit non-response may reduce the longitudinal component, thus increases the variance of the estimates. It can also produce bias if non-respondents have significantly different labour features with respect to respondents.

Another important aspect is that the sample design drives the specification of the longitudinal population of inference for the LFS. In fact, longitudinal data concerns only people who are resident, both at the beginning and at the end of the period, in the same municipality. Hence the “actual” longitudinal population is defined as the initial population excluding deaths, emigrants and people who changed municipality.

Given the longitudinal population defined as above, it is necessary to classify all the individuals interviewed at the first quarter into two groups:

- Eligible: they represent part of the longitudinal population that should be re-interviewed at the second wave;
- Not-eligible: they left from the initial population during the observed period thus they do not represent part of the longitudinal population.

This is a very important theoretical concept but, as the LFS is not a panel survey, the individuals can not be distinguished in eligible and not-eligible. As consequence, this means that it is not possible to use logistic regression models to compensate for longitudinal non-response (wave non-response).

The last problem is that the longitudinal component produces both cross-sectional and longitudinal estimates referred to the longitudinal population. These cross-sectional estimates have to be consistent (i.e. not higher than) with the “official” estimates provided by the cross-sectional samples at the beginning and end of the observed period. The differences between these two kind of cross-sectional estimates must be non-negative because they refer to the labour activity status at the beginning and end of the period for people who left the initial cross-sectional population and for people who entered in the final cross-sectional population, respectively. Given that longitudinal estimates have higher variability than “official” estimates then it is not possible to control their consistency completely. However, it is possible to reduce the risk of obtaining inconsistent results by using specific constraints in a weighting process.

3. Weighting procedure

Some methods dealing with longitudinal non-response are based on logistic models (Rizzo et al, 1996). These models are not really useful for the LFS because the eligible unmatched individuals, that should be considered for adjusting the weights of similar matched individuals, are not distinguishable from the non-eligible unmatched ones.

Generalised raking procedures have been used to take into account for, and so reduce, the bias due to wave non-response. Raking involves modifying the original weights in order to simultaneously satisfy several marginal constraints (or to control totals) while minimizing the distance between original and adjusted weights. Final longitudinal weights have been obtained after two raking-ratio weighting stages: the first accounting for bias due to municipality non-response and due to differences in rotation groups; the last adjusting for bias due to individual non-response and to make weighted sample totals conform to the longitudinal population.

All the individuals *linkable* at the beginning of the period (i.e. all the individuals of the two rotation groups which overlap at 12 month and resident in those municipalities which provided interviews for both waves) have been selected; they can be considered like a random sub-sample of the whole cross-sectional sample. Let π_i denote the probability of being selected for the cross-sectional sample at the beginning of the period and let $k_i = 1/\pi_i$ denote the base cross-sectional weight. Then, for *linkable* individuals, base longitudinal weights k_i^* are obtained from k_i applying the following correction

$$k_i^* = \left(k_i / \sum_{i \in \text{Linkable}} k_i \right) {}_1P \quad (1)$$

where ${}_1P$ is the cross-sectional population at the beginning of the period.

In order to ensure consistency between longitudinal and cross-sectional “official” estimates, the first raking procedure makes *linkable* individuals at the first quarter representing exactly the same cross-sectional population of the whole cross-sectional sample. Moreover, it makes them provide the same cross sectional “official” estimates for a number of relevant figures (cross-classification of sex, region, age group, labour activity status, education, etc.). Thus, from the base longitudinal weight k_i^* and for all the *linkable* individuals the initial weight $g_i^* = k_i^* \delta_i^*$ has been obtained as result of a minimum constrained problem as follows:

$$\sum_{i \in \text{Linkable}} g_{i,rse}^* = \sum_{i=1}^{N_1} w_{i,rse} = {}_1P_{rse} \quad (2)$$

$$\sum_{i \in \text{Linkable}} g_i^* X_i = \sum_{i=1}^{N_1} w_i X_i \quad (3)$$

$$\min \left\{ \sum_{i \in \text{Linkable}} D(k_i^* \delta_i^*, k_i^*) \right\} \quad (4)$$

where w_i is the first quarter cross-sectional final weight, ${}_1P_{rse}$ are the known totals for the first quarter cross-sectional population for area r , sex s and age group e , N_1 is the first quarter cross-sectional sample dimension, X_i are auxiliary variables, D is a logarithmic distance function (Deville, Sarndal, 1992).

In the second stage, the final longitudinal weight $w_i^* = g_i^* \gamma_i^*$, only for *linked* individuals, have been obtained applying a different raking step to make weighted longitudinal component totals conform to totals derived from the longitudinal population $\sum_{i \in \text{linked}} w_{i,rse}^* = {}_lP_{rse}$ under the following constraints

$$\min \left\{ \sum_{i \in \text{Linked}} D(g_i^* \gamma_i^*, g_i^*) \right\} \quad (5)$$

where ${}_lP_{rse} = {}_1P_{rse} - m_{rse} - c_{rse}$ is the known longitudinal population.

4. Conclusions

The procedure described as above has been applied to nine yearly longitudinal datasets from April 1993 to April 2002 and transition matrices have been produced. It has been found that it reduces the negative effects of non-response bias providing consistent estimates for almost all the relevant figures. Few inconsistencies have been found for not-significant estimates and for those variables more likely subject to classification errors. Further research will be carried out for the estimation of the variances for the gross flows estimates through non-parametric methods.

References

- Ceccarelli C., Discenza A. R., Rosati S., Paggiaro A., Torelli N. (2002) Le matrici di transizione della rilevazione trimestrale sulle forze di lavoro, *Nota metodologica*, Istat, Roma
- Paggiaro A., Torelli N. (1999) Una procedura per l'abbinamento di record nella rilevazione trimestrale delle forze di lavoro, *working paper*, 15, CLEUP, Padova.
- Tate P.F., Clarke P.S., (1999) Methodological issues in the production and analysis of longitudinal data from the Labour Force Survey, *GSS Methodology Series*, 17.
- Falorsi S., Rinaldelli C. (1998) Uso di un software generalizzato per il calcolo delle stime e degli errori di campionamento, *Statistica Applicata*, Vol. 10, n.2.
- Rizzo L., Kalton G., Brick J.M. (1996) A Comparison of Some Weighting Adjustment Methods for Panel Nonresponse, *Survey Methodology*, 22, 43-53.
- Deville J.C., Sarndal C.E. (1992) Calibration Estimators in Survey Sampling, *Journal of the American Statistical Association*, 87, 376-382.