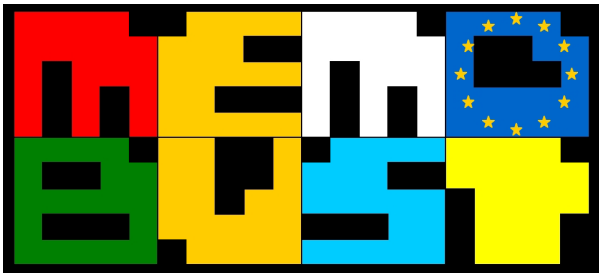

This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Editing for Longitudinal Data

Contents

General section

1. Summary

We refer to longitudinal data as repeated observations of the same variables on the same units over multiple time periods. They can be collected either prospectively, following subjects forward in time, or retrospectively, by extracting multiple measurements on each unit from historical records. The process of Editing and Imputation can exploit the longitudinal characteristic of the data as auxiliary information, useful at both the editing and the imputation stages. This theme describes the editing process applied to longitudinal data, that could be performed for all aforementioned types of data, with special focus on Short Term Statistics context.

2. General description

2.1 Longitudinal data

Another term for longitudinal data is panel data. This definition focuses on the particular sample, which units are selected to be observed several times with some degree of regularity. The occurrence of those observations can be once along several years (every four years or biannual) or once a year (annually) or several times during the same year (quarterly or even monthly). Panel data are mostly used to describe patterns of change within and between the statistical units under observation, in other cases to highlight and to identify differences and changes over time of a specific parameter of the population under study. In general, for each unit $i = 1,\ldots,n$ there are $t = 1,..,T$ different measurements, one for each wave of interview. The period t can be a month, a quarter or a year; the first two cases drive to infra-annual longitudinal data. As a consequence, given the period t, a vector of cross-sectional observations is available, while as regards the i-th observation a vector of longitudinal data is available and a strong correlation is expected among its values. According to the type of required estimates, different types of panel are considered, so it can always follow the same units or rotate some of them after a period (rotating panel). The different design will create different type of longitudinal data set.

In the context of business statistics, longitudinal data can be used both in structural and in short-term analysis. The difference between Structural Business Statistics (SBS) and Short Term Statistics (STS) actually depends on the combination of the survey occurrence and the type of final target parameter; see also the modules “General Observations – Different Types of Surveys” and “Repeated Surveys – Repeated Surveys”. In the SBS context, totals, means, levels are usually the object of the estimates; in the STS the main objective is usually to publish regular series of statistics on changes of totals for specific domains. These are frequently published in the form of index numbers, whose main purpose is to measure net changes between two periods. In these cases the rationale for a panel design is to improve the precision of estimates, because the minor variance of estimates is assured by the presence of historical correlation between data referred to the same units over the period in which the observations take place; see also the topics “Sample Selection” and “Weighting and Estimation”. On the other hand, also from an operational point of view, the use of a panel for an infra-annual survey can yield important cost savings. Indeed, to interview the same units is often less expensive than starting afresh, at each wave, the contacts on new units.

2.2 *Introduction to editing for longitudinal data*

In general, two main aspects are crucial in an editing process framework:

1) the rule to identify an acceptance region for a test variable;

2) the technique used to change a value detected as wrong during the process.

In a longitudinal context, these aspects have to be fitted to the specific target parameter, which is often given by the estimation of the change of a population parameter (mostly the mean) concerning a quantitative not-negative variable y. It is strongly recommended to use the available historical information of the observation units for two main reasons:

1) a strong correlation is expected among different measurements of the same variable on the same units, thus any detecting rule can rely on relevant information about the unit profile and can result in being more efficient;

2) since most of the time the target parameter is the change of a main parameter along time, any observed change between sequential periods on the observations can be used as a precious source of information with regards the final estimation.

In general, the editing process in a longitudinal context must take into account the characteristics of the change under investigation and the timeliness constraints. The control rules can be defined taking into account comparisons between values of the same variable on the same unit at different times, i.e., the two values y_t and y_{t-k}, where t is a month or a quarter, t-k is a previous period and k varies according to the variable features and/or to the type of change under observation. Additional specifications are generally required, they are briefly described in the following.

2.3 *Editing scheme in a longitudinal context*

When the editing process is set on longitudinal data, there are some issues which assume a strategic meaning:

1) Longitudinal and cross-sectional checks can be carried out at the same time; this is because longitudinal surveys keep a statistical relevance for cross-sectional analyses as well. For instance, a certain variable x may have a direct connection with the target variable y and, as a consequence, a specific cross-sectional check is needed. In this case, a troublesome decision concerns the priority level among the cross-sectional and the longitudinal checks, even though the last ones should come first. Thus, it is important to coordinate them in order to avoid the risk to oversize the overall number of checks as well as the amount of changes carried out on the original micro-database (Granquist and Kovar, 1997). On the other hand only cross-sectional checks may be applicable in case of "new" units, for which no past data are available.

2) Given the target parameter and the characteristics of the variable under investigation, at each reference time t there is the need to specify which are the previous periods to be considered in the editing process. For example, for monthly data the periods t-1 and t+1 or t-12 and t+12, most of the times because of the presence of significant seasonal components.

3) Economic units may change their demographic features over time (such as change of their ownership, location, economic activities carried out, number of local units, employment and so on) as a result of events of different nature (i.e., mergers or splits). Statistical units interested by

these changes could lose their "longitudinal" identity and their data cannot be compared in a longitudinal data analysis process. As a consequence suspected changes may come up, which are not the results of real mistakes, but they are due to structural changes of the unit economic profile along time. In a longitudinal survey context – in particular, in a short-term survey framework – it is often difficult: a) to identify cases when there are anomalous increases or decreases due to demographic changes and not to real measurement errors (lack of updated information even from the business register); b) to apply a proper amendment to microdata able to overcome the non-comparability of data over time.

4) In a short-term survey framework, the required timeliness for the elaboration of the indicators becomes a hard constraint for the editing strategy, as it strongly reduces the available time to check all the microdata. It is a good solution to identify a sub-set of "critical" units, for which a deeper analysis can guarantee the required quality. This approach is generally defined as *selective editing*, which presumes the definition of a *score function* to rank the observations according to their impact on the target estimates; see the module "Statistical Data Editing – Selective Editing". Several score functions are proposed in literature, the difference among them is mainly given by the way to measure the impact on the final estimates, that anyway usually depends on: i) the given sampling weights; ii) the size of the possible error; iii) the longitudinal behaviour of each respondent.

2.4 Type of edits

The error detection process usually consists of a set of integrated error detection methods dealing each with a specific type of error (EDIMBUS, 2007), which results are flags pointing to missing, erroneous or suspicious values. Error detection is often based on the use of edit rules, that are restrictions to the values of one or more data items that correspond to missing, invalid or inconsistent values potentially in error (cf. "Statistical Data Editing – Main Module"). In a longitudinal context, the coherence of individual historical data is the basic rationale to analyse the data, because the units are believed to be strongly characterised by their own longitudinal profile. According to this point of view, the data of each unit at the occasion t can be checked by comparison with other values observed on the same unit at other times, i.e., belonging to its profile, with regards to an expected value or range.

In the following, the typology of edits is described according the needs and the features of a longitudinal context:

- Consistency checks: their purpose is to detect whether the value of two or more variables on the same unit are in contradiction, hence, whether the values of two or more data items do not satisfy some predefined expected relationship. In this regard, comparisons with other sources which produce comparable microdata are included. Data items can refer also to measurement on the same unit in different periods, it is important that this reference data has been previously checked for errors[1]. The reference data used and the way in which the comparison takes place depend on the target parameter.

[1] If the past value y_{t-k} refers to the previous year, past data can be supposed to have been fully checked on the basis of information available from sources external to the survey, so that normally suspect ratios y_t/y_{t-k} lead to change the actual value y_t (but not the past value). However, this rule is not rigid and past data may be changed as well (that is the case of wrong reporting by some units which can review past values even one year later).

- Balance edits: often the value of a variable at time t can be obtained by the sum of the values in the previous period and the registered flow in the reference period for that variable; e.g., the number of persons employed at the end of month t-1, plus the number of persons who started working between months t-1 and t, minus the number of persons who stopped working between months t-1 and t, must be equal to the number of persons employed at the beginning of month t.
- Check for unity measure errors: some errors are due to misunderstandings about the measure according to which a variable x is collected, e.g., thousand instead of billion and so on. In these cases, there is a thousand-error if one of the following relations is verified:

$$\text{abs}(x_t) > h \cdot [\text{abs}(x_{t-k})] \qquad \text{for some } k \in \{1,\ldots,P\} \tag{1a}$$

$$h \cdot [\text{abs}(x_t)] < \text{abs}(x_{t-k}) \qquad \text{for some } k \in \{1,\ldots,P\} \tag{1b}$$

  where $x_{t-k}>0$, abs(x) is the absolute value of the variable x and h is a constant to be chosen properly by the expert.
- Ratio edits. These edit rules are bivariate restrictions taking the general form $a \leq x / y \leq b$, where x and y are numerical variables and a and b are constants. In a longitudinal context, the comparison is based on the two measurements y_t and y_{t-k} , k will vary according to case under study (type of data, characteristics of the variable, etc.).
- A further type of edit is related to a specific feature of longitudinal surveys, because it is possible to ask twice for the same data, with reference to the same variable for the same period. Normally, it happens when a certain value is asked in two consecutive waves at times t-1 and t. Let $y_{it(t-1)}$ be the value of the variable y on the unit i asked in the wave t even though referred to the t-1 period, then a frequent longitudinal check is given by:

$$y_{it(t)} = y_{it(t-1)} \tag{2}$$

  This option may help both to check for the quality of supplied longitudinal information and to take under control changes of some accounting figures inside the unit; it is also very useful to achieve longitudinal data from units characterised by wave non response, e.g., those units which may be non-respondent in t-1 and respondent in t, or vice-versa. This solution has to be defined accurately, in order to be worth without increasing the statistical burden on the respondent units.

2.5 Methods for longitudinal data

In a longitudinal context, one of the most relevant test variables is the "individual trend" or "individual change", defined as:

$$c_{it} = y_{it} / y_{it-k} \tag{3}$$

As a consequence most data controls are based on the study of (3) and on rules to check whether the individual trend is too large or too low. The main issue is to define a criterion to decide whether a given level satisfies or not the acceptation rules. The unit trend information can be used in different ways, a couple of them is shortly resumed as follows.

2.5.1 The Hidiroglou-Berthelot method for detecting outliers

The empirical distribution of all the individual trends can supply useful information for the editing process, by comparing each c_{it} with some main indicators of such distribution. In this regards, the

Hidiroglou-Berthelot method (Hidiroglou and Berthelot, 1986) proposes a way to establish an acceptance interval for c_{it}, based on a function of its interquartile, in order to detect outliers.

Firstly, for each occasion t the median of all the c_{it} is elaborated, defined as $q_{0.5}(c_t)$. Afterwards, a transformation is applied to every c_{it}, to ensure more symmetry of the distribution tails:

$$s_{it} = \begin{cases} 1 - q_{0.5}(c_t)/c_{it}, & \text{if} \quad 0 < c_{it} < q_{0.5}(c_t) \\ q_{0.5}(c_t)/c_{it} - 1, & \text{if} \quad c_{it} \geq q_{0.5}(c_t) \end{cases} \tag{4}$$

Let also define:

$$E_{it} = s_{it} \cdot \{\max(y_{it}, y_{it-1})\}^{U} \tag{5}$$

which is the "effect" concerning unit i at time t; it is based on the "individual trend" component s_{it} defined by (4) and the "size" component due to the y-levels of the same unit. The parameter $U \in [0,1]$ is a tuning parameter which should balance the magnitude of the size component with respect to the individual trend. Then, given the first and the third quartile, $q_{0.25}(E_t)$ and $q_{0.75}(E_t)$, the following values are defined:

$$D_1 = \max\{q_{0.5}(E_t) - q_{0.25}(E_t), A \cdot q_{0.5}(E_t)\} \tag{6}$$

$$D_3 = \max\{q_{0.75}(E_t) - q_{0.5}(E_t), A \cdot q_{0.5}(E_t)\} \tag{7}$$

where the constant A is chosen to avoid difficulties which can arise when the differences $q_{0.5}(E_t) - q_{0.25}(E_t)$, and $q_{0.75}(E_t) - q_{0.5}(E_t)$ are small (generally it is set to 0.05).

Hence, the acceptance region is defined as follows:

$$(q_{0.5}(E_t) - A \cdot D_1,\ q_{0.5}(E_t) + A \cdot D_3) \tag{8}$$

and each observation y_{it} which falls out of such interval is considered to be an outlier.

It is worthwhile to underline how the identification of anomalous ratios c_{it} due to errors (not necessarily outlier observations) may be carried out according to an analogous methodological scheme.

2.5.2 Score functions ranking

In case a selective editing scheme has to be defined, the basic rationale is the evaluation of the impact of the change of each unit on the overall trend, considering its size and its sampling weight. This kind of analysis can be carried out ranking the units on the basis of a score function, which takes into account the above mentioned dimensions. Thus, a simple score function to be applied to each unit depends on the three dimensions:

Score = (longitudinal trend) x (sampling weight) x (size).

In the following, a score function is described that takes these elements into account, for which a transformation of the individual trend c_{it} is defined in order to take into account different options of needs. A preliminary transformation is made to assign high priority to units characterised by either a very high or a very low change:

$$d_{ij} = \max(c_{it}, 1/c_{it}) = \max(y_{it}/y_{it-k}, y_{it-k}/y_{it}) \tag{9}$$

New units, for which no historical data are available, will be assigned $c_{it}=1$.

Then, the following conversion will be used to define the final score function:

$$r_{it} = |k_{1i} d_{it} - k_{2i}|$$

where k_{1i} and k_{2i} can be chosen according to any needs expressed by the given survey, a typical choice is to put both k_{1i} and k_{2i} equal to 1.

Thus, the score function for a generic unit i and a given time t can be built up as follows:

$$\Phi_{it} = r_{it}^{\alpha} w_{it}^{\beta} z_{it}^{\gamma} \qquad (10)$$

where w is the sampling weight and z is a "size" variable (for instance, turnover, production, number of persons employed). Parameters α, β and γ should be used in order to balance the relative importance of each score component on the final score Φ. Normally it is recommended to use parameter values chosen from the interval [0,1] (Gismondi and Carone, 2008). After the calculation of the score (10) for each unit, scores can be ordered in a non-decreasing ranking: the units occupying the "first positions" in the ranking will be detected as influent suspicious units, to be checked with priority or even re-contacted. Some techniques for assessing the number of influent units have been proposed by McKenzie (2003), Philips (2003), Chen and Xie (2004).

2.6 The case of categorical data

There are particular kinds of business longitudinal surveys for which categorical variables play a fundamental role. That may happen when the main goal:

a) is still the evaluation of the change of a quantitative variable, but a preliminary step consists in the assessment of the presence (or absence) of a certain phenomenon (binary variable: 1=present, 0=absent);

b) consists in the evaluation of a set of opinions and their developments over time (qualitative variables).

An example of the kind a) is the survey on job vacancies. The main goal is the estimation of the number of job vacancies at the end of each quarter, but a preliminary step consists in assessing if an enterprise is searching for new personnel or not. There are the following possibilities:

- The firm declares an amount of job vacancies higher than zero, that implies the firm is searching for new staff. In this case no problem is encountered.
- The firm declares zero job vacancies. This value may be right, but it may be wrong as well, for instance, because the firm is not able to correctly count the number of job vacancies (and prefers to declare zero in order to tackle the question quickly). A signal in favor of a potential error may be given by a simple ex post longitudinal check: the comparison between the number of persons employed at times (t+1) and (t). If the former amount is higher than the latter, it is not possible that the number of job vacancies declared at time (t) was zero.
- The firm does not declare anything. Also in this case, longitudinal checks may be useful for making proper changes, but they may not be enough and the binary variable presence/absence of job vacancies will be object of estimation (for instance, using a logistic model where the explicative variables are often given by past responses provided by the same unit) or will be asked again to the firm (when it will be possible, according to budget and time constraints).

An example of the kind b) is given by tendency surveys. Tendency surveys concern enterprises and consumers and are aimed at asking a series of qualitative questions related to economic situation, household budget, purchases planning, employment, prices, etc. Questions ask for opinions concerning the development of each issue with respect to a previous period. Normally response modalities are: i) strong increase, ii) increase, iii) no change, iv) decrease, v) strong decrease. Macro figures are calculated as weighted differences between optimistic opinions i)+ii) and the pessimistic ones iv)+v). In tendency surveys main quality checks do not refer explicitly to past longitudinal data. This may be due to the use of rotated samples and/or to the weak correlation between responses provided by the same unit in two consecutive survey waves. The basic control is that for each unit and each question one and only one response must be provided.

3. Design issues

The design of the editing and imputation process should be part of the design of the whole survey process. In the frame of editing and imputation procedures three main logical phases are usually carried out, based on the following actions:

1. Identification and elimination of errors that are evident and easy to treat with sufficient reliability, that can involve both interactive and automatic methods;
2. Selection and treatment of influential errors through a careful inspection of influential observations; automatic treatment of the remaining non influential errors, through a selective editing procedure;
3. Check of the final output looking for influential errors that have been undetected in the previous phases or introduced by the procedure itself, that involves macro-editing procedures.

In a longitudinal context, the identification and the calculation of a set of indicators based on macrodata may be based on ratios between the same macrodata related to two different periods, where macrodata of the previous period are supposed to be good (already validated at previous occasions). If the macro indicator falls inside an acceptation range, then no other controls are needed, otherwise it is necessary to go back to microdata and to run again all or a part of controls already activated in the previous micro-editing phase a). Usually, acceptations intervals for macro indicators are determined according to subjective choices by survey experts.

Finally, in the last phase, provisional publication figures are elaborated and analysed using historical data or external sources. If the aggregate figures are implausible, the individual records are examined in order to check for further outliers or error affecting influential records; in these cases data can be modified if necessary. The errors detected at this stage may have been not individuated in the earlier phases of the editing process, or may have been introduced by the process itself. Anyway, also every treatment of these kinds of errors is always made at micro level. If the provisional figures are plausible, the detection of errors and their treatment process is concluded.

The edited file is used in the subsequent statistical process for aggregation purposes, for the estimation of totals and for further analyses.

4. Available software tools

5. Decision tree of methods

6. Glossary

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

7. References

Chen, S. and Xie, H. (2004), Collection Follow Up Score Function and Response Bias. *Proceedings of the SSC Annual Meeting – Survey Methods Section*, Statistics Canada, 69–76.

EDIMBUS (2007), *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*. http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/RPM_EDIMBUS.pdf.

Gismondi, R. and Carone, A. (2008), Statistical criteria to manage non-respondents' intensive follow up in surveys repeated along time. *Rivista di Statistica Ufficiale*, 1/2008, 5–29.

Granquist, L. and Kovar, J. (1997), Editing of Survey Data: How Much Is Enough? In: L. E. Lyberg, P. Biemer, M. Collins, E. D. De Leeuw, C. Dippo, N. Schwartz, and D. Trewin (eds.), *Survey Measurement and Process Quality*, John Wiley & Sons, New York, 415–435.

Hidiroglou, M. A. and Berthelot, J. M. (1986), Statistical Editing and Imputation for Periodic Business Surveys. *Survey Methodology* **12**, 73–83.

McKenzie, R. (2003), *A Framework for Priority Contact of Non Respondents*. Available at: www.oecd.org/dataoecd.

Philips, R. (2003), The Theory and Application of the Score Function to Prioritize and Limit Recontacts in Editing Business Surveys. *Proceedings of the SSC Annual Meeting – Survey Methods Section*, Statistics Canada, 121–126.

Interconnections with other modules

8. Related themes described in other modules

1. General Observations – Different Types of Surveys
2. Repeated Surveys – Repeated Surveys
3. Sample Selection – Main Module
4. Statistical Data Editing – Main Module
5. Statistical Data Editing – Selective Editing
6. Weighting and Estimation – Main Module

9. Methods explicitly referred to in this module

1.

10. Mathematical techniques explicitly referred to in this module

1.

11. GSBPM phases explicitly referred to in this module

1. 2.5 Design statistical processing methodology
2. 5.3 Review, validate and edit

12. Tools explicitly referred to in this module

1.

13. Process steps explicitly referred to in this module

1. Data validation

Administrative section

14. Module code

Statistical Data Editing-T-Longitudinal Data

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	05-02-2013	first version	Roberto Gismondi, Maria Liria Ferraro, Anna Rita Giorgi, Fabiana Rocci	Istat (Italy)
0.2	30-05-2013	second version	Roberto Gismondi, Maria Liria Ferraro, Anna Rita Giorgi, Fabiana Rocci	Istat (Italy)
0.3	20-08-2013	third version (accepted corrections)	Roberto Gismondi, Maria Liria Ferraro, Anna Rita Giorgi, Fabiana Rocci	Istat (Italy)
0.4	15-11-2013	fourth version	Roberto Gismondi, Maria Liria Ferraro, Anna Rita Giorgi, Fabiana Rocci	Istat (Italy)
0.4.1	20-12-2013	preliminary release		
0.4.2	08-01-2014	final release	Roberto Gismondi, Maria Liria Ferraro, Anna Rita Giorgi, Fabiana Rocci	Istat (Italy)
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:13