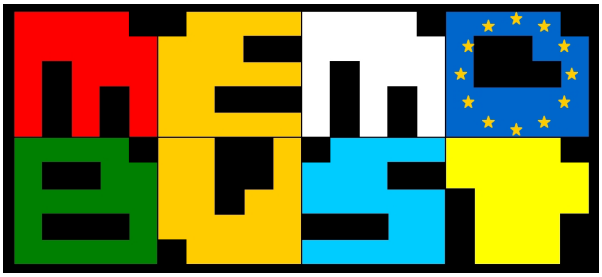

This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Imputation for Longitudinal Data

Contents

General section

1. Summary

We refer to longitudinal data when the same variables of the same units are measured several times at different moments. The common trait is that the entity under investigation is observed or measured at more than one point in time, possibly regularly, in order to study how it develops over time. The data are collected either prospectively, following subjects forward in time, or retrospectively, by extracting multiple measurements on each unit from historical records. Also data from registers can be referred to as longitudinal data, indeed it is possible to match historical data about the same units once they are available with some degree of regularity.

This theme is due to describe the methods for imputation of missing longitudinal data, that could be performed for all aforementioned types of data. Particular emphasis is focused on the Short Term Statistics context.

2. General description

2.1 Longitudinal data

Longitudinal data are typically the result of a repeated survey, whose purpose is to collect data on the same observation units along several years (e.g., every four years or biannual) or once a year (annually) or several times during the same year (e.g., quarterly or even monthly). In the context of business statistics, longitudinal data can be used both in structural and in short term surveys. The combination of the periodicity and the type of parameter to be estimated can determine the difference between Structural Business Statistics (SBS) and Short Term Statistics (STS) (see the modules "General Observations – Different Types of Surveys" and "Repeated Surveys – Repeated Surveys"). In a short-term statistics context the parameter to be estimated is usually the change of a certain indicator along time.

In general, longitudinal data can be represented as data collected on the same units several times in a consecutive sequence, hence for each unit $i=1,\ldots,n$ belonging to the sample, 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 intra-annual longitudinal data. It is clear that, given the period t, a vector of cross-sectional observations is available, while as regards the i-th observation a vector of longitudinal data on the same unit is available and a strong correlation is expected among these values (see the module "Statistical Data Editing – Editing for Longitudinal Data").

2.2 Introduction to imputation for longitudinal data

In statistical surveys, respondents sometimes do not provide answers to one or more questions, while they are required to do that. Commonly, two cases are distinguished: the *item non-response* (or *partial non-response*) is when the unit answers to the survey, but it does not provide information about one or more questions; the *unit non-response* case is when the observation unit does not respond at all. In a longitudinal context, these cases can vary also with respect to the specific time t the data are related to, hence, the missing values come into two forms:

a) scattered missing values: item or total non-response, because units do not answer to some questions or to the total questionnaire in one or more waves, but they deliver the whole records in other waves. Most of the times the high timeliness of the STS increases late answers with respect to the deadline, so that their data are available afterwards;

b) panel drop-out: starting from a specific time *t* some units stop to answer. This phenomenon is called panel *attrition* (Kalton, 2009).

In the case of longitudinal data, the unit dropout is often the greatest concern, because it could hide a major reason for not answering and it should be considered to systematically behave in a different way compared to the units which give response to the survey, even if not at every wave. In these cases, it is suggested to investigate the event, to discover whether the unit has been modified by a demographic event (see the module "Statistical Registers and Frames – The Populations, Frames, and Units of Business Surveys") that could change the composition of the panel.

Where imputation of missing values is required, there are two possible approaches according to the dimension. On one side, for each occasion *t* a set of cross-sectional data is available, for which all the described methods are applicable. On the other side, for each unit *i* a longitudinal vector is available, for which also other methods can be applied that would take into account the information from other measurements on the same units.

There are two main reasons to use longitudinal imputation techniques instead of the cross-sectional methods:

1. Earlier or later observations of the same object are generally very good predictors for the missing value. This means that the quality of the imputation can strongly be improved.
2. To correctly estimate changes of a variable over time (typically the final aim of a short-term survey), the imputation of missing values should take into account information about the previous and the future values of the given variable on the same unit under observation, that supplies useful evidence about their change over time.

It must be observed that the use of cross-sectional methods is unavoidable in case of missing or incorrect information referred to units included for the first time in a rotating panel, as no historical data are available for these units.

Imputation of missing values can be derived from other characteristics of the unit under study (see the module "Imputation – Deductive Imputation"), when also values recorded in other occasions are available the same rule can be applied. In other cases, auxiliary information is available and it makes prediction model of the missing values possible, which is supposed to generate the data (see the module "Imputation – Model-Based Imputation"). These models can be applied also in the case of a longitudinal context, once the proper auxiliary variable has been settled to be the measurement of the same variable on the given unit in another occasion. The choice of the imputation method usually depends on the characteristics of the variable under observation. In the longitudinal context the different pattern of seasonality should also be taken into account, as it determines important features of the variable (for instance, the number of monthly hours worked depends on the number of working days in the same month).

Many methods are based on the assumption that data are originated from a multivariate normal distribution. These methods should be applied carefully to data coming from business surveys,

because the above mentioned hypothesis is not valid in case of concentration of enterprises. In particular cases such as for very big enterprises, it is worth identifying a specific imputation method which takes into account the profile of the units themselves in order to improve quality of final estimates.

This is the reason why an a priori analysis of each variable under study is recommended, in order to choose the proper kind of historical data to be used for the imputation as auxiliary information

2.3 Imputation methods

Imputation methods considerably depend on the type of data set, its extent and the characteristics of the missingness mechanism. Those for longitudinal data usually take into account the historical information of each unit to define any type of imputation method (both for the deductive imputation and as auxiliary information). Let y_{it} be a missing value of unit i at period t on variable y. Then y-values of unit i at previous and subsequent periods can be used to create an imputed value $\tilde{y}_{it}$. The longitudinal imputation methods are briefly described in the following sections.

2.3.1 Last observation carried forward

In this case, the last observed value of a unit is used for the values of the later periods that must be imputed, that is called Last Observation Carried Forward (LOCF). It is often used in practice, even though it may have some problems (Israëls et al., 2011; Watson and Starick, 2011).

This method is mainly applicable to categorical variables, for which it is known that their change is very little over time. For the quantitative variables, it risks to produce an overly stable picture of the actual situation.

2.3.2 Interpolation or historical imputation

In this case, missing observations can be estimated from both previous and later observations; obviously, in the case of current surveys data can be imputed only using previous observations. Different versions of the method include correction based on a trend component (Israëls et al., 2011).

In the case data exhibit a specific periodical pattern, it is recommended to use data from the same period (in short-term statistics the historical data of one season ago, i.e., one year ago, one month ago).

For the unit i, $\tilde{y}_{it}$ is determined by a function of the K observations from the past and L observations from the future. Interpolation can be used for quantitative variables in a situation where it is difficult to make any model assumption on the variable under study, because there is neither correlation with previous measurement of the same variable nor with other variables in the same context. For quantitative variables, the following rather general formula is suggested:

$$\tilde{y}_{it} = \frac{\sum_{k=1}^{K} w_{-k}\, y_{it-k} + \sum_{l=1}^{L} w_l\, y_{it+l}}{\sum_{k=1}^{K} w_{-k} + \sum_{l=1}^{L} w_l} \qquad (1)$$

with weights $w_{-1} \geq w_{-2} \geq \ldots \geq w_{-K}$ and $w_1 \geq w_2 \geq \ldots \geq w_L$; this means that y_{it} has a smaller weight in both directions from period t, as periods k and l are further away from period t. The weights can be freely

selected, for example, it is possible to choose $K=L$ and $w_k=w_{-k}=1/k$. When only information from the past is used or in the case of panel drop-out, the weights $w_1,\ldots,w_L$ are all equal to zero.

If an intra-annual value has to be estimated, the interpolation formulas can be adjusted in order to take into account the seasonal pattern.

The general formula (1) can be applied in several cases, one example is the linear interpolation between the preceding and the subsequent observation of the same unit, for which the equality $w_1=w_{-1}$ is usually considered:

$$\tilde{y}_{it} = \frac{w_1(y_{it-1} + y_{it+1})}{2w_1} = \frac{y_{it-1} + y_{it+1}}{2} \tag{2}$$

A proposal to determine the weights w_{-1} and w_1 is based on the observed changes on the respondent units of the sample: an indicator variable is created which equals to 1 when the reported change between waves t and t-1 is smaller than the reported change between waves t and t+1 for the complete cases and 0 otherwise. Then, it is possible to calculate the proportion p, which is the share of the interviewed sample for which the change between waves t and t+1 is smaller than the change between the previous wave t and t-1. Hence, the weight $w_1= p$ reflects the probability to change between t and t+1, while $w_{-1}= 1-p$ is about the change between t-1 and t, both reflecting the probabilities associated with the occurrence of change between waves found in the complete cases (Watson and Starick, 2011).

2.3.3 Mean imputation

A missing value is replaced by the mean of valid data. It can be applied both in the longitudinal and cross-sectional view. According to the first one it can be seen as a specific case of the interpolation, where the weights simply represent the presence of each data. The cross-sectional approach is very useful when longitudinal data are not available and the assumption of similar behaviour between respondents and not respondents is valid.

Let y_{it} the response for subject i at occasion t, let y_{it-k} and y_{it+l} be the response of the same unit at time $t-k$ and time $t+l$, and r_{it-k} and r_{it+l} equal to 1 if y_{it-k} and y_{it+l} are observed, 0 otherwise. If y_{it} is missing, it can be replaced by the mean of the nearest preceding and subsequent observations as follows:

$${}_L\tilde{y}_{it} = \frac{\sum_{k=1}^{K} r_{it-k}\, y_{it-k} + \sum_{l=1}^{L} r_{it+l}\, y_{it+l}}{\sum_{k=1}^{K} r_{it-k} + \sum_{l=1}^{L} r_{it+l}} \tag{3}$$

where the time t can vary both along previous observations or future observations. In this case, each missing unit will be replaced by a different value that is strictly correlated to its longitudinal profile. On the other side, the cross-sectional mean response for unit i at time t is equal to:

$${}_{CS}\tilde{y}_{it} = \frac{\sum_{j} r_{jt}\, y_{jt}}{\sum_{j \in obs} r_{jt}} \tag{4}$$

where y_{jt} is the observed value of the j-th respondent at time t and obs is the sample of respondent observations. In this case a cross-imputation is done and the same mean is imputed for each missing value; in this term, it can lead to a peak in the distribution. An alternative version of this method is to

impute a class mean, where the classes may be based on some explanatory variables. This method is influenced by the existence of patterns and similarities between enterprises and, therefore, it has to be carefully evaluated before being used. Anyway, it offers a very good tool in the case where new units have entered the panel and no longitudinal information is available for them. Disadvantages of such procedures are that distributions of survey variables are compressed and relationships between variables may be distorted (Little and Rubin, 2002).

2.3.4 Ratio imputation

Let us suppose that the variable *y,* to be imputed, is strongly correlated to a single auxiliary variable *x* and let a coefficient *R* represent the relationship between the variables *y* and *x* such that *y=Rx* for every unit in the target population. For longitudinal data, the most common situation is that *x* measures a past observation of the same variable *y*, for which it is reasonable to take the assumption that the observation at period *t* is proportional to the observation at period *t-1*. To update the past value to the current time *t* the observed growth on the respondents is used, with respect to the past observed value at time *t*-1. After the pattern of the variable has been determined, it could happen that variable *y* is proportional to the same variable observed at the same month (or quarter) in the previous year, hence, the choice will fall on past observations referred to times *t*-12 or *t*-4 (an example is the case of the hours worked). As a consequence, a missing value can be estimated by increasing the previous observation according to the same proportion of the one observed on the respondent units from time *t-1* to time *t*.

In these terms, the past value y_{it-1} can be used as the auxiliary information to impute y_{it} and the constant *R* is used to link the two historical values. Generally, *R* is not known and it is estimated at every *t* using only those units for which values at both occasion *t* and *t*-1 are known:

$$\tilde{y}_{it} = \tilde{R}_t y_{it-1} = \frac{\sum_{j \in obs} y_{jt}}{\sum_{j \in obs} y_{jt-1}} y_{it-1} \qquad (5)$$

where y_{jt} is the observed value of the *j*-th respondent at time *t* and *obs* is the sample of respondents observations. According to the previous formula, the proportional constant is equal to the ratio between the means of y_t and y_{t-1} calculated using the units respondent in both periods[1].

2.3.5 Regression imputation

The regression of the variable of interest is based on covariates and the resulting equation is used to estimate the missing values. An advantage of longitudinal data is that, in general, the past and/or future observed values of a variable are very good predictors of missing values.

The regression imputation may use both quantitative and categorical variables, in the second case the logistic regression must be used instead of the linear regression. It is considered a good imputation method for business surveys (Kovar and Whitridge,1995), but it should be controlled in case of new developments in the business cycle that are not included in the model.

[1] Where, for example, the variable *y* strongly depends on the number of working days in the reference period (nwd_t), the use of a further multiplier is recommended such as: nwd_t/nwd_{t-1}.

For a missing value y_i, a regression model is assumed for the prediction of y by means of information given by the observed value of the same variable y at previous time t-1, t-2… The regression model is as follows:

$$y_i = \alpha + \beta_{t-1} y_{it-1} + ... + \beta_{t-k} y_{it-k} + \varepsilon_i \quad (6)$$

with α, β_{t-1}, …, β_{t-k} are unknown parameters, $\varepsilon_i \sim N(0, \sigma_i^2 I)$ is the unit residual which is supposed to follow a multivariate normal distribution, where I is the identity matrix and σ_i^2 is the unit model variance. In the presence of longitudinal data, we are generally interested in the correlation between the observations at different periods; therefore it is important that the imputation method retains the correlation between the observations. Where the changes over time are under study, if the disturbance term is not used, the significance of the changes will be strongly overestimated.

Model (6) can be seen as a particular case of the general regression model, where only the lagged values of the variable y are used as auxiliary variables. Regression imputation may also be applied including other auxiliary variables x correlated with the y under study in model (6) as well.

The mean imputation and the ratio imputation can be seen as special cases of the regression imputation (see the module "Imputation – Model-Based Imputation"): in the mean imputation no auxiliary variables are used; in the case of the ratio imputation the model is based also on another auxiliary variable x.

2.3.6 Donor imputation

The donor imputation methods involve replacing missing values with values from a "similar" responding unit of one or more variables for a non-respondent (called the recipient) with observed values from a respondent (the donor), that is similar to the non-respondent with respect to characteristics observed on both cases. In some versions, the donor is selected randomly from a set of potential donors, which we call the donor pool, as the *random hot deck method.* In other versions a single donor is identified and values are imputed from that case, as the *nearest neighbour method* based on some metric, where there is no randomness involved in the selection of the donor (see the module "Imputation – Donor Imputation").

The missing variable values are replaced by the values of one of the respondents, the possibility to impute several values on the same unit, also in its longitudinal profile, makes these methods particularly suitable for longitudinal data. As a rule, one donor is chosen to ensure consistency within the same record. In nearest neighbour imputation, a distance $d(i,j)$ is defined between two objects i and j, where i is the item non-respondent and j an arbitrary item respondent. A possible measure for the similarity between a non-respondent enterprise and a possible neighbour is based on the correlation of historical data. An advantage of the method is that the results are plausible values, because the donor has been checked in advance and so not too many further controls are needed.

2.3.7 Little and Su method

The Little and Su method can be used for missing values in a quantitative variable y, which can be modelled as a combination of period effect and an individual effect and for which stochastic imputation is desired. It is a nearest neighbour technique, that takes into account both cross-sectional and longitudinal information in defining the nearest neighbours. Imputations can be based on row effects (units) and column effects (periods), where the sum of periods reproduces the whole

observation year. The residual is taken from another unit which, in terms of the row effect, is most similar to the unit that is imputed. The assumption is that units that are similar with respect to the row effect are also similar with respect to residuals. In the ideal case, the donor (of the residuals) has as many attributes equal to the recipient as possible.

This method is reasonably easy to use and can deal with different patterns of missing data, including multiple missing values per single unit. More details on the calculation method are described in the specific method module "Imputation – Little and Su Method").

2.4 Evaluation techniques

An analysis of the imputed data is usually recommended, most of the proposed indicators are based on the comparison between the imputed values and the true values that the non-respondents would have supplied. In the STS context sometimes the non-response is actually a late answer, i.e., it is not in time with respect to the official deadline for the estimates, but it is available immediately after. Hence, such a comparison is possible at least on the set of late responses. On the other hand, a measurement can also be performed on data created randomly according to a simulation scheme, in this way data are not influenced by any characteristics of the late respondents, and the comparison would be done between the simulated data and the ones derived from the imputation method (Little and Su, 1989).

2.5 Quality indicators of the output data

The indicators are usually based on a measure of distance between the two kinds of data. They can be evaluated either at a micro level, or regarding a parameter elaborated at macro level or comparing the eventual difference between the distributions of the two final sets of data.

In general, the usual indicators are based on the following criteria:

a. *Predictive Accuracy*: to assess how the imputed value $\tilde{y}$ (estimate) is close to the reference (true) value y^*:

a.1 the first evaluation criterion, based on the Pearson correlation between $\tilde{y}$ and y^*, this criterion works well for data that are reasonably normal. As r gets closer to 1 the imputation method is judged to be good; if data are highly skewed this measure is not recommended as it could be influenced by the presence of outliers and influential values.

a.2 Another criterion assesses the preservation of the change between waves, by comparing the cross-wave correlations for the imputed and true values. The imputation method is better as the cross-wave correlations from the imputed data are closer to the true cross-wave correlations.

b. *Distributional accuracy*: to measure the distribution accuracy by analysing whether the imputation method preserves distribution of the true values:

b.1 the Kolmogorov-Smirnov distance is calculated between the empirical distribution for both the imputed and the true values. The imputation method is judged to be better as the distance is smaller.

b.2 It is also important to compare the distribution in the dataset that includes the imputed values with the one that includes only true values (this measure includes all cases rather than just those imputed). A measure is based on the change in the variable "decile group membership" from one wave to another. A Chi-Square test is used where the observed cell frequencies are those from the imputed

dataset and the expected cell frequencies are the true cell frequencies. The best imputation method will have the lowest χ^2.

3. Design issues

4. Available software tools

Mean, ratio and regression imputation can be implemented using almost any statistical software. Several R packages are available that can perform imputation, for example, *StatMatch* and *Mice*.

In SAS there are IVEware (Imputation and Variance Estimation) and BANFF. The first uses a multivariate sequential regression approach for multiply imputing item missing values in a data set. The second is a generalised system for statistical editing and imputation developed at Statistics Canada.

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

De Waal, T., Pannekoek, J., and Scholtus, S. (2011), *Handbook of Statistical Data Editing and Imputation*. John Wiley & Sons, New Jersey.

EUROSTAT (2006), *Methodology of Short Term Business Statistics: Interpretation and Guidelines*. Methods and Nomenclatures.

Israëls, A., Kuijvenhoven, L., van der Laan, J., Pannekoek, J., and Schulte Nordholt, E. (2011), *Imputation*. Methods Series Theme, Statistics Netherlands, The Hague.

Kalton, G. (2009), Designs for Surveys over Time. In: *Sample Surveys: Design, Methods and Applications*, Elsevier, Amsterdam, 89–108.

Kennon, R., Copeland, K. R., and Valliant, R. (2007), Imputing for Late Reporting in the U.S. Current Employment Statistics Survey. *Journal of Official Statistics* **23**, 69–90.

Kovar, J. and Whitridge, P. (1995), Imputation of Business Survey Data. In: B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge, and P .S. Kott (eds.), *Business Survey Methods*, John Wiley & Sons, New York, 403–423.

Little, R. J. A. and Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, second edition. John Wiley & Sons, New York.

Little, R. J. A. and Su, H.-L. (1989), Item Non-response in Panel Surveys. In: D. Kasprzyk, G. Duncan, and M. P. Singh (eds.), *Panel Surveys*, John Wiley and Sons, 400–425.

Watson, N. and Starick, R. (2011), Evaluation of Alternative Income Imputation Methods for a Longitudinal Survey. *Journal of Official Statistics* **27**, 693–715.

Interconnections with other modules

8. Related themes described in other modules

1. General Observations – Different Types of Surveys
2. Repeated Surveys – Repeated Surveys
3. Statistical Registers and Frames – The Populations, Frames, and Units of Business Surveys
4. Statistical Data Editing – Editing for Longitudinal Data
5. Imputation – Model-Based Imputation
6. Imputation – Donor Imputation

9. Methods explicitly referred to in this module

1. Imputation – Deductive Imputation
2. Imputation – Little and Su Method

10. Mathematical techniques explicitly referred to in this module

1.

11. GSBPM phases explicitly referred to in this module

1. 5.4 Imputation

12. Tools explicitly referred to in this module

1. R
2. SAS

13. Process steps explicitly referred to in this module

1.

Administrative section

14. Module code

Imputation-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	23-08-2013	review based on the received comments	Roberto Gismondi, Maria Liria Ferraro, Anna Rita Giorgi, Fabiana Rocci	Istat (Italy)
0.3	31-10-2013	review	Roberto Gismondi, Maria Liria Ferraro, Anna Rita Giorgi, Fabiana Rocci	Istat (Italy)
0.4	25-11-2013	review	Roberto Gismondi, Maria Liria Ferraro, Anna Rita Giorgi, Fabiana Rocci	Istat (Italy)
0.4.1	29-11-2013	preliminary release		
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 18:17