This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Little and Su Method

Contents

General section 3

1. Summary 3
2. General description of the method 3
   - 2.1 Introduction 3
   - 2.2 Description of the Little and Su method 4
   - 2.3 Conclusion 5
3. Preparatory phase 5
4. Examples – not tool specific 5
   - 4.1 Example of the Little and Su method 5
5. Examples – tool specific 7
6. Glossary 7
7. References 7

Specific section 8

Interconnections with other modules 9

Administrative section 10

General section

1. Summary

In the following we describe the Little and Su method which is applicable to impute longitudinal data. It takes into account both trend information derived from the data and single units levels. In Section 2 some background of the method are given, while in Section 4 an example of an application of this imputation method is described.

2. General description of the method

2.1 Introduction

Text In the case of repeated measures on a single variable, relatively efficient and simple imputations can often be based on the variable classified by unit and by period (wave). In this context the Little and Su imputation method actually incorporates information about the overall trend of the data and the single unit levels of the unit under study (Little and Su, 1989). It is a nearest neighbour technique, that takes into account both cross-sectional and longitudinal information in defining the nearest neighbours. Furthermore, a residual component is taken from another unit which is most similar to the unit that is imputed in terms of the unit characteristics.

According to this method, the variable is classified by row (meant as unit level) and by column (meant as the period), on which the information about the unit and the trend, respectively, are elaborated.

The two main effects can be combined in different ways. If the missing values are well fitted by a model with additive row and column effects, then imputations may be based on an additive row + column fit:

$$\text{imputation} = (\text{row effect}) + (\text{column effect}) + (\text{residual}) \tag{1}$$

If a multiplicative model, or equivalently an additive model for the logarithm of the variable, seems more appropriate to fit missing values, then imputations may be based on a multiplicative row × column fit:

$$\text{imputation} = (\text{row effect}) \times (\text{column effect}) \times (\text{residual}) \tag{2}$$

The choice of an additive or a multiplicative model depends on the characteristics of missing data, i.e., if data to be imputed have to be not negative, a multiplicative model has to be applied. This is the common case of data coming from business surveys: turnover, number of persons employed, wages and so on. An example can be found in Little and Su (1989).

In the Little and Su method the row and column effects are proportional to row and column means; the column effect describes the mean change over time and is therefore also called the ‘period effect’, while the row effect describes the single unit level corrected for the period effect (Frick et al., 2003).

In particular, the column effect for a certain period is based on the ratio between the period y mean and the average y mean calculated through the whole year: the higher the column effect, the higher the “seasonal” weight of the period concerned will be.

The row effect for a certain unit is given by the *y* mean of all the available longitudinal observations for that unit, where each period observation has been divided by its specific column (period) effect. The row effect is the “longitudinal profile” of the unit concerned.

The residual is taken from another unit which, in terms of the row effect, is the most similar to the unit of which data are going to be imputed. The assumption is that units that are similar with respect to the row effect are also similar with respect to residuals.

2.2 Description of the Little and Su method

As said before, this method (Israëls et al., 2011) can be used for missing values in a quantitative variable *y*, which can be modeled as a period effect combined to a single unit effect and for which imputation is desired. It is reasonably easy to use and can deal with different patterns of missing data, including multiple missing values for each unit. Some problems in applying this method can occur in cases of observed values are all equal to zero for rows with values to be imputed.

In the following an implementation of the model is described.

The column effect c_t gives the mean change of the variable *y* over time and is estimated by:

$$c_t = \frac{\bar{y}_t}{\frac{1}{M}\sum_{t=1}^{M}\bar{y}_t} \tag{3}$$

where $\bar{y}_t$ is the mean of the observed y_{it} at period *t, M* is the number of periods (or waves) for which the average is considered to be significant. The row effect r_i for unit *i* is represented by:

$$r_i = \frac{1}{m_i}\sum_t \frac{y_{it}}{c_t} \tag{4}$$

where the sum is calculated over the m_i available y_{it} for unit *i* over all the periods it is observed.

The residual is derived considering all the units for which the periods, missing for unit *i,* are observed. All these units are sorted according to the row effect value and, among them, the one presenting a row effect closest to that of unit *i,* say unit *j*, is selected.

The residual of unit *j* is represented by:

$$e_{jt} = \frac{y_{jt}}{r_j c_t} \tag{5}$$

In the case of additive model (1), the final estimation is:

$$\tilde{y}_{it} = r_i + c_t + e_{jt} \tag{6}$$

on the other hand, in the case of multiplicative model (2), the final estimation is:

$$\tilde{y}_{it} = r_i c_t e_{jt} \tag{7}$$

It is important to notice that, in this case, a zero row effect will result in a zero imputed value.

In both (6) and (7) the three terms represent the row, column, and residual effects, respectively. In particular the first two terms estimate the predicted mean, and the last term is the component of the imputation from the matched case.

Considering (5), expression (7) can also be written as:

$$\tilde{y}_{it} = r_i c_t \frac{y_{jt}}{r_j c_t} = \frac{r_i}{r_j} y_{jt} \tag{8}$$

From (8) it can be derived that, if the multiplicative model (2) is applied, the final estimation is proportional to the y_{jt} value (y value for the closest unit), adjusted by the ratio between the row effects of the units i and j.

2.3 Conclusion

In general, the method has the following useful features:

a) the imputed values incorporate information about trend from the column effects, and single unit level from the row effects;

b) the method does not require separate modelling for different pattern of missing data, dealing with all patterns simultaneously;

c) the method is comparatively easy to implement and this is an important consideration with large complex data sets.

3. Preparatory phase

4. Examples – not tool specific

4.1 Example of the Little and Su method

A practical example of the use of the Little and Su method in a longitudinal study can be found in this section. Suppose to have the following small sample of fictitious responses to current wages and salaries. In Table 1 there are all cases.

From this example, we see that observation 1 did not respond to the current wages and salaries questions in wave 1, but provided responses in subsequent waves. Observations 5 and 6 also partially responded and wages and salaries information are not provided in two and in one waves, respectively. The first step in the Little and Su method consists in calculating the column effects based on complete cases only, that is, units that were interviewed in 3 waves and responded in all 3 waves for the variables of interest; in the example there are 7 complete cases.

The Little and Su method incorporates trend information into the imputed amounts via the column effects. In this example, the wave 1 column effect of 0.70 indicates that the mean current wages and salaries in wave 1 is 30% lower than the overall mean current wages and salaries, and the means in waves 2 and 3 are 6% and 24% higher than the overall mean, respectively.

Table 1

OBS	Wages & salaries		
	Wave 1	Wave 2	Wave 3
1		400	420
2	675	235	700
3	345	690	800
4	200	480	210
5	200		
6	350	370	
7	400	450	470
8	0	790	790
9	360	450	600
10	135	130	200

In the following, the row effects are calculated: for each unit the row effect is the mean (computed on the number of recorded cases) of the reported values divided by the correspondent column effect. In our example, the row effect for unit 1 is ((400/1.06+420/1.24)/2). The sample is then ordered by increasing row effects (Table 2). In this way, for each observation to be imputed, it is possible to identify the closest donor as the closest complete case.

Table 2

OBS	Wages & salaries			
	Wave 1	Wave 2	Wave 3	
10	135	130	200	**159**
5	200			**286**
4	200	480	210	**303**
1		400	420	**358**
6	350	370		**425**
7	400	450	470	**458**
8	0	790	790	**461**
9	360	450	600	**474**
2	675	235	700	**584**
3	345	690	800	**596**
	0.70	**1.06**	**1.24**	

The following step consists in imputing the missing value by multiplying the actual value for the variable of interest of the donor with the row effect of the recipient divided by the row effect of the donor. That is:

- Obs1 - Wave 1: 200*358/303 = 236.30 ~ 236
- Obs5 - Wave 2: 480*286/303 = 453.07 ~ 453
- Obs5 - Wave 3: 210*286/303 = 198.22 ~ 198
- Obs6 - Wave 3: 470*425/458 = 436.14 ~ 436

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

Frick, J. R. and Grabka, M. M. (2003), *Missing Income Data in the German SOEP: Incidence, Imputation and its Impact on the Income Distribution*. DIW Berlin.

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

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.

Specific section

8. Purpose of the method

Check erroneous values in microdata on logical grounds.

9. Recommended use of the method

1.

10. Possible disadvantages of the method

1.

11. Variants of the method

1.

12. Input data

1.

13. Logical preconditions

1. Missing values
    1.
2. Erroneous values
    1. Not allowed. All observed values have to be correct.
3. Other quality related preconditions
    1.
4. Other types of preconditions
    1.

14. Tuning parameters

1.

15. Recommended use of the individual variants of the method

1.

16. Output data

1.

17. Properties of the output data

1.

18. Unit of input data suitable for the method

19. User interaction - not tool specific

1.

20. Logging indicators

1.

21. Quality indicators of the output data

1.

22. Actual use of the method

1.

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Imputation – Imputation for Longitudinal Data

24. Related methods described in other modules

1.

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

1.

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

1. GSBPM Sub-process 5.3: Review, validate and edit

27. Tools that implement the method described in this module

1.

28. Process step performed by the method

Administrative section

29. Module code

Imputation-M-Little and Su

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	05-02-2013	first version	Roberto Gismondi Fabiana Rocci Anna Rita Giorgi Maria Liria Ferraro	Istat (Italy)
0.2	20-08-2013	second version	Roberto Gismondi Fabiana Rocci Anna Rita Giorgi Maria Liria Ferraro	Istat (Italy)
0.3	30-10-2013	review	Roberto Gismondi Fabiana Rocci Anna Rita Giorgi Maria Liria Ferraro	Istat (Italy)
0.3.1	19-11-2013	preliminary release		
0.3.2	29-11-2013	minor corrections		
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	21-3-2014 18:17