7 th WORKSHOP ON LABOUR FORCE SURVEY METHODOLOGY

DATA PROCESSING AND DATA QUALITY

Madrid, Spain, 10 – 11 May 2012

C. Data processing: Methods for editing and imputation, weighting, non-response adjustment

C1– Method for the imputation of the earning variable in the Belgian LFS

Astrid Depickere – Belgium

Presentation to the 7th workshop on Labour Force Survey methodology
Madrid, 10-11 May 2012

A Method for the imputation of the earnings variable in the Belgian LFS

Astrid Depickere, Statistics Belgium
Anja Termote, Statistics Belgium
Pieter Vermeulen, Statistics Belgium

1. Introduction: the wage variable in LFS

In the 1999 questionnaire of the Labour Force Survey, a new question was introduced aimed at measuring the net monthly wage of the respondent. The sensitive nature of the question and its optional character resulted in a very high number of missing values. This seriously limited the use of this variable.

Over the years, some actions were taken in order to reduce the number of item non response on this question, like changing the interviewer instructions and asking interviewers to persuade the respondent to respond to the question. This resulted in a reduction of the share of item non response from around 50% in 1999 to 25 % in 2011.

Additionally, from 2009 on, we started imputing the missing values of the earnings variable. This imputation method is the specific subject of this paper.

2. Imputation method

As imputation method we have chosen for a regression imputation, where the regression parameters are obtained from a regression model performed on data from an external source, being the Structure of Earnings Survey (SES).

The European Structure of Earnings Survey (SES) provides detailed information on the level and structure of remuneration of employees, their individual characteristics and the enterprise or local unit to which they belong. It is a 4-yearly survey conducted under Council Regulation 530/1999 and Commission Regulation 1916/2000. Whereas the survey is 4-yearly under the Regulation, Statistics Belgium decided to collect the data on an annual basis, except for NaceRev 1 sections M, N and O (from 2009 on NaceRev 2 sections P-S).

An alternative option to the SES would have been to use the non missing cases in the LFS itself to obtain the regression equation. However, given the high number of missing values and the less accurate measurement of the wage variable (which is not the core variable of the LFS), we chose to use the SES as the source for our regression model.

Whereas the use of SES has the advantage that it provides us with a better regression equation, it has a number of drawbacks, for which a specific solution had to be sought.

The first is the gap of almost two years between the reference periods of these surveys. At the time of delivery of the LFS dataset to Eurostat, the most recent SES data are two years old. We

resolved this problem by applying an indexation on the basis of the quarterly Labour Cost Index.

A second issue we had to resolve was the fact that not every reference year the entire market is covered by the SES. More specifically, the sections M, N and O of NaceRev 1 (or, from 2009 on sections P-S of NaceRev 2) are only included once every four years (being the reference years of the Eurostat regulation). We resolved this by deriving the regression parameters for the missing years on the basis of the non missing years.

Finally, a third difficulty is that SES only measures gross wages, whereas LFS measures net wages. The solution to this problem was to apply the administrative rules to go from gross wages to net wages, taking into account as much as possible the information from LFS (e.g. the number of children, the partner's employment situation). We are however limited to the information available in LFS.

3. Steps

3.1. Obtain regression equation from SES

The SAS Proc GLM procedure was used to obtain the regression equation. Different models were compared and the model that was retained was the following:

logGMW = sex age age^2 isco_3d nace_2d isced_6cl region size pt_pct

with:

- *logGMW*: the logarithm of the Gross Montly Wage is used as dependent variable. Using the logarithm instead of the actual Wage results in a better regression model.
- *sex*: gender of the respondent
- *age* and *age*^{2}: apart from the age of the respondent, the squared age is also used as a predictor. Using both allows modelling the typical shape of the association between wage and age.
- *isco_3d*: respondents job according to ISCO, 3 digits
- *nace_2d*: economic activity of the company, according to NACE, 2 digits
- *isced_6cl*: respondents highest level of education (6 ISCED classes)
- *region*: Region of employment (3 classes)
- *size*: sizeclass of the local unit of the company (6 classes)
- *pt_pct*: the percentage of part-time work. For full time workers this value is 100%.

The regression model has a R-squared of 75%, which means that 75% of the variation in the wage is accounted for (predicted) by the independent variables.

3.2. Impute Wage variable in LFS

The regression equation obtained from step 1 was then applied on the LFS data in order to compute an imputed wage variable for each respondent of the LFS. The result is an estimated Gross Monthly Wage value. Next, the indexation coefficient (by 1 digit NACE) obtained from the Labour Cost Index was applied to this value, in order to correct for the two year time gap between the LFS and SES data.

3.3. Prepare LFS dataset to go from Gross to Net Wage

According to Belgian legislation, the level of a person's Net Wage is determined by several aspects like the number of persons in charge, the partnership and the employment position (and wage) of the partner.

Because the LFS data are collected on a household level, we could determine a number of variables on the household level, like the number of children in the household, whether a respondent has a partner or not and what is the employment position of the partner. The information is somehow imperfect, because we are limited to the information about the relationship of a respondent to the reference person in the survey and we lack information about the relationships among the members that are not reference person.

3.4. Determine Net Wage

By applying the Gross/Net calculation algorithm, we obtain a value for the Net Monthly Wage for each of the respondents in the LFS dataset. The calculation of an imputed value for each respondent (i.e. not only those who have a missing value on the wage variable) allows us moreover to compare the result of the imputation with what we have observed among the non missing cases. Therefore, the method is not only used for imputation, but also in the data editing process (outlier detection).

4. Evaluation

If we compare the estimated values of the Wage variable before and after imputation, we see that the results are quite similar. The estimated mean value before imputation differs only slightly from the estimated mean value after imputation. A comparison of the main percentiles gives again a very similar picture, as shown in the table below, especially for the values closes to the median. Deviations between the two estimates become bigger on the extreme sides of the distribution (Pctl 99 and Pct 1), reflecting the smaller variation in the data after imputation.

Table 1: Mean values and Percentiles of Wage variable: comparison of values before and after imputation.

Analysis Variable : Q91										
	Mean	99th Pctl	95th Pctl	90th Pctl	75th Pctl	50th Pctl	25th Pctl	10th Pctl	5th Ptcl	1st Pctl
After imputation	1630.32	4000	2783	2330	1900	1530	1256	916	749	410
Before imputation	1641.46	4200	2800	2300	1900	1500	1250	980	780	350

This artificial reduction of variance and sampling errors is a typical effect of many imputation methods. More sophisticated methods, like multiple imputation methods, try to cope with this. However, given that the wage variable is not among the core values of the LFS, we can conclude that the described method performs quite satisfactory and serves the purposes of most uses of the survey data.

Table 2: Variance and standard errors of Wage variable: comparison of values before and after imputation.

Analysis Variable : Q91		
	Variance	Std Error
After imputation	513142.60	7.4030644
Before imputation	593941.47	9.8034195