

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

*Workshop on LFS methodology,
Madrid 2012, May 10-11*

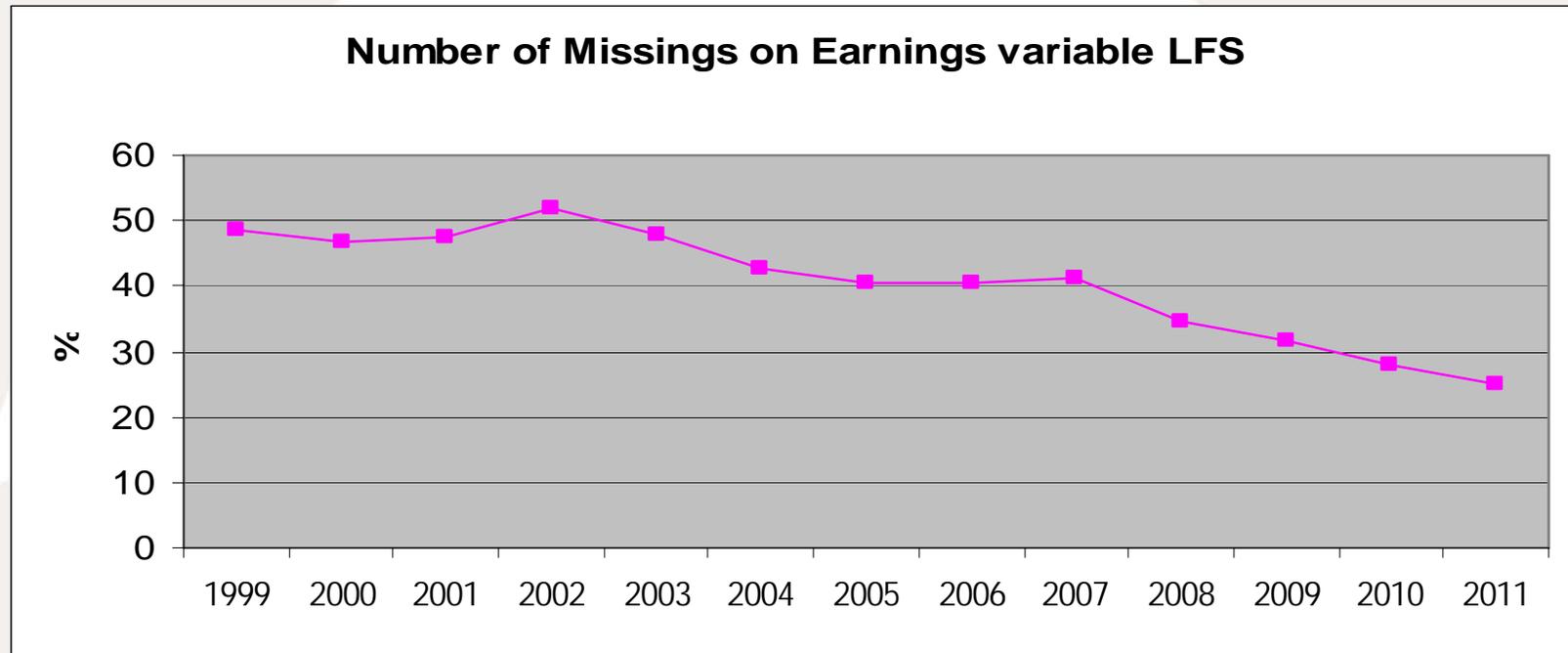
Astrid Depickere, Anja Termote, Pieter Vermeulen

Outline

1. Introduction
2. Imputation
3. Imputation method for Earnings variable in LFS
4. Implementation: different steps
5. General evaluation

Introduction

- The Earnings Variable in the Labour Force Survey (LFS) : very high number of missing values. (24,9% in 2011)



- In 2009:
 - Some actions were undertaken to reduce the number of missings
 - Start imputation of the earnings variable

Imputation

- **Imputation** = replacing missing values with ‘**credible**’ data from a **donor**.
 - What is ‘**credible**’ data? Using what we know in order to say something about we do not know
 - **Donor?**
 - Same source: borrowing information from the nonmissing observations to impute for the missing observations
 - External source: using information from another source to impute for the missings

- **Imputation techniques:**
 - Single imputation: *generate a single replacement value for each missing data point.*
 - Multiple Imputation: *creates several copies of the data set and imputes each copy with different plausible estimates of the missing values.*

Imputation method for Earnings variable in LFS (1)

- Regression imputation using an external source: the Structure of Earnings Survey (SES):
 - Regression imputation (or conditional mean imputation) replaces missing values with predicted scores from a regression equation.
 - We use the information about the effects of different personal and job characteristics on the wage level from the SES,
 - in order to predict a wage level for the missing observations in the LFS.
- Why SES (instead of LFS)?
 - A better measurement of wage variables in SES than in LFS. Earnings are the core variables in SES, whereas they are not in LFS.
 - High number of missings in LFS: insufficient representativity of the regression model

Imputation method for Earnings variable in LFS (2)

- Some particular issues that needed to be resolved:
 - Two year gap between delivery of SES data and LFS data
 - ⇒ Indexation on the basis of the Labour Cost Index
 - SES is a yearly survey but does not always cover the entire market. Some sectors are included only once every four years (ESTAT year).
 - ⇒ Coefficients for the missing years are derived on the basis of the last nonmissing year
 - SES only measures gross wages, whereas for LFS nett wages are needed.
 - ⇒ Applying a gross/nett calculation (taking into account as much as possible the information in LFS on individual an his household)

Implementation: different steps (1)

Step 1: Obtain regression equation from SES

- SAS proc GLM
- Different models were compared
- Final model has a R-squared of 75%
- Only main effects, no interactions
- Regression parameters were converted into a formula for the prediction of a Gross Monthly Wage

logGMW = sex age age2 isco_3d pct_pt nace_2d isced_6cl region size

Dependent variable =
variable to be predicted

Independent variables = predictors

Implementation: different steps (2)

Step 2: Impute Wage variable in LFS

- Regression equation is applied
- Result = Gross Monthly Wage value for the missing observations in the LFS survey
- Apply indexation (by NACE_1d) obtained from the Labour Cost Index

Step 3: Prepare LFS dataset for Gross/Nett calculation

- Update calculation according to legislative rules: Nett wage is a function of the Gross wage, number of persons in charge, partnership & employment position (and wage) of the partner
- Derive household variables

Implementation: different steps (3)

Step 4: determine Nett Wage

- By applying the gross/nett calculation, a Nett Monthly Wage value is obtained (for all observations)
- Validation of the result: compare imputed values to observed values (for the nonmissing observations)
- The method not only serves as an imputation method, but can also be used for data editing (e.g. evaluation of outliers)

General evaluation

- Effect of imputation on **estimates** (descriptive values): bias remains very small => strong coherence between the sources
- Imputed (but biased) data better quality than original ones?

Analysis Variable : Q91										
	Mean	99th Pctl	95th Pctl	90th Pctl	75th Pctl	50th Pctl	25th Pctl	10th Pctl	5th Pctl	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

General evaluation (2)

- Effect of imputation on **variance and sampling error**: artificial reduction of variance, true variance is underestimated

<u>Analysis Variable : Q91</u>		
	<u>Variance</u>	<u>Std Error</u>
<u>After imputation</u>	513142.60	7.4030644
<u>Before imputation</u>	593941.47	9.8034195

- Solution could lie in the use of a different technique:
 - Stochastic regression imputation
 - Multiple imputation