

Non-response Recovery by Imputation using Temporal Extrapolation of (administrative) Profit & loss account data in the Structural Business Survey

Statistics Belgium, Methodology Dept., guy.vekeman@statbel.mineco.fgov.be

Introduction

- Stratified simple random sampling
- Companies employing ≥ 50 persons: open-end take-all stratum.


Information from administrative records: Profit & loss account data from balance sheet deposited yearly at the Belgian National Bank.

Accounting totals are the key values that are inserted into the survey questionnaire.

Any breakdown of those totals needs to be estimated. Two imputation systems considered, consist in:

- Using ratio estimation on data of that same respondent from the previous survey
- Using ratio estimation on summed up data of similar respondent companies (same NACE class if possible and same size band)

Obtaining comparable data

- Variable definitions need to be uniform
 - No accounting standardization
 - Questionnaire items need to be added or subtracted : e.g. for turnover:
 - $EU12110 = (\text{Turnover from normal operations})$
 - (Subsidies booked in turnover)
 - + (Miscellaneous recurrent income)
 - (Compensations and subsidies)
 - (Profits realized on sales of material assets, compared to accounting value)
 - (Indemnities received)
- 
- Administrative totals

How good are both administrative result sheet totals as a proxy for the EU12110 ?

- Turnover from normal operations: $\rho = 0.999$,
- Miscellaneous recurrent income: $\rho = 0.332$ (both calculated on 25.552 companies).

Using additional individual data

Focus on **subsidies received** :

- Poor correlation with turnover from normal operations (0,084) for total set of enterprises
- Correlation coefficient of 0,966 for selection of 389 companies receiving any subsidies
- Determining factors ? Nace, size band ? No reliable 'predictive' factors.

Constraints for various imputation schemes

- Appropriate imputation schemes need to supply data that respect consistency controls
- Introducing accounting totals: correct values for the accounting totals (key values)
- Ratio estimation for the breakdown of those accounting totals.
- Ratios drawn from:
 - Questionnaire items from a prior survey. (Temporal extrapolation)
 - Aggregates of questionnaire items from 'similar' responding companies: drawn in the same or a closely related stratum.

Using temporal extrapolation on the breakdown of accounting totals

- Advantages:
 - The simplest way to exploit all knowledge about the individual company is to use a prior questionnaire that was completed correctly and combine it with the accounting totals known from the yearly account.
 - Non-accounting information can be copied from the prior questionnaire.
- Implicit assumption: temporal correlation (between successive years) of the breakdown of accounting totals of a given company is on average better than the correlation of the breakdown calculated from a set of companies from the same or a neighbouring stratum in the same survey.

Using appropriate aggregated data

Stratification is a clear advantage for imputation:

Data of responding companies from the same stratum can be used to calculate ratios to be used for the breakdown of the totals of the result account of the non-responding company. Mean ratios are calculated from aggregated data.

Advantages:

- Straightforward method, easy to implement.
- Aggregated totals are guaranteed to be non-zero (accounting totals: always same sign).

Evaluation of the two procedures

- Evaluate the quality of the imputation procedure on any example by comparing the imputed values to true values.
- Measuring in a proper way the ‘distance’ between the true values and both sets of imputed values. Heuristic formula: Weighted Euclidian distance.

$$d_{e,t} = \sqrt{\sum_{j=1}^n w_j^2 \cdot (e_j - t_j)^2}$$

$$D_{e,t} = \frac{2 \cdot \sqrt{\sum_{j=1}^n w_j^2 \cdot (e_j - t_j)^2}}{\sum_{j=1}^n w_j \cdot t_j}$$

Both procedures were used on the *datasets of a major food processing company*.

The temporal extrapolation of the breakdown of accounting totals yielded

following distances: $d_{e,t} = 36,0 \cdot 10^6$ $D_{e,t} = 35,7 \cdot 10^{-6}$

The use of aggregated data from the NACE-3 x size band stratum gave following

distances: $d_{e,t} = 274 \cdot 10^6$ $D_{e,t} = 272 \cdot 10^{-6}$

The second estimate is still acceptably well, but its distance from the true values exceeds the former one by a factor of 7,5.

Implementation

Both imputation procedures are actually applied at Statistics Belgium.

- The temporal extrapolation of the breakdowns of accounting totals is used for data recovery with the larger non-respondents.

Implementation in visual-basic module in MS-Excel.

The procedure can be completed by a qualified operator. This takes half the time it requires to check a real questionnaire for consistency and input the data.

- Ratios calculated from aggregated data are used for data recovery mainly for SME's above 10 employees in industry and above 20 in other sectors, where sample sizes per stratum tend to be somewhat larger.

Procedure needs to be performed by a statistician on a PC. Due to the use of ODBC to gain access to the database on a mainframe computer it is intrinsically slow. However, after initialisation, it requires little input and can be operated as a 'background task' or on a dedicated PC.

- No imputation procedure is used to compensate for non-response of small SME's. Reweighting the respondents is thought not to affect the true totals in this case. (Assumption of 'missing completely at random')

Final Remark: The response rate is declining as a result of budget cuts in the inspection department, which takes over data collection if a second reminder is not fruitful. Imputation procedures are therefore used more frequently than what they were meant for initially. Due to a cascaded use of imputation procedures, imputed values may gradually diverge from true data.