Robust Regression, MissForest and Calibration combined with Non-Linear Optimization with Constraints to impute VAT Turnover

Jacques Saliba

Federal Statistical Office FSO/ Data Science, AI and Statistical Methods/ Statistical Methods

UNECE Conference of European Statisticians
Expert meeting on Statistical Data Editing | October 5, 2022

Outline

Introduction

- Goal: allocate a yearly turnover to $\sim 700'000$ business units in CH
- For $\sim 55\%$ business units, the turnover is known from paid value-added tax (VAT) representing $\sim 62\%$ of total turnover
- For the remaining business units:
  - Imputation of missing turnovers ($\sim 43.2\%$ business units representing $\sim 3\%$ of total turnover)
  - VAT group members: adjustment of turnovers based on the above mentioned imputation ($\sim 1.1\%$ business units representing $\sim 35\%$ of total turnover)
- One model version out of 8 detailed in the paper will be presented in the following.

Auxiliary variables

- Number of employees
- Number of full-time equivalents
- Classification of economic activities (NOGA $\sim$ NACE) for business units
- Customs data (import, export) in CHF
- Total wages based on the old-age and survivor's insurance $t-1$, has few missing values

Imputation of totally missing turnovers

1. First imputation step: Linear robust regression
   - Consider "model" business units with more than 20 employees + known turnover
   - Based on NOGA, build imputation classes containing at least 30 model business units
   - For an imputation class I, the turnover y_i is modelled as a linear combination of x_i (number of employees) and s_i (total wages):

$$y_i = \alpha_I + \beta_I x_i + \gamma_I s_i + \epsilon_i.$$

   The MM robust method was used to reduce the effect of outliers on parameter estimation.

Imputation of totally missing turnovers

2. Second imputation step: MissForest algorithm (Stekhoven and Bühlmann [2012])
   - Imputation of the turnover of business units with ≤ 20 employees, using the auxiliary variables (beside the ones previously mentioned):
     - Number of employees size classes
     - Quantiles and average of total wages, in each NOGA2

Distribution of turnovers within VAT groups

- In VAT groups, the VAT is paid by the group head unit for all the group members.
- For a VAT group G, we denote $z^{(1)}$ the known total turnover.
- Imputed turnovers of its members are denoted by $y_1, \cdots, y_k$ and we have

$$\sum_{j=1}^{k} y_j = z^{(2)}.$$

- A basic way to get the desired total turnover $z^{(1)}$: Multiply all y_j by $r := \frac{z^{(1)}}{z^{(2)}}$.

Calibration

A calibration method with a linear truncated distance (Deville and Särndal [1992]) is used:

- For a VAT group G, assign initial weights $= 1$ to each member's turnover.
- Use Lagrange multiplier to find weights g_i's as close as possible to 1 such that

$$\sum_{i \in G} g_i y_i = z^{(1)} \text{ and } \sum_{i \in G} D(g_i, 1) \text{ is minimal}$$

with the pseudo-distance $D(.,.)$ with fixed bounds L and H given by

$$D(a,b) = \begin{cases} \frac{(a-b)^2}{2b} & \text{if } Lb < a < Hb. \\ \infty & \text{otherwise}, \end{cases}$$

Calibration

- Initial weights $= 1 \Longrightarrow$ final weights $g_i \in [L, H]$
- Choose L and H such that $\{1, r\} \in [L, H]$
- After calibration, the distributed turnover of a business unit i becomes $y_i^c := g_i \times y_i$.

Linear optimization with additional constraints

- Goal: Try to adjust distributed turnovers in order to satisfy productivity bounds.
- Compute quantiles p_5 and p_{95} of productivity (turnover/#employees) in each NOGA2 crossed with number of employees size classes
- Using NlcOptim in R, try to find weights g_i' as close as possible to g_i such that

$$\sum_{i \in G} g_i' y_i = z^{(1)} \text{ and } p_5 \leq \frac{g_i' \times y_i}{x_i} \leq p_{95}.$$

- Reiterate NlcOptim with productivity percentile pairs $\{4, 96\}, \{3, 97\}, \{2, 98\}$ and $\{1, 99\}$.
- If no solution is found for a VAT group, keep $y_i^c = g_i \times y_i$ as distributed turnover.

Results

- We compare the results of the imputed and distributed turnovers with their corresponding turnover from the survey of the production and value added statistics (WS) for 2019.
- The WS turnover is defined slightly differently from VAT turnover. The R^2 of robust regression between non-imputed VAT turnover and their corresponding WS turnover is ~ 0.7.

- Denote by Old-imp the basic imputation model: robust linear regression in NOGA2 with only employees as auxiliary variable and distributed turnovers using ratio $r := \frac{z^{(1)}}{z^{(2)}}$.
- Denote RF_B20 the application of the robust regression, the MissForest, the calibration and optimization as outlined previously.

Table: R^2 between original/imputed/distributed VAT turnovers and WS turnovers (2019)

	distributed	imputed	original
Old-imp	0.283	0.255	0.702
RF_B20	0.385	0.337	0.702

Conclusions

- The quality of the imputation model is enhanced by using MissForest to impute turnovers of small business units
- The distribution model of turnovers within VAT groups is enhanced by using a calibration method
- More realistic imputation values result from adjusting the calibration weights to productivity bounds

Potential improvements:

- Use of past years VAT and WS data to improve the distributed turnovers
- Adding more explanatory variables to the robust regression and to the MissForest
- Sharpening the selection of units and tuning the parameters of the MissForest

References

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Jean-Claude Deville and Carl-Erik Särndal. Calibration estimators in survey sampling. *Journal of the American statistical Association*, 87(418):376–382, 1992.

Daniel J. Stekhoven and Peter Bühlmann. MissForest - non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.