Utrecht University

Methodology and Statistics for the Biomedical, Behavioural and Social Sciences

Master's Thesis

Improvements of classical ratio-imputation methods using robust statistics and machine learning techniques

Author:
Agnese Giacomello
Student Number: 6078249

Supervisor:
Dr. Jeroen Pannekoek

December 14, 2018

1 Abstract

When dealing with survey data, individual values are in practice frequently missing. To allow for a more precise statistical analysis, missing data need to be taken into account properly. The classical approach for business data in Official Statistics is to replace missing values based on a ratio imputation model, which involves a single predictor variable. The methods in this thesis aim at developing a more accurate ratio imputation procedure while maintaining a simple model structure. Extensions of the classical method involve the use of the robust Huber and Tukey estimators and the application of the machine-learning technique *boosting* to the case of interest, which allows for the inclusion of multiple predictors. With a focus on business data in Official Statistics, the accuracy performance of the robust and *boosting* techniques will be compared to the classical approach.

2 Introduction

Missing data are an ongoing problem in real data research. The presence of missing values can be due to a variety of factors, related both to the way in which data are collected and to the information contained in the data itself. As an example, missing observations are more likely to occur on private or sensitive topics as in the case of income-related surveys (de Waal, Pannekoek, & Scholtus, 2011). To ensure a correct research, missing values need to be properly handled in the analysis process. Statistical methods can be used to infer the loss of information, so to limit the damage that the missingness produces on the research output.

In the past thirty years a thorough research was developed with the goal of finding methods that lead to unbiased estimates of the missing information. Little and Rubin (1986) had a major impact in the developments of the currently used techniques. They provided the basis of the actual missing data processing, which can be non-extensively summarised in two categories: case deletion and imputation methods (Qin, Zhang, Zhu, Zhang, & Zhang, 2007). The former, which consists of the deletion of cases in which some information is missing, can often result in estimation bias, in particular with small sample sizes (Kang, 2013). Imputation, which dates back at least to 1966 (Palmer & Jones, 1966), aims at inferring the most likely value for the variable of interest by using the amount of information contained in the non-missing values. In the presence of *item nonresponse*, where only some observations of a single respondent are missing, the common approach is to replace missing values by imputation procedures (Kalton & Kasprzyk, 1986; Allison, 2002).

Imputation methods evolved in a number of different approaches, one of which is the focus of this thesis: ratio imputation (Shao, 2000). In particular, for business data in Official Statistics ratio imputation is the most commonly used method to deal with *item nonresponse* (Takahashi, 2017). Reasons for that lie within the kind of data involved, which are usually non-negative (e.g. costs, turnover, etc.) and proportional to a measure of size of the business itself (e.g. number of employees). The ratio model here presents a number of advantages: (i) it does not involve an intercept, therefore protecting against the prediction of negative values that would not fit business data; (ii) it comprises the variance proportional to the mean characteristic usually found in business variables (i.e. larger firms have larger errors); (iii) its simple model structure allows for an easy and direct interpretation of results. As will be further explained in section 3, the classical ratio imputation estimation method is based on the ratio of the

means of a single predictor and an output variable. With a focus on business data in Official Statistics, the aim of this thesis is to find a ratio imputation procedure which leads to more accurate estimates while maintaining its simplicity. In this regard, the Huber (Huber, 1964) and Tukey (Tukey, 1977) robust estimators will be considered and compared to the classical method. These methods involve the use of a single predictor in the imputation process, while offering protection against outliers.

A further extension that will be investigated is the application of the machine-learning technique *boosting* to the case of interest, which allows for the inclusion of multiple predictors while maintaining a simple and interpretable model structure.

Performance of these methods will be compared and estimation accuracy will be assessed based on two measures: the Mean Absolute Percentage Error (MAPE) and the Root Mean Squared Error (RMSE).

3 Theory

3.1 Ratio Imputation

In ratio imputation, a single predictor variable x is used in the prediction of a target variable y, where the target variable represents the outcome of interest. In the ratio model, the ratio between target and auxiliary variable is assumed to be approximately constant and it is caught by the quantity R. In the classical ratio model, R is estimated as the ratio between the mean of y and x computed on the units for which both auxiliary and target values are available, i.e. $\hat{R} = \bar{y}/\bar{x}$. Given a dataset with N units, the missing y-values can therefore be imputed following the ratio model as in Equation 1:

$$\tilde{y}_i = \hat{R}x_i, \tag{1}$$

for $i = (1, ..., N)$.

To explain how the above-presented $\hat{R}$ estimator is derived, the notion of *loss function* needs to be introduced. The loss-function L_w is the weighted sum of squares of the observed residuals and as such, the objective is its minimisation. How residuals are weighted defines the different types of estimators. In the classical weighted least squares (wls) case, the loss function weights the sum of squares of the residuals proportionally to the reciprocal observed variance, leading to the function in Equation 2:

$$L_w = \frac{1}{2}\sum_{i=1}^{N}\left(\frac{e_i}{\sqrt{\sigma^2 x_i}}\right)^2, \tag{2}$$

where $e_i = y_i - \hat{R}x_i$ are the model residuals. The minimisation of this quantity, achieved by setting its derivative to zero with respect to $\hat{R}$, leads to the wls estimator $\hat{R} = \bar{y}/\bar{x}$.

Although this ratio estimator is simple and intuitive, it involves only a single predictor and it does not prevent the influence of outliers in the estimation process. This, together with the waste of information when multiple target variables are available, may result in biased estimates.

Robust Ratio Imputation

The classical *wls* method works well for normally-distributed residuals. However, in the presence of outliers this condition is often not met and can result in an erroneous influence on the ratio estimate, leading to a lower imputation accuracy. Robust statistics allow to reduce outliers' influence by down-weighting their impact in the estimation process. In this context, outliers are considered observations with high residuals from the fitted ratio model in Equation 1 (Kelly, 1992).

Two robust estimators have been investigated: Huber (Huber, 1964) and Tukey (Tukey, 1977). These estimators achieve their robustness by modifying the way residuals are weighted, and consequently their loss function. To get an insight on how these estimators reduce the influence of residuals compared to *wls* estimation, the function ρ of the residuals can be inspected in Table 1. The sum of the ρ function over all the observations represents the loss function, i.e. $L_m = \sum_{i=1}^{N} \rho(e_i)$.

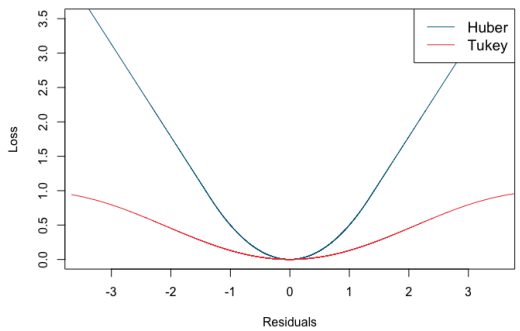
TABLE 1: Function ρ of the residuals for the *Wls*, Huber and Tukey estimators

	$\rho(e_i)$
Wls	$\rho(e_i) = \frac{1}{2}(\frac{e_i}{\sqrt{\sigma^2 x_i}})^2$
Huber	$\rho_H(e_i) = \begin{cases} e_i^2/2 & \lvert e_i\rvert \leq k \\ k\lvert e_i\rvert - \frac{1}{2}k^2 & \lvert e_i\rvert > k \end{cases}$
Tukey	$\rho_T(e_i) = \begin{cases} k^2\{1-[1-(\frac{e_i}{k})^2]^3\}/6 & \lvert e_i\rvert \leq k \\ k^2/6 & \lvert e_i\rvert > k \end{cases}$

Where $L_m = \sum_{i=1}^{N} \rho(e_i)$, k is a predefined constant and $e_i = y_i - \hat{R}x_i$ are the model residuals.

The choice of the tuning constant k regulates the amount of robustness of these estimators: smaller values of k lead to more resistance to outliers, and vice-versa (Wang & Bai, 2007). A suggested choice is to pick $k = 1.345\sigma$ for the Huber and $k = 4.685\sigma$ for the Tukey one (Kelly, 1992), where σ is the standard deviation of the residuals. This choice has been shown to be optimal for symmetrical distributions. In Figure 1, a visualisation of both the Huber and Tukey loss function can be found, which is based on the k values suggested by Kelly (1992).

FIGURE 1: Loss function L_m of the Huber and Tukey estimators



3.2 Boosting (introduction)

Boosting is a step-wise Machine Learning algorithm that, given a target variable and a set of predictors, allows to add predictors sequentially one at a time, so to optimise the model performance.

Given a set of p predictors $(x_1, ..., x_p)$ and a target variable y, boosting achieves the optimal prediction of y given x by minimization of the loss function $\eta(y, f(x))$.

This translates in the estimation of the function:

$$f*() = argmin_f E(\eta(y, f(x))), \tag{3}$$

where $\eta(y, f(x))$ is assumed to be differentiable with respect to $f(x)$.

Applied to the ratio model, the optimal prediction of y given x is in practice achieved through the following steps:

1. consider $\hat{R}_j$ the ratio estimate computed through one of the estimation methods presented in subsection 3.1 using target variable y and predictor variable x_j (with $j = (1, ..., p)$);
2. find the predictor x_j (with $j = (1, ..., p)$) which minimizes the prediction error of the ratio model in Equation 1, leading to the model:

$$\hat{y}_1 = \hat{R}_j x_j, \quad \text{with residuals} \quad r_1 = y - \hat{y}_1 \tag{4}$$

3. fit all the remaining predictors to the model residuals r_1 in Equation 4 and pick the one (x_k, with $k = (1, ..., p-1)$) which minimizes its error, such that $\hat{r}_1 = \hat{R}_k x_k$. The estimate of the target variable is now updated by the inclusion of this predictor as:

$$\hat{y}_2 = \hat{y}_1 + \hat{r}_1 = \hat{R}_j x_j + \hat{R}_k x_k, \quad \text{with residuals} \quad r_2 = y - \hat{y}_2 \tag{5}$$

4. repeat step 3 until the optimal m_{stop} iteration is achieved (where the optimal m_{stop} value ranges between 1 and p). The package *mboost* (Hofner, Mayr, Robinzonov, & Schmid, 2014) of the software R (R Core Team, 2013) can be used to derive this optimal value.

Following the algorithm above, it is possible to include information from more than one predictor while keeping a relatively simple model structure. The high accuracy that boosting achieves (Schapire, 1999) and the possibility to include multiple predictors makes it an appealing method to improve ratio imputation's accuracy.

4 Methods

4.1 Data description

With the goal to estimate totals and means with the highest accuracy, analyses are based on a dataset from the *Dutch Structural Business*, where 274 units are recorded on six variables. These include two predictor variables (*Turnover* and *Number of employees*) and four target variables (*Cost of purchases*, *Cost of depreciation*, *Cost of employees* and *Other costs*). As it is often the case with business data, these variables are all right skewed and strictly positive.

4.2 Estimation methods

A ratio model was fitted to the data for every combination of target-predictor variables. In each model, the ratio R was estimated based on the wls, Huber and Tukey estimators.

To assess estimators' accuracy on the out of sample predictions, missing values were randomly introduced on each target variable individually with a 10% proportion. Considering the random nature of this procedure, it was repeated for 500 iterations. At each iteration, the ratio R was estimated using the 90% non-missing values for each target-predictor variables combination and each estimator. The ratio estimates were therefore used in the prediction of the missing y-values following Equation 1.

The predicted values were compared to the (known) true y-values to assess estimators' performance. At each iteration, accuracy was assessed based on the following measures:

- The **Mean Absolute Percentage Error** (MAPE), which provides a standardized accuracy measure in terms of percentages. Lower values correspond to better predictions. It is defined as:

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{|y_i - \hat{y}_i|}{|y_i|}\right) * 100, \tag{6}$$

where:

- y_i is the actual observation;
- $\hat{y}_i$ is the unit prediction (obtained using Equation 1);
- N is the sample size.

- The **Root Mean Squared Error** (RMSE), which corresponds to the standard deviation of the residuals, is defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}. \tag{7}$$

It measures the average magnitude of the error, and the squared term causes high errors to have a stronger influence on the final error estimate.

The final error for each estimator and variables' combination was obtained by averaging the accuracy measures over the 500 iterations.

5 Results

5.1 Ratio estimates

At first, the *wls*, Huber and Tukey estimators were used to derive the different R estimates on the complete data. Results have been summarized in Table 2 for a few combination of target and predictor variables.

TABLE 2: Comparison of the ratio R based on different estimators

	R estimator		
Target - Predictor variables	*wls*	Huber	Tukey
Cost of Employees - Number of Employees	37.86	37.65	37.29
Cost of Purchases - Turnover	0.87	0.83	0.82
Cost of Depreciations - Turnover	0.0070	0.0072	0.0076
Other Costs - Number of Employees	33.01	29.01	27.31

5.2 Classical and Robust statistics: accuracy comparison

As explained in subsection 4.2, after the introduction of the nonresponses the MAPE and RMSE were used to compare the accuracy of the different estimators.

Table 3 includes the errors' results for the same variables reported in Table 2, where lower values indicate a lower prediction error. These results were obtained after averaging over the 500 iterations.

TABLE 3: Imputation error assessment, based on MAPE and RMSE

	MAPE			RMSE		
Target - Predictor variables	*wls*	Huber	Tukey	*wls*	Huber	Tukey
Cost of Employees-Number of Employees	1.34	1.30	1.26	34.97	33.84	32.84
Cost of Purchases - Turnover	1.17	1.08	1.07	519.62	476.97	471.25
Cost of Depreciations - Turnover	10.20	8.90	8.14	36.10	31.50	28.82
Other Costs - Number of Employees	2.78	2.64	2.68	63.31	60.02	61.01

5.3 Boosting

Results on *boosting* will be included at a later stage.

6 Conclusion

So far, classical and robust techniques have been compared and discussed. Table 2 shows how the use of robust statistics impacts the ratio estimate.

Imputation performance after the randomly induced missing mechanism suggests that robust statistics always perform better than the *wls* estimator, based on the MAPE and the RMSE. Most of the times, on the current data Tukey estimator showed a lower prediction error compared to the Huber one.

In robust estimation, as described in section 3.1, the proposed tuning constants k have been shown to be optimal for symmetrical distributions. Considering the asymmetrical nature of these data, a different procedure needs to be developed on the choice of these constants.

Boosting will also be applied to the same data and results will be compared to the above, based on the same accuracy measures.

Analyses on new datasets will be necessary to assess the effectiveness of results.

References

Allison, P. D. (2002). *Quantitative applications in the social sciences: Missing data*. Thousand Oaks, CA, USA: Sage Publications, Inc. doi: 10.4135/9781412985079

de Waal, T., Pannekoek, J., & Scholtus, S. (2011). *Handbook of statistical data editing and imputation*. Wiley.

Hofner, B., Mayr, A., Robinzonov, N., & Schmid, M. (2014). Model-based boosting in R: A hands-on tutorial using the R package mboost. *Computational Statistics*, *29*, 3–35. doi: 10.1007/s00180-012-0382-5

Huber, P. J. (1964). Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, *35*(1), 73–101. doi: 10.1214/aoms/1177703732

Kalton, G., & Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, *12*(1), 1-16. doi: 12-001-X198600114404

Kang, H. (2013). The prevention and handling of the missing data. *Korean journal of anesthesiology*, *64*(5), 402-6. doi: 10.4097/kjae.2013.64.5.402

Kelly, G. E. (1992). Robust Regression Estimators-The Choice of Tuning Constants. *The Statistician*, *41*(3), 303. doi: 10.2307/2348552

Little, R. J. A., & Rubin, D. B. (1986). *Statistical analysis with missing data*. New York, NY, USA: John Wiley & Sons, Inc. doi: 10.1002/9781119013563.ch10

Palmer, S., & Jones, C. (1966). A look at alternate imputation procedures for cps noninterviews. *Bureau of the Census*, 66-459.

Qin, Y., Zhang, S., Zhu, X., Zhang, J., & Zhang, C. (2007). Semi-parametric optimization for missing data imputation. *Applied Intelligence*, *27*(1), 79–88. doi: 10.1007/s10489-006-0032-0

R Core Team. (2013). R: A language and environment for statistical computing.

Schapire, R. E. (1999). A brief introduction to boosting. In *Proceedings of the 16th international joint conference on artificial intelligence - volume 2* (pp. 1401–1406). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Shao, J. (2000). Cold deck and ratio imputation. *Survey Methodology*, *26*(1), 79-85. doi: 12-001-X20000015180

Takahashi, M. (2017). Implementing multiple ratio imputation by the emb algorithm. *Journal of Modern Applied Statistical Method*, *16*(1), 657-673. doi: 10.22237/jmasm/1493598900

Tukey, J. (1977). *Exploratory Data Analysis*. Reading, MA, USA: Addison-Wesley.

Wang, Y.-G., & Bai, Z. D. (2007). Robust estimation using the huber function with a data-dependent tuning constant. *Journal of Computational and Graphical Statistics*, *16*(2), 468-481. doi: 10.1198/106186007X180156