Working Paper

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Editing
(Neuchâtel, Switzerland, 18-20 September 2018)

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

Prepared by Jeroen Pannekoek, Statistics Netherlands

I. INTRODUCTION

1. For the production of its statistical outputs, Statistics Netherlands (SN) and most other national statistical institutes, processes many data sets with missing values. Often, imputation is applied to fill-in the missing data. The imputation methods used are most often relatively simple regression and hot-deck methods. For economic data, ratio-imputation is the most frequently used method.

2. The ratio-model (De Waal et al., 2011, ch. 7.4) is an intuitively plausible model for business statistics. It states that many important variables like turnover, costs, revenues etc., scale with a measure of 'size' of the units. Formally, the ratio-model can be viewed as a special kind of regression model with a single predictor variable (the measure of 'size', for instance number of employees) no intercept and variance proportional to the mean. The model underlies a common estimation method (ratio-estimation) as well as an imputation method (ratio-imputation).

3. This relatively simple model works well for homogeneous data sets without outliers. When outliers do occur however, they affect the estimated ratio and consequently the accuracy of the imputations for the non-outlying units. As a remedy we will apply robust regression techniques to estimate the ratio-model and investigate its performance as an imputation method.

4. Starting from the single predictor robust ratio-models we will also try to improve the predictive power by combining a few of these simple models, using a machine learning technique known as boosting. This method combines the predictions from several models by sequentially adding predictors that most reduce the residuals from the previous model.

5. The performance of the proposed imputation methods will be demonstrated on a data set from the Netherlands structural business statistics and will be compared with traditional ratio-imputation methods.

II. ROBUST RATIO-IMPUTATION

6. A ratio-model for a target variable y and a predictor variable x states that the ratio between y and x is approximately constant, that is, for each unit i we have $y_i \approx Rx_i$, with R the ratio between y and x. The usual estimator for the ratio R is $\hat{R} = \bar{y}/\bar{x}$, with $\bar{x}$ and $\bar{y}$ the means of x and y for a sample

of units with both x and y observed. Ratio-imputation replaces missing y-values by the prediction $\hat{y}_i = \hat{R}x_i$

A. Estimating a ratio by wls regression

7. The estimator $\hat{R}$ defined above can be derived as the weighted least squares (wls) estimator of the regression coefficient in the model

$$y_i = Rx_i + \epsilon_i, \tag{1}$$

where ϵ_i is a random disturbance with $E\epsilon_i = 0$ and $var(\epsilon_i) = \sigma^2 x_i$. Since $E(y_i) = Rx_i$, the variance of the residuals is thus assumed to be proportional to the mean. This increase in variance with the size of the unit, here measured by the value of x (or $\hat{y}$) is often observed in business statistics. The wls estimator $\hat{R}$ of R takes this variance heterogeneity into account by minimising the weighted, by the reciprocal variances, sum of squares of the observed residuals, that is, the loss-function

$$L_w = \frac{1}{2}\sum_i (e_i/\sqrt{x_i})^2 = \frac{1}{2}\sum_i r_i^2, \tag{2}$$

with residuals $e_i = y_i - \hat{R}x_i$ and weighted residuals r_i. Setting the derivative of L_w w.r.t. $\hat{R}$ to zero gives the estimating equation

$$\partial L_w/\partial\hat{R} = \sum_i r_i \left(\partial r_i/\partial\hat{R}\right) = -\sum_i r_i\sqrt{x}_i = \sum_i -y_i + \hat{R}x_i = 0, \tag{3}$$

the solution of which is precisely the commonly used intuitive estimator $\hat{R} = \bar{y}/\bar{x}$.

8. A well known drawback of least squares estimators of regression coefficients is their sensitivity to outliers and several alternatives have been proposed. A well-founded class of robust alternatives are based on the the so-called M-estimators introduced by Huber (1964), see also Huber (2009), and in the next section we will apply commonly used robust M-estimators to the weighted regression approach to estimating the ratio.

B. Estimating a ratio by robust wls regression

9. Commonly used M-estimators for regression minimise the sum of a function ρ of the residuals, de objective can thus be expressed as the minimisation of the loss-function $L_M = \sum_i \rho(e_i)$. The least squares special case is obtained by $\rho(e_i) = \frac{1}{2}e_i^2$. The ρ-function for robust alternatives is chosen such that large residuals have smaller contributions to the loss-function function L_M. Huber proposed, as a robust alternative to least squares, the following ρ-function:

$$\rho_{H}(e_i) = \begin{cases} \frac{1}{2}e_i^2 & |e_i| \le k \\ k|e_i| - \frac{1}{2}k^2 & |e_i| > k. \end{cases} \tag{4}$$

For larger residuals ($> k$) Huber's loss is a function of the absolute values of the residuals whereas for values smaller than $\le k$ it is the standard ols loss. Thus, large outlying residuals have less influence than in the ols-loss, but they still do effect the estimation results. This possible drawback is taken care of by Tukey's (bisquare) ρ-function given by:

$$\rho_{B} = \begin{cases} \frac{k^2}{6}\left\{1 - \left[1 - (e_i/k)^2\right]^3\right\} & |e_i| \le k \\ k^2/6 & |e_i| > k, \end{cases} \tag{5}$$

which, for values larger than k, does not depend on the value of the residuals and therefore completely rejects gross outliers in the estimation procedure.

10. The tuning constant k in these loss-functions depends on the scale of the residuals and can be written as $C\sigma$, with σ the standard deviation of the residuals. Common choises for C are 1.345 for the Huber loss and 4.685 for the bisquare. These choices correspond to 95% efficiency in the case of normally distributed residuals.

11. To obtain a robust alternative for the ratio-estimator, based on M-estimation, we note that, in line with the 'variance increasing with size' assumption, the ratio-estimator minimises weighted residuals $r_i = e_1/\sqrt{x_i}$ rather than e_i. Robust ratio estimators, by using M-estimation, can therefore be defined by replacing e_i by r_i in (4) and (5).

12. The robust estimators can be found by solving the estimating equation obtained by setting the derivative of the loss-function L_M w.r.t $\hat{R}$ to zero.

$$\partial L_M/\partial \hat{R} = \sum_i \psi(r_i)\left(\partial r_i/\partial \hat{R}\right) = \sum_i \psi(r_i)\sqrt{x_i} = 0, \tag{6}$$

with $\psi(r_i) = \partial\rho(r_i)/\partial r_i$. The solution to this estimating equation is the robust estimated ratio-regression parameter. To obtain this solution we can rewrite (6) as $\sum_i w_i r_i \sqrt{x_i} = 0$ with weights $w_i = \psi(r_i)/r_i$ and notice that it is a weigthed version of the estimating equation (3) and it solves the weighted least square loss-function $\sum_i w_i r_i^2$. For the solution an iteratively reweighted least squares procedure is needed. The procedure is iterative because it is weighted least squares with weights that depend on the residuals and therefore depend on the current parameter estimates and need to be updated iteratively. The solution satisfies

$$\hat{R}_M = \frac{\sum_i w_i y_i}{\sum_i w_i x_i}, \tag{7}$$

which shows that the robust ratio-estimator is the ratio of the *weighted* means of y and x with weights that are equal for the x and y values of the same unit.

13. The weights for the Huber loss-function are given by

$$w_{H.i} = \begin{cases} 1 & |r_i| \le k \\ k/|r_i| & |r_i| > k. \end{cases} \tag{8}$$

For larger residuals ($> k$) Huber's loss is a function of the absolute values of the residuals whereas for values $\le k$ it is the standard ols-loss. Thus, large outlying residuals have less influence than in the ols-loss, but they still do effect the estimation results. This possible drawback is taken care of by Tukey's (bisquare) ρ-function given by:

$$w_{T.i} = \begin{cases} \left[1 - (r_i/k)^2\right]^2 & |r_i| \le k \\ 0 & |r_i| > k, \end{cases} \tag{9}$$

which, for values larger than k, does not depend on the value of the residuals and therefore effectively rejects gross outliers in the estimation procedure.

C. Application to ratio-imputation of SBS-variables

14. In this subsection we will illustrate the possible benefits of using the robust methods by applying them to a data set from the Netherlands structural business statistics. The data set for this relatively simple example contains units from the business category 'Wholesalers'. Two variables, *Turnover* (total turnover) and *Employees* (total number of employees) are used as predictor variables. These variables are often used as auxiliary variables for imputation because they characterise the size of a business in monetary value and in employment. Moreover, these total variables usually contain

hardly any missing values and, at least for *Employees*, there may be values available from administrative sources. As target variables for imputation we have chosen four variables related to the breakdown of the total costs or expenditures of a company: *Personnel* (costs of personnel), *Purchases* (costs of purchases), *Depreciations* (value of depreciations) and *Other* (other costs). The data set used here consists of 274 records without missing values on the variables of our example.

15. For the example data set we have calculate different estimators for a number of ratios between the target variables and the predictor variables. Apart from the robust estimators *Huber* and *Tukey* these also include the conventional ratio of the means, *Means* and the median of the ratios y_i/x_i which is a simple robust alternative that is sometimes advocated in the context of score functions for selective editing. The results are in Table 1. The predictor variable (denominator) used for each of the target variables (numerator) was chosen because it turned at the be the best predictor (in the sense of the relative percentage difference criterion defined in the next paragraph).

TABLE 1. Different estimators of the ratio of 'Target variable' to 'Predictor variable'

Ratio	Means	Median	Huber	Tukey
Personnel costs / Employees	37.8	38.6	37.6	37.2
Cost of purchases / Turnover	0.87	0.78	0.83	0.82
Depreciations / Turnover	0.0067	0.0086	0.0071	0.0075
Other costs / Employees	0.045	0.075	0.055	0.056

The robust estimators can be higher (predominantly outliers with low values) lower (predominantly outliers with high values) or about the same as the non-robust ratio of the means (Means). The Median is in all cases the most extreme estimator in the sense that it deviates most from Means.

16. Using these estimators, predictions were calculated for the four target variables and all 274 units. To asses how well these ratio-models perform in predicting the target variable, the mean of the predicted values was compared to the observed mean. The relative percentage difference between these means ($100 \times |\text{mean predicted} - \text{observed mean}|/\text{observed mean}$) is displayed in Table 2.

TABLE 2. Goodness of fit: % relative difference between observed mean and mean of predicted values

Target variable	Means	Median	Huber	Tukey
Personnel costs	0	1.0	0.3	0.8
Cost of purchases	0	5.0	2.2	2.8
Depreciations	0	14.0	2.9	5.6
Other costs	0	4.7	6.0	8.8

The mean of the predictions by the ratio of means is by definition always equal to the mean of the observed values. The robust ratio estimators can differ considerably from the observed mean. This is in particular the case if the median of the ratios is used.

17. The question for imputation purposes, however, is the accuracy of the out of sample predictions. To evaluate the imputation performance we have therefore simulated a random missing value mechanism with 10% missing values for each of the target variables separately. This percentage missing values, including values that are discarded because they are erroneous, is common among these variables. The above results pertained to using the optimal predictor(Turnover or Employees) in any case. In practice this is will not always be the case. Predictors are chosen beforehand and based on there availability. Therefore we used both predictors to evaluate the imputation performance.

18. The imputation performance is assessed by comparing the imputed mean by the observed mean, were the imputed mean is the mean of the data after filling in the 10% missing values by predictions from the imputation method. Since the 10% missing values are assigned randomly, we have repeated this procedure 500 times, each time evaluating the percentage relative difference between the observed and imputed means and then averaging over these 500 values. Table 3 shows the results.

TABLE 3. Imputation performace: % relative difference between imputed mean and observed mean

Predictor:Turnover				
	Personnel cost	Cost of purchases	Depreciations	Other costs
Means	7.5	1.1	8.5	8.0
Median	8.9	1.2	7.7	8.5
Huber	7.2	1.0	7.6	7.4
Tukey	7.0	1.0	7.2	6.9
Predictor: Employees				
Means	1.3	8.1	2.7	2.5
Median	1.1	5.6	2.6	2.4
Huber	1.2	5.6	2.5	2.5
Tukey	1.2	5.7	2.6	2.6

19. The results show that the robust estimators Huber and Tukey show an improvement over the conventional Means estimator, except for the imputation of Other costs by the predictor Employees, where Tukey is slightly worse than Means. This is not true for the Median estimator that can have a considerably larger error than Means. The gain of Huber and Tukey over Means is not very substantial if the imputation error of Means is already small, which is the case for the imputation of Cost of purchases by Turnover and Personnel cost, Depreciations and Other costs by Employees. It should be noted however that in this data set 'obvious errors' such as unity measure errors have already been removed.

III. COMBINING PREDICTORS BY BOOSTING

20. One way of improving the accuracy of prediction is to use multiple predictors. The classical ratio-imputation method, however, is confined to a single predictor variable only. Since the single ratio-model can be cast as a single regression model (section A), we could consider a multiple regression model to make use of multiple x-variables. But this does not results in an obvious and unique generalisation of the ratio-model. The difficulty is that the ratio model requires the weighting with $1/x$ which does not directly generalises to the situation with multiple x-variables.

21. Another way of improving the accuracy is by applying the ratio-model separately to different strata. This is actually often used in practice, where separate ratios for different size classes or NACE-groups often yield a much better fit than the same ratio for all groups. A difficulty with this approach is that it is not always obvious before hand how to divide the data in strata such that the within strata variation in ratios is small. Moreover, the optimal division in strata may very well differ among target variables. Automatic classification of units to homogeneous strata can be performed by regression trees (see, Hastie et al., 2009). An application of regression trees for imputation in official statistics is described in Zabala (2015).

22. For these reasons we will investigate the gradient boosting technique (Friedman, 2001) that can combine different predictors in a straightforward and flexible way, including the automatic classification of units in homogeneous strata by regression trees. Boosting is an ensemble technique. It combines the results of a collection of simple predictors for the same target variable with the purpose of improving over the prediction of any single variable alone. Ensemble techniques are generally divided into bagging and boosting. While bagging uses averaging techniques to combine a number of predictions and mainly reduces variance, boosting is more directly targeted at optimising predictions. Boosting uses predictors sequentially, by combining predictors in a stepwise manner, adding in each step the predictor that gives the largest reduction of the prediction errors.

A. The boosting approach to model building

23. Gradient boosting uses stage-wise additive fitting (Friedman, 2001). Suppose that there are a number of simple *base* predictors $\hat{b}_j(x)$ depending on some of the available predictor variables in x. Such predictors could be a ratio-estimator using a (robust) estimated ratio and, say, variable x as a predictor. Or it may be a prediction based on an estimated regression tree using one or more of the predictor variables. The gradient boosting approach to model building starts by evaluating the prediction error of each of the available base predictors. The first stage model then is the model with the best predictor only:

$$\hat{y}_1 = \hat{b}_1(x), \tag{10}$$

with residuals $r_1 = y - \hat{y}_1$.

24. At the second stage, all base predictors are fitted to the first stage residuals and the one with minimum mean squared residual is chosen as the second base-predictor, thus arriving at the second stage model

$$\begin{aligned} \hat{r}_1 &= \hat{b}_2(x) \text{ and} \\ \hat{y}_2 &= \hat{b}_1(x) + \hat{b}_2(x) \end{aligned} \tag{11}$$

with residuals $r_2 = y - \hat{y}_2$. This process of repeatedly fitting base predictors to current residuals is continued until some fitting criterion is met, thereby taking account of the dangers of over-fitting for predictive purposes. The AIC criterion and cross-validation are often used for deriving stopping rules.

25. Note that, contrary to stepwise model selection in regression problems, the parameters of previous steps remain the same and are not updated when predictors are added. Clearly this makes the meaning of the estimated parameters dependent on the order in which they are added to the model, but this is not so much an issue with boosting since the purpose is only to optimise the prediction.

B. Defining the base predictors

26. The base predictors (or *base learners*) that will be used in our application include, of course, a base-predictor using a robust ratio estimator, which will be denoted by

$$\hat{b}_{rr}(x) = \hat{R}x, \tag{12}$$

were x denotes the predictor variable and $\hat{R}$ the robust estimator of the ratio, for which the Huber-estimator will be used in the following.

27. We will also consider a regression tree with a single split-point on the x-variable (a stump), which estimates the target variable by fitting two means, one for values of x smaller than the split-point and one for values of x larger than the split-point. The split-point is chosen to achieve the best fit of

this simple two-means model to the observed y-values. This tree-mean (or two-mean) base-predictor is denoted as

$$\hat{b}_{tm} = \hat{\mu}_1 \times I(x \leq s_x) + \hat{\mu}_2 \times I(x > s_x) \tag{13}$$

with I the 0-1 indicator function with $I(.) = 1$ of its argument is true and zero otherwise, s_x the split-point on variable x and $\hat{\mu}_1$, $\hat{\mu}_2$ estimators of the means in the two subgroups based on the observed y-values.

28. The base-predictor $\hat{b}_{tm}$ makes limited use of the information in the x-variable which is in contrast to the known (or assumed) strong relation between the predictor and target variables that underlies the ratio-model. Therefore we may modify the standard regression tree-based predictor by using a robust ratio-estimator in each of the two categories of x defined by the binary split. This tree-ratio base-predictor is given by

$$\hat{b}_{tr} = \hat{R}_1 \times I(x \leq s_x) + \hat{R}_2 \times I(x > s_x), \tag{14}$$

with $\hat{R}_1$ and $\hat{R}_2$ robust estimators for the ratio with the categories of x defined by the split.

C. Application of adding base predictors to the ratio-models

29. Using the boosting algorithm we can extend the simple ratio-model by adding base predictors in an automated stage-wise procedure. Here we have limited the algorithm to selecting models with at most four base predictors. In Table 4 we show the imputation performance for a number of models for each of the target variables. Apart from the best model (in terms of the % difference between observed and imputed means) for each of the target variables, for which the result is shown in bold, we also show the results for a number of other models to compare with.

30. The simplest model is the Huber estimator already shown in Table 3, for the predictor Turnover this is the first model in the top panel of Table 4 ($\hat{b}_{rr}(T)$). and for the predictor Employees this is the first model in the bottom panel ($\hat{b}_{rr}(E)$). The second model in both panels extends the first model by adding the robust ratio base-predictor for the other predictor variable not in the first model. The third model adds to the robust ratio-estimator the tree-ratio base-predictor with the same x-variable and then the ratio-predictor for the other variable. The last model also adds the tree-ratio base-predictor for the other variable.

31. The simplest robust ratio-model is the best model for the target variable Depreciations imputation with extension of this model with multiple predictors did not lead to improvement in estimating the mean. The model with two robust-ratio base predictors, one for each predictor variable, is best for the target variable Other, but the improvement over the model with only the robust-ratio predictor with Employees as predictor variable is small. The third model which includes a tree–ratio base-predictor as well as the two ratio base predictors is best for the variables Purchases and Personnel cost. The improvement over the single predictor model is large for Purchases were the estimation error is almost halved.

32. The estimated robust ratio for Turnover is, for each of the target variables, the same for all models because of the stage-wise fitting. For the variable Purchases, for instance, the ratio is 0.83 averaged over the 500 samples. Adding to the simple robust ratio base-predictor the tree-ratio base-predictor with the same x-variable (a model with as first two base-predictors: $\hat{b}_{rr}(T) + \hat{b}_{tr}(T)$) adds the parameters -0.02 and 0.15 with a split-point at T=146731. This model implies that the estimated ratio of Purchases to Turnover is 0.81 units with smaller values of Turnover and 0.98 for larger values. Apparently, the margins are smaller for larger wholesalers. The best model for Purchases is obtained

by also including the base learner $\hat{b}_{rr}(E)$. This model improves on the predictions of the previous one by adding the predictions of the residuals obtained by including the variable Employment in the model. The small numbers of base-predictors in these models leads to easily interpretable models.

TABLE 4. Imputation performance: % relative difference between imputed mean and observed mean

Extensions of ratio-imputation by *Turnover*				
Model	Personnel cost	Purchases	Depreciations	Other
$\hat{b}_{rr}(T)$	7.15	1.02	7.59	7.39
$\hat{b}_{rr}(T) + \hat{b}_{rr}(E)$	7.03	0.97	7.04	6.88
$\hat{b}_{rr}(T) + \hat{b}_{tr}(T) + \hat{b}_{rr}(E)$	2.51	**0.52**	3.76	2.95
$\hat{b}_{rr}(T) + \hat{b}_{tr}(T) + \hat{b}_{rr}(E) + \hat{b}_{tr}(E)$	3.18	0.59	4.33	3.34
Extensions of ratio-imputation by *Employees*				
Model	Personnel cost	Purchases	Depreciations	Other
$\hat{b}_{rr}(E)$	1.22	5.64	**2.46**	2.48
$\hat{b}_{rr}(E) + \hat{b}_{rr}(T)$	1.18	4.40	2.47	**2.33**
$\hat{b}_{rr}(E) + \hat{b}_{tr}(E) + \hat{b}_{rr}(T)$	**1.13**	4.16	2.65	2.64
$\hat{b}_{rr}(E) + \hat{b}_{tr}(E) + \hat{b}_{rr}(T) + \hat{b}_{tr}(T)$	1.28	5.68	2.55	2.77

IV. **Conclusion**

33. In this paper we have investigated methods to improve on the often applied ratio-imputation method. Two lines of improvement were followed. One is the robustness of the estimated parameter(s) of the imputation model against outliers and the other is to increase the precision of the imputation by combing different predictors.

34. With respect to robustness we derived ratio-estimators based on well known robust-regression techniques in particular two M-estimators according to proposals of Huber and Tukey. These robust estimators can improve imputations considerably in some cases and outperform the conventional non-robust estimator as well as a the median of ratios which is a simple robust alternative. Especially the Huber-estimator seems a safe choice because it performed never worse than the standard non-robust estimator, at least in our limited application. This may be a more general property that is related to its known high efficiency. Nevertheless, other robust-estimators could also be investigated as well as the tuning parameters for which we have only used the 'default' proposals so far.

35. A stage-wise gradient boosting framework was used the combine different predictors to increase the precision of imputations. This framework is very flexible and makes it easy to combine, for instance, ratio-estimators with different predictors, regression-trees and ratio-estimators within the leaves of a regression-tree.

36. The methods were tried on a application but the results seem promising and much more extensive applications will be investigated. For other applications we may also come up with other and more base-predictors than the few investigated here. That would also mean that we could make more use of the model selection procedures usually applied in the boosting framework than we did with the limited number of base-predictors considered here.

References

de Waal, T., J. Pannekoek, and S. Scholtus (2011), *Handbook of Statistical Data Editing and Imputation*. John Wiley & Sons, New York.

Hastie, T. R. Tibshirani and J. Friedman (2009), *The elements of statistical learning: data mining, inference and prediction. 2nd ed.*. Springer, New York.

Huber, P. J. (2009). *Robust Statistics (2nd ed.)*. John Wiley & Sons, New York.

Friedman, J. (2001), Greedy function approximation: A gradient boosting machine. *Ann. of Statist.* 29, 1189-1232.

Zabala, F. (2015), Let the data speak: Machine learning methods for data editing and imputation. *UNECE Work Session on Statistical Data Editing* (Budapest, Hungary)