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

Workshop on Statistical Data Editing

Neuchâtel, Switzerland, 18-20 September 2018

Jeroen Pannekoek

Introduction

Ratio imputation

- Most often used imputation method for business statistics.
- Based on a model of "scaling with size". Many variables are proportional to size: Turnover, Costs, Employees

Imputation for a unit is performed by using the model prediction:

$$\hat{y}_{miss,i} = \hat{R} x_{obs,i} \quad \text{and} \quad \hat{R} = \frac{\bar{x}_{obs}}{\bar{y}_{obs}}$$

Currently we are investigating improvements in two directions:

- Robust estimator of R
- Using more auxiliary variables than in the single predictor ratio-model

But keeping the easy configuration and interpretation of the standard model

Contents

Robust ratio-imputation

Example using data from SBS

Combining robust ratio-predictors by boosting

Example continued.

Robust ratio-imputation

Ratio-model can be viewed as a weighted regression model:

Model: $y = Rx + \epsilon$ with $var(\epsilon) = \sigma^2 x$ Prediction: $\hat{y}_i = \hat{R}x$

Estimation by wls, **minimise**: $\sum_{i \in obs} (\hat{y}_i - y_i)^2 / x_i$ => $\widehat{\boldsymbol{R}} = \frac{\sum_i y_i}{\sum_i x_i}$

Apply robust-regression techniques to estimate ratio-model:

M-estimation (Huber): large residuals are downweighted.
We apply this to the 1/x-weigted residuals of the ratio estimator:

Estimation by irwls, **minimise**: $\sum_{i \in obs} w_i (\hat{y}_i - y_i)^2 / x_i$ => $\widehat{\boldsymbol{R}} = \frac{\sum_i w_i y_i}{\sum_i w_i x_i}$

Ratio of weighted means with weights equal for x and y values of the same unit. Different from separate robust estimators of $\bar{x}$ and $\bar{y}$

Weight functions

Huber:
$|r| < 1.3\sigma$ -> w=1
declining gradually

Tukey:
$r = 0$ -> w=1
$|r| > 4.7\sigma$ -> w=0
cut-off

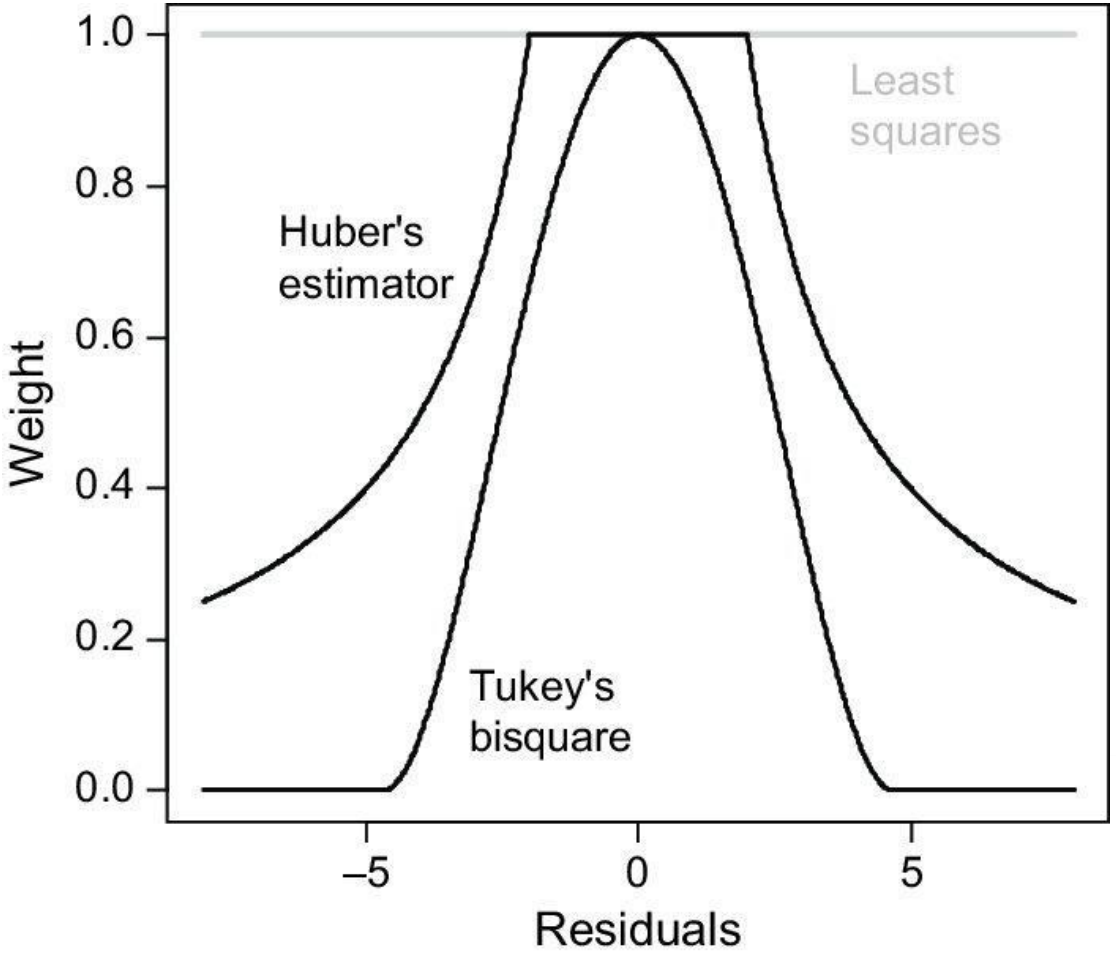


Illustration: ratio-imputation of SBS-variables

Data: 271 units of SBS for Wholesalers

- **Introduce non respons**: 10% at random in a target variable
- **Impute by ratio-imputation**
- **Estimate the mean** using the imputed data (replacing the non-respons by predictions).
- **Evaluate imputation error** in estimating the mean:

  Error is % absolute difference between imputed mean and true mean .
- **Repeat 500 times** (random nonrespons) and average error measure.

Example ratio-imputation of SBS-variables

Imputed by ratio imputation with 4 different estimators $\hat{R}$

Means: $\bar{y}/\bar{x}$ Median: $med(y_i/x_i)$ Huber/Tukey-Robust

Averaged % imputation error in the mean.

Predictor *Turnover*

Target	Means	Median	Huber	Tukey
Personnel costs	7.5	8.9	7.2	7.0
Cost of purchases	1.1	1.2	1.0	1.0
Depreciations	8.5	7.7	7.6	7.2
Other costs	8.0	8.5	7.4	6.9

Example ratio-imputation of SBS-variables

Imputed by ratio imputation with 4 different estimators $\hat{R}$

Means: $\bar{y}/\bar{x}$ Median: $med(y_i/x_i)$ Huber/Tukey-Robust

Averaged % imputation error in the mean.

Predictor *Employees*

Target	Means	Median	Huber	Tukey
Personnel costs	1.3	1.1	1.2	1.2
Cost of purchases	8.1	5.6	5.6	5.7
Depreciations	2.7	2.6	2.5	2.6
Other costs	2.5	2.4	2.5	2.6

Other improvement: combining predictors by boosting

Boosting combines different predictors for the same variable.
Uses stage-wise sequential fitting.

Apply to robust ratio-estimators:

1. Fit first model to y: $\hat{y}_1 = \hat{R}_1 x_1$ with residuals $r_1 = y - \hat{y}_1$
2. Fit second model to residuals r_1: $\hat{r}_1 = \hat{R}_2 x_2$
   Improve the first estimate $\hat{y}_1$ by adding the estimated residuals
   $\Rightarrow \hat{y}_2 = \hat{y}_1 + \hat{r}_1 = \hat{R}_1 x_1 + \hat{R}_2 x_2$ with residuals $r_2 = y - \hat{y}_2$

Additive model. But different from regression!

- Stage-wise sequential fitting different from simultaneous fitting.
- Very flexible in choise of “base-predictors” to be combined.

Base predictors

1. Robust ratio-estimator. R(X)

   Fits a robust ratio-model. Use X=*Employees* and X=*Turnover*

2. Tree-model with two leaves (stump). Tree-mean. TM(X)

   Split on (x) and fit two means: $\bar{y}_1$ for x < split(x)

   $\bar{y}_2$ for x ≥ split(x)

3. Tree-model with two ratio estimators. Tree-ratio. TR(X)

   Fits two ratio-models: $\hat{y}_1 = \hat{R}_1 x$ for x < split(x)

   $\hat{y}_2 = \hat{R}_2 x$ for x ≥ split(x)

Application to SBS data

Number of models tested with up to 4 base predictors.

Single robust ratio: R(T)

Best model : R(T) + R(E) + TR(T)

Overfitted model : R(T) + R(E) + TR(T) +TR(E)

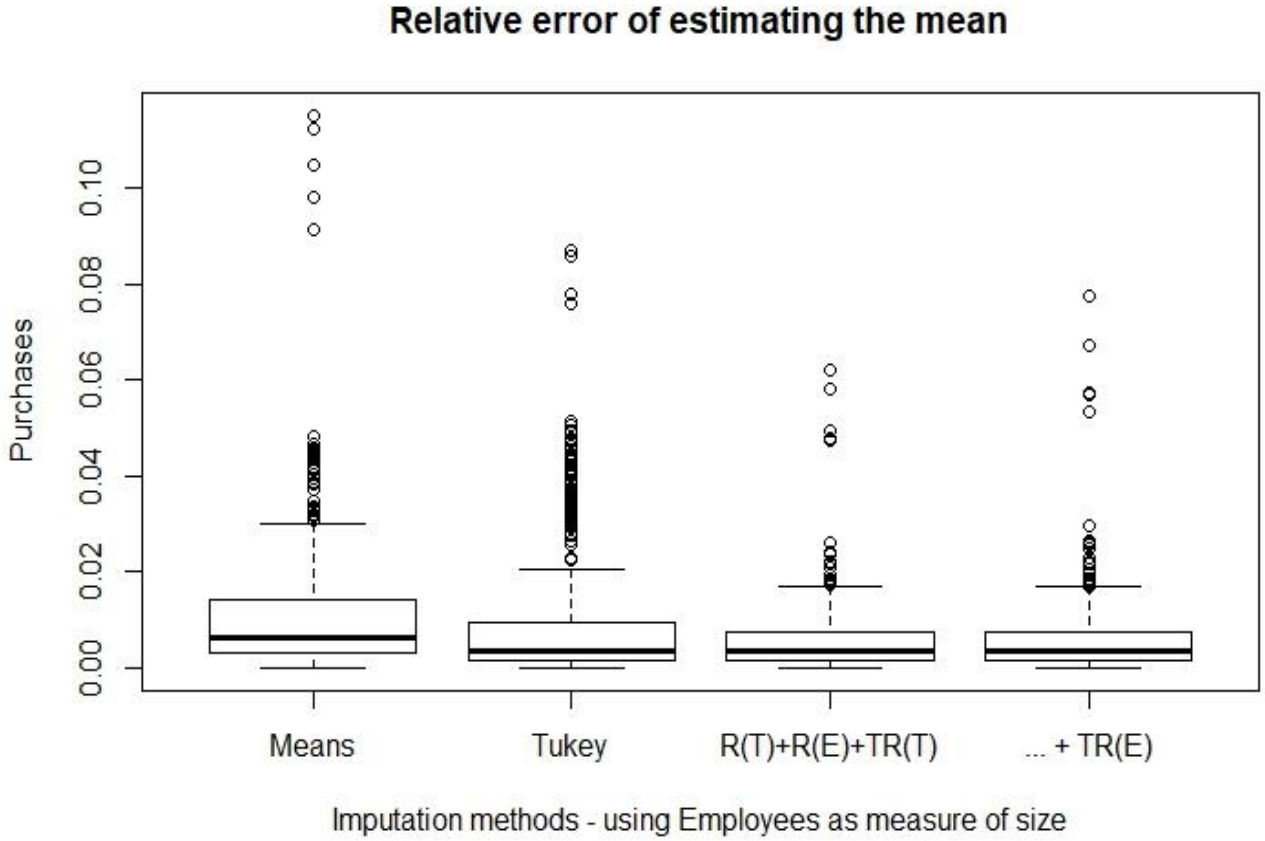
Averaged % imputation error in the mean.

Target	Single	Best	Overfitted
Personnel costs	7.2	2.5	3.2
Cost of purchases	1.0	0.5	0.6
Depreciations	7.6	3.8	4.3
Other costs	7.4	3.0	3.3

Avaraged parameters of “best” model

For costs of purchases:

$$\hat{y} = 0.83T + \begin{cases} 0.15T \\ -0.02T \end{cases} - 15E = \begin{cases} 0.98T \\ 0.81T \end{cases} - 15E$$


Relative error of estimating the mean
Purchases
0.00
0.02
0.04
0.06
0.08
0.10
Means
Tukey
R(T)+R(E)+TR(T)
... + TR(E)
Imputation methods - using Employees as measure of size

Concluding remarks

- Ratio estimators can be made more robust by applying *M*-estimation techniques. Works better than a median.
- Combining robust ratio estimators can be done easily by boosting: sequential stage-wise fitting of the residuals of a previous model. Easy to implement and understand.
- In a limited experiment we saw considerable improvements of simple extensions of the traditional model in cases where this model didn't perform well.
- We will do a follow-up study with much more data-sets and further investigate the tuning of the parameters of the methods and automated model selection.