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 \( \hat{R} \)
• 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: $\sum_{i \in obs} (\hat{y}_i - y_i)^2 / x_i \Rightarrow \hat{R} = \frac{\sum_i y_i}{\sum_i x_i}$

Apply robust-regression techniques to estimate ratio-model:

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

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

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:
$|r| < 1.3\sigma \rightarrow w=1$
declining gradually

Tukey:
$r = 0 \rightarrow w=1$
$|r| > 4.7\sigma \rightarrow w=0$
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*

<table>
<thead>
<tr>
<th>Target</th>
<th>Means</th>
<th>Median</th>
<th>Huber</th>
<th>Tukey</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personnel costs</td>
<td>7.5</td>
<td>8.9</td>
<td>7.2</td>
<td>7.0</td>
</tr>
<tr>
<td>Cost of purchases</td>
<td>1.1</td>
<td>1.2</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Depreciations</td>
<td>8.5</td>
<td>7.7</td>
<td>7.6</td>
<td>7.2</td>
</tr>
<tr>
<td>Other costs</td>
<td>8.0</td>
<td>8.5</td>
<td>7.4</td>
<td>6.9</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Target</th>
<th>Means</th>
<th>Median</th>
<th>Huber</th>
<th>Tukey</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personnel costs</td>
<td>1.3</td>
<td>1.1</td>
<td>1.2</td>
<td>1.2</td>
</tr>
<tr>
<td>Cost of purchases</td>
<td>8.1</td>
<td>5.6</td>
<td>5.6</td>
<td>5.7</td>
</tr>
<tr>
<td>Depreciations</td>
<td>2.7</td>
<td>2.6</td>
<td>2.5</td>
<td>2.6</td>
</tr>
<tr>
<td>Other costs</td>
<td>2.5</td>
<td>2.4</td>
<td>2.5</td>
<td>2.6</td>
</tr>
</tbody>
</table>
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 $y$: $\hat{y}_1 = R_1 x_1$ with residuals $r_1 = y - \hat{y}_1$
2. Fit second model to residuals $r_1$: $\hat{r}_1 = R_2 x_2$

   Improve the first estimate $\hat{y}_1$ by adding the estimated residuals
   $\Rightarrow \hat{y}_2 = \hat{y}_1 + \hat{r}_1 = R_1 x_1 + R_2 x_2$ 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 choice of “base-predictors” to be combined.
Base predictors

1. Robust ratio-estimator. \( R(X) \)
   Fits a robust ratio-model. Use \( X=\text{Employees} \) and \( X=\text{Turnover} \)

2. Tree-model with two leaves (stump). Tree-mean. \( TM(X) \)
   Split on \( (x) \) and fit two means:
   \[ \bar{y}_1 \text{ for } x < \text{split}(x) \]
   \[ \bar{y}_2 \text{ for } x \geq \text{split}(x) \]

3. Tree-model with two ratio estimators. Tree-ratio. \( TR(X) \)
   Fits two ratio-models:
   \[ \hat{y}_1 = \hat{R}_1 x \text{ for } x < \text{split}(x) \]
   \[ \hat{y}_2 = \hat{R}_2 x \text{ for } x \geq \text{split}(x) \]
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.

<table>
<thead>
<tr>
<th>Target</th>
<th>Single</th>
<th>Best</th>
<th>Overfitted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personnel costs</td>
<td>7.2</td>
<td>2.5</td>
<td>3.2</td>
</tr>
<tr>
<td>Cost of purchases</td>
<td>1.0</td>
<td>0.5</td>
<td>0.6</td>
</tr>
<tr>
<td>Depreciations</td>
<td>7.6</td>
<td>3.8</td>
<td>4.3</td>
</tr>
<tr>
<td>Other costs</td>
<td>7.4</td>
<td>3.0</td>
<td>3.3</td>
</tr>
</tbody>
</table>
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

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.