Generalized robust ratio estimator for imputation

Kazumi Wada (kwada@nstac.go.jp)*, Keiichiro Sakashita (ksakashita@nstac.go.jp) *

Keywords: M-estimators, Outlier, Iteratively reweighted least squares.

1. INTRODUCTION

The robust ratio estimator described in this paper was developed for the imputation of the 2016 Economic Census for Business Activity in Japan.

The 2016 Census was conducted by the Ministry of Internal Affairs and Communications and the Ministry of Economy, Trade and Industry on June 1, 2016. It aims to identify the structure of establishments and enterprises in all industries on a national and regional level, and to obtain basic information to conduct various statistical surveys by investigating the economic activity of these establishments and enterprises.

The major corporate accounting items, such as sales, expenses and salaries, surveyed by the census require imputation to avoid bias. Although ratio imputation is a leading candidate, it is well known that the ratio estimator is very sensitive to outliers; therefore, we need to take appropriate measures for this problem.

2. METHODS

Conventional ratio estimator has a heteroscedastic error term that is proportional to the variance of x. We first segregate the homoscedastic error term with no relation to x, from the original error term. It is necessary to robustify the estimator by means of M-estimation for regression. The reformed estimator can be expanded into different error terms with regards to its relationship with x. The different error terms give dissimilarity to the characteristics of the estimator. A few examples are briefly described below.

2.1. Ratio Imputation

Ratio imputation is a special case of regression imputation [1]. When there are missing values in the target variable *y*, a single auxiliary variable *x* without missing values is used to estimate the missing *y* values. Therefore, *x* must be chosen from the variables that are highly correlated with *y*. The imputation model is as follows:

$$y_j = rx_j + \epsilon_j, \tag{1}$$

where $i = 1, \dots, N$ of (x, y) is observed on each of the N units in the domain for imputation. Because the true ratio r is usually unknown due to the missing values of *y*, the estimated ratio

$$\hat{r} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i},$$

is used to substitute the missing *y* values such that

$$\hat{y}_i = \hat{r}x_i,$$

* National Statistics Center, Japan

where $i = 1, \dots, n$ of (x, y) is observed on each of the n units in the domain excluding those with missing values. The ratio estimator of the model (1) is BLUE (the best linear unbiased estimator) under the following two conditions: (i) the relationship between the variables *y* and *x* is a straight line through the origin and (ii) the variance of *y* about this line is proportional to *x* [2].

2.2. Generalization of the ratio estimator

The model

$$y_i = a_0 + a_1 x_i + \varepsilon_i,$$

is an example of a simple regression where a_0 is the intercept and a_1 is the slope. The error term of the model is supposed to be normal with a mean of 0 and a constant variance, which can be written as $\varepsilon_i \sim N(0, \sigma^2)$. Meanwhile, the error term ϵ_i of the ratio estimator model in (1) should be proportional to $\sqrt{x}$, i.e. the variance of ϵ_i is proportional to *x* and can be written as $\epsilon_i \sim N(0, x\sigma^2)$. Because these two error terms have the relationship $\epsilon_i = \sqrt{x_i}\varepsilon_i$, the ratio model, Eq. (1), can be written in the following form:

$$y_i = r x_i + \sqrt{x_i}\varepsilon_i. \tag{2}$$

We refer to ε_i in the ratio estimator model hereafter as the quasi-error term because the true error term is ϵ_i. Then, we can extend the model (2), to have an error term that is proportional to x_i^{β} as follows:

$$y_i = r x_i + x_i^{\beta}\varepsilon_i. \tag{3}$$

The corresponding ratio estimator becomes

$$\hat{r} = \frac{\sum_{j=1}^{n} y_j x_j^{1-2\beta}}{\sum_{j=1}^{n} x_j^{2(1-\beta)}}. \tag{4}$$

This model (3) and its estimator (4) broaden the definition of the conventional ratio estimator. Eq. (3) contains various models according to the value of β. A few examples are shown in Table 1. The original ratio estimator corresponds to case B'.

Table 1. Variations in the estimator depending on $\boldsymbol{\beta}$

Case	$\boldsymbol{\beta}$	Model	Estimator	Quasi-error term
A'	$\beta = 1$	$y_i = r x_i + \varepsilon_i x_i$	$\hat{r} = \frac{1}{n}\sum \frac{y_i}{x_i}$	$\varepsilon_i = \frac{y_i}{x_i} - r \sim N(0, \sigma^2)$
B'	$\beta = 1/2$	$y_i = r x_i + \varepsilon_i \sqrt{x_i}$	$\hat{r} = \frac{\sum y_i}{\sum x_i}$	$\varepsilon_i = \frac{y_i}{\sqrt{x_i}} - r\sqrt{x_i} \sim N(0, \sigma^2)$
C'	$\beta = 0$	$y_i = r x_i + \varepsilon_i$	$\hat{r} = \frac{\sum y_i x_i}{\sum x_i^2}$	$\varepsilon_i = y_i - r x_i \sim N(0, \sigma^2)$

2.3. Characteristics of the models

Cases A', B' and C' have different features. Of these models, we focus in particular on A' and B', because C' is a regression model without an intercept. Characteristics of regression models are well known and those with homoscedastic error terms do not fit the targeted census data. The ratio estimator A' is obtained by means of the ratios of

each observation. Even though the estimator A' has the possibility of having a very large variance due to the formula, it is not significantly affected by large values of *x* and/or *y*, unlike B'. The conventional ratio estimator B' is a ratio of the sums or means of *x* and *y*. The estimator is very stable, i.e. it is less likely to have a very large variance. However, the estimation is highly dependent on the large-scale observations of *x* and *y*.

2.4. Robustification

The robustified generalized ratio estimator of (4) is derived by means of M-estimation as follows:

$$\hat{r}_{rob} = \frac{\sum w_i y_i {x_i}^{1-2\beta}}{\sum w_i {x_i}^{2(1-\beta)}},$$

where w_i is obtained by Tukey's biweight function as shown below:

$$w\left(\frac{\check{\varepsilon}}{\sigma}\right) = w(e) = \begin{cases} \left[1 - \left(\frac{e}{c}\right)^2\right]^2 & |e| \le c \\ 0 & |e| > c. \end{cases}$$

From (3), the quasi-residuals based on the homoscedastic quasi-error term are derived such that

$$\check{\varepsilon}_i = \frac{y_i - \hat{r} x_i^{\beta}}{x_i^{\beta}}.$$

The cases with $\beta = 1$ and $\beta = 1/2$ are shown in Table 2. The corresponding models are similar to those for cases A' and B'.

Table 2. The robustified estimators

Case	β	Estimator	Quasi-residuals
A	$\beta = 1$	$\hat{r}_{robA} = \frac{\sum w_i (y_i/x_i)}{\sum w_i}$	$\check{\varepsilon}_i = \frac{y_i}{x_i} - \hat{r}_{robA}$
B	$\beta = 1/2$	$\hat{r}_{robB} = \frac{\sum w_i y_i}{\sum w_i x_i}$	$\check{\varepsilon}_i = \frac{y_i}{\sqrt{x_i}} - \hat{r}_{robB}\sqrt{x_i}$

3. RESULTS

The purpose of the study was to apply a robustified ratio estimator to the 2016 Economic Census. Model selection between the estimators A and B was made using previous census data. The estimator B was chosen based on the simulation results.

Then, random number simulations were conducted to confirm the performance. These simulations were performed with *x* uniform random numbers from 1000 to 1100, the ratio $r = 2$ and the quasi-error terms ε_i were random numbers following the t-distribution of the degrees of freedom: 1, 2, 3, 5, 10 and infinite. The objective variable *y* was calculated based on the model B' using the above mentioned components. For each simulation, 100,000 data-sets of size n = 100 were generated with a given degree of freedom of the t-distribution for the quasi-error term. Table 3 shows the number of iterations needed to compute the estimator B. At least two iterations are necessary because the initial value calculation is counted as one iteration. As the tails of the quasi-error terms become longer,

the number of iterations tends to increase; however, it is obvious that the conversion is sufficiently fast. Table 4 shows the accuracy of the estimator B compared to the estimator B'. It illustrates that the estimator B successfully improved the root mean square error (RMSE) when the quasi-error term had longer tails. Moreover, the loss of efficiency for normal error terms was 5%. In addition, it was confirmed that the estimation is not biased compared to the true value of *r*.

Table 3. Computational efficiency of the estimator B

Repeat counts	Degree of freedom (t-distribution)					
	1	**2**	**3**	**5**	**10**	**Inf.**
2	22006	70557	90232	98180	99818	99995
3	75142	29435	9768	1820	182	5
4	2852	8	0	0	0	0
Total	100000	100000	100000	100000	100000	100000

Table 4. RMSE and relative efficiency

Estimator	Degree of freedom (t-distribution)					
	1	**2**	**3**	**5**	**10**	**Inf.**
B' (not robust)	9000000	1.46	0.28	0.16	0.12	0.10
B (robust)	0.66	0.24	0.17	0.14	0.12	0.10
Relative Efficiency	0.00	0.16	0.61	0.87	0.98	1.05

The details of the results will be presented and described in the full paper.

4. CONCLUSIONS

The proposed robustified ratio estimator broadens the conventional definition of the ratio estimator with regards to the variance of the quasi-error term in addition to effectively alleviating the influence of outliers.

The estimator of $\beta = 1/2$ was adopted to represent the major corporate accounting items of the 2016 Economic Census for Business Activity. Robust estimators usually have degraded efficiency under the condition of normal error when the original ratio estimator is the most efficient; however, the degradation of efficiency is limited to 5% for the adopted estimator. Because the surveyed data tend to have longer tails, the application of the robust estimator is expected to contribute to the accuracy of the Census results.

REFERENCES

[1] T. De Waal, J. Pannekoek, and S. Scholtus, Handbook on Statistical Data Editing and Imputation, Wiley handbooks in survey methodology. Hoboken, New Jersey: John Wiley & Sons, (2011), 244-245.

[2] W. G. Cochran, Sampling Techniques, 3rd edition, John Wiley & Sons., (1977), 158-159.

[3] P. W. Holland and R. E. Welsch, Robust Regression Using Iteratively Reweighted Least-Squares, Communications in Statistics – Theory and methods, (1977), A6(9), 813-827.