Austrian Journal of Statistics
January 2021, Volume 50, 74–87.
http://www.ajs.or.at/
doi:10.17713/ajs.v50i1.994

Robust Estimation for a Generalised Ratio Model

Kazumi Wada
NSTAC

Keiichiro Sakashita
NSTAC

Hiroe Tsubaki
ISM

Abstract

It is known that data such as business sales and household income need data transformation prior to regression estimate as the data has a homoscedastic error. However, data transformations make the estimation of mean and total unstable. Therefore, the ratio model is often used for imputation in the field of official statistics to avoid the problem.

Our study aims to robustify the estimator following the ratio model by means of M-estimation. Reformulation of the conventional ratio model with homoscedastic quasi-error term provides quasi-residuals which can be used as a measure of outlyingness as same as a linear regression model. A generalisation of the model, which accommodates varied error terms with different heteroscedasticity, is also proposed.

Functions for robustified estimators of the generalised ratio model are implemented by the iterative re-weighted least squares algorithm in R environment and illustrated using random datasets. Monte Carlo simulation confirms accuracy of the proposed estimators, as well as their computational efficiency. A comparison of the scale parameters between the average absolute deviation (AAD) and median absolute deviation (MAD) is made regarding Tukey's biweight function. The results with Huber's weight function are also provided for reference.

The proposed robust estimator of the generalised ratio model is used for imputation of major corporate accounting items of the 2016 Economic Census for Business Activity in Japan.

Keywords: ratio imputation, M-estimation, outlier, iteratively re-weighted least squares, R.

1. Introduction

Ratio imputation is a special case of regression imputation (De Waal, Pannekoek, and Scholtus (2011), pp.244–245). When there are missing values in the target variable y, the observed auxiliary variable x is used to estimate missing y values. Therefore, x must be chosen from the variables that are highly correlated with y. The imputation model is

$$y_i = \beta x_i + \epsilon_i, \tag{1}$$

where data $i = 1, \ldots, n$ of (x, y) are observed n units in the imputation class of size N. The true ratio β is obtained by $\bar{y}/\bar{x}$; however, it is usually unknown due to the existence of missing values in y. The estimated ratio

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

is used to substitute for the missing y values such that

$$\hat{y}_i = \hat{\beta} x_i.$$

The ratio model (1) is a best linear unbiased estimator (BLUE) under the following two conditions: (i) the relationship between variables y and x is a straight line through the origin and (ii) the variance of y about this line is proportional to x (Cochran (1977), pp. 158-159.). The model (1) looks like a single regression model without intercept,

$$y_i = \beta x_i + \varepsilon_i, \tag{2}$$

however; the error term of a linear regression model is $\varepsilon_i \sim N(0, \sigma^2)$ while that of the ratio model is $\epsilon_i \sim N(0, x_i\sigma^2)$.

The ratio model is useful for imputation, as it accommodates heteroscedastic data without transformation. On the other hand, the ratio model is easily affected by outliers just like regression models (e.g. Farrella and Salibian-Barrerab (2006)).

In this paper, the idea of M-estimation for regression models is briefly explained, and reformulation of the ratio model is described so that it has the homoscedastic error term as same as a regression model. Then generalisation of the ratio model, and a robustified estimator for the generalised ratio model is proposed in section 2. Estimation of the proposed model by the iteratively re-weighted least squares (IRLS) algorithm is explained in section 3. Reasons for selecting Tukey's biweight function and its tuning constant with relation to the scale parameter are also discussed in the section. R functions based on the robustified estimator are implemented and evaluated in section 4. Monte Carlo simulation is conducted with random datasets to compare their accuracy and computational efficiency with different scale parameters regarding Tukey's biweight function. The results with Huber's weight function are also provided in Appendix. Application to a real dataset is illustrated in section 5, and section 6 concludes the paper.

2. Methodology

2.1. M-estimation for regression models

A regression model,

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_i p + \varepsilon_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \tag{3}$$

has a homoscedastic error term ε_i, which is assumed to be normal with a mean of 0 and constant variance, $V(\varepsilon_i) = \sigma^2$, where $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ip})^\top$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^\top$, and p is number of explanatory variables. The estimation equation of $\boldsymbol{\beta}$ can be expressed as $\sum_{i=1}^n (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})\mathbf{x}_i = 0$ (e.g. Huber and Ronchetti (2009), p. 155).

Huber (1973) extended his idea of M-estimation for a location parameter (Huber 1964) to the case of linear regression. The proposed estimation equation is

$$\sum_{i=1}^n w_i (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})\mathbf{x}_i = 0,$$

where w_i is the weight function $w_i = w(e_i)$ based on the standardised residuals $e_i = r_i/\hat{\sigma}$. The idea of M-estimation is controlling influence of outliers by weights w_i derived by a weight function. The value of w_i, which is within the range between 0 and 1, is determined according to the magnitude of a standardised residual. A smaller weight is allocated to an outlying observation, and then the observation has less influence to the parameter estimation.

2.2. Reformulation and gerenalisation of the ratio model

The obstacle of M-estimation for the ratio model is its heteroscedastic error term ϵ_i. Because of the error term's characteristic, residuals of the ratio model cannot be used as a measure of outlyingness. Therefore, we first reformulate the conventional ratio model so that it has a homoscedastic error term like a regression model.

The error term of a regression model (3) is assumed to be normal with a mean of 0 and constant variance, which can be written as $\varepsilon_i \sim N(0, \sigma^2)$. Meanwhile, the error term ϵ_i of the ratio model (1) is proportional to $\sqrt{x}$; i.e., the variance of ϵ_i is proportional to x and can be written as $\epsilon_i \sim N(0, x\sigma^2)$. The ratio model can be expressed in the following form:

$$y_i = \beta x_i + \sqrt{x_i}\varepsilon_i, \tag{4}$$

as these two different error terms have the relationship of $\epsilon_i = \sqrt{x_i}\varepsilon_i$. We refer to ε_i in the ratio model hereafter as the quasi-error term because the true error term of the model is ϵ_i. Then, we also propose extending the model (4) to obtain an error term that is proportional to x_i^γ as follows:

$$y_i = \beta x_i + x_i^\gamma \varepsilon_i. \tag{5}$$

The corresponding estimator of the generalised ratio model is

$$\hat{\beta} = \frac{\sum_{i=1}^n y_i x_i^{1-2\gamma}}{\sum_{i=1}^n x_i^{2(1-\gamma)}}, \tag{6}$$

and its quasi-residual $\check{r}_i$,

$$\check{r}_i = \frac{y_i - \hat{\beta} x_i}{x_i^\gamma}. \tag{7}$$

The model (5) and the estimator (6) broaden the definition of the conventional ratio model. Model (5) encompasses different models according to the value of γ. A few examples are shown in Table 1. The original ratio estimator corresponds to the case B'.

Table 1: Variations in the estimator depending on γ

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

Cases A', B' and C' have different features. In this paper, we discuss about cases A' and B' in particular, since our focus is on the models with a heteroscedastic error term. Case C' is a regression model without an intercept and has a homoscedastic error.

As the ratio β of case B' is estimated by the sum of y divided by the sum of x regarding observed data, the value is mostly decided by very large-scale observations. This estimator has a relatively small variance compared with case A' inherent to its definition, even when x and y contain extreme values. One may be able to demonstrate under what conditions estimator A' has smaller sampling variance than estimator B'. On the other hand, influence of very large observations is much smaller in the case A' since the estimation is made by the mean of ratios of each observation; however, this definition makes the value of $\hat{\beta}$ relatively unstable especially when there are very small observations in x. Scatter plots of data sets following these models are shown as Figure 1 to render the differences. The size of the data sets are $n = 1000$, and the explanatory variables of these data sets follows $x \sim N(5, 1)$ and $\beta = 2$. Objective variables y are derived based on each model with normally distributed quasi error term $\varepsilon \sim N(0, 1)$.

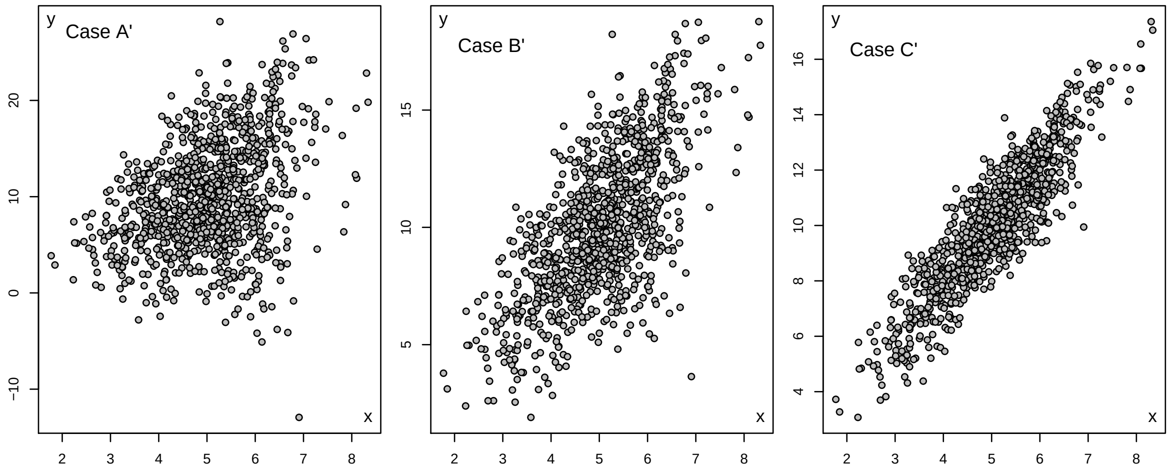

Figure 1: Random data following each model.

2.3. Robustification

The robustified estimator for the generalized ratio model (5) is derived by means of M-estimation as follows:

$$\hat{\beta}_{rob} = \frac{\sum w_i y_i x_i^{1-2\gamma}}{\sum w_i x_i^{2(1-\gamma)}}, \tag{8}$$

where w_i is a weight function of quasi residuals $\check{r}$. The role of the weight function is to alleviate influence of the observations with large residuals. There are a variety of choices in Holland and Welsch (1977) and Zhang (1997), for examples. The following two are the most popular functions among them. One is Tukey's biweight function,

$$w_i = w\left(\frac{\check{r}_i}{\hat{\sigma}}\right) = w(e_i) = \begin{cases} \left[1 - (e_i/c)^2\right]^2 & |e_i| \le c \\ 0 & |e_i| > c, \end{cases} \tag{9}$$

described in Beaton and Tukey (1974), and the other is Huber's weight function,

$$w_i = w\left(\frac{\check{r}_i}{\hat{\sigma}}\right) = w(e_i) = \begin{cases} 1 & |e_i| \le k \\ k/|e_i| & |e_i| > k, \end{cases}$$

proposed by Huber (1964). The standardised residuals e_i are quasi-residuals $\check{r}_i$ divided by an estimated scale parameter $\hat{\sigma}$. The selection of a scale parameter and tuning constants c and k are discussed in the next section. Quasi-residuals $\check{r}_i$ based on the homoscedastic quasi-error term ε_i are obtained by 7.

The cases with $\gamma = 1$, $\gamma = 1/2$ and $\gamma = 0$ are shown in Table 2. The corresponding models are similar to those for cases A', B' and C'.

Table 2: Robustified estimators.

Case	γ	Estimator	Quasi-residual
A	$\gamma = 1$	$\hat{\beta}_{robA} = \sum w_i(y_i/x_i)/\sum w_i$	$\check{r}_i = y_i/x_i - \hat{\beta}_{robA}$
B	$\gamma = 1/2$	$\hat{\beta}_{robB} = \sum w_i y_i/\sum w_i x_i$	$\check{r}_i = y_i/\sqrt{x_i} - \hat{\beta}_{robB}\sqrt{x_i}$
C	$\gamma = 0$	$\hat{\beta}_{robC} = \sum w_i y_i x_i/\sum w_i x_i^2$	$\check{r}_i = y_i - \hat{\beta}_{robC} x_i$

3. Implementation

3.1. Selection of weight function

It is important to think of the purpose of estimation and policy toward outliers for selecting a weight function. Among the two described in the previous section, we adopt Tukey's biweight function which can eliminate the influence of extreme outliers, since our purpose is imputation. The underlying policy corresponding to Tukey's biweight function is to assume that outliers are not representative for the part of the population under scrutiny. The estimation is made to complete missing data, and elimination from the estimate for imputation does not mean exclusion of the outlying observations from the survey results.

In contrast, Huber's weight function may be prefered if the purpose is a population estimate, since this function does not eliminate any observation from the estimation. Survey observations should not be wasted for the population estimates unless they are erroneous or invalid.

Figure 2 shows the difference between these weight functions. While the tails of the biweight function reach zero when the absolute value of the standardized residuals exceeds a certain threshold, the tails of Huber's weight function only approach zero at infinity.

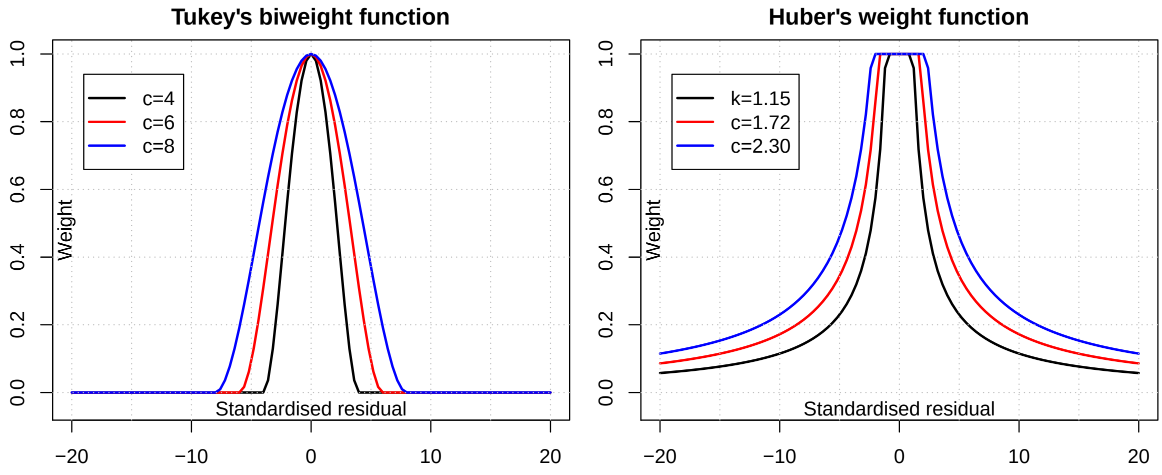

Figure 2: Features of the major weight functions.

3.2. Scale parameter and tuning constant

Beaton and Tukey (1974) use interquartile range as the scale parameter for Tukey's biweight function. Bienias, Lassman, Scheleur, and Hogan (1997) adopts average absolute deviation (AAD)

$$\sigma_{\text{AAD}} = \frac{1}{n}\sum_{i=1}^{n} |\check{r}_i|, \tag{10}$$

instead and recommends the range of tuning constant c from 4 to 8. As for Huber's weight function, Huber (1964) adopts median absolute deviation (MAD)

$$\sigma_{\text{MAD}} = \text{median}(|\check{r}_i - \text{median}(\check{r}_i)|).$$

Holland and Welsch (1977), which proposes the IRLS algorithm, provides tuning constants for 95% asymptotic efficiency under the standard normal distribution for both of these functions with σ_{MAD} as their scale parameter.

It is known that the scale parameters based on standard deviation σ_{SD} and σ_{AAD} have the relation

$$\sigma_{\mathrm{AAD}} = \sqrt{2/\pi} \cdot \sigma_{\mathrm{SD}} \approx 0.80 \cdot \sigma_{\mathrm{SD}},$$

since

$$\frac{\sigma_{\mathrm{AAD}}}{\sigma_{\mathrm{SD}}} = \frac{\mathrm{E}|z|}{\sqrt{\mathrm{E}(z^2)}} = \sqrt{\frac{2}{\pi}}.$$

Similarly,

$$\sigma_{\mathrm{MAD}} = \Phi\left(3/4\right) \cdot \sigma_{\mathrm{SD}} \approx 0.67 \cdot \sigma_{\mathrm{SD}},$$

since

$$\frac{\sigma_{\mathrm{MAD}}}{\sigma_{\mathrm{SD}}} = \frac{1}{\Phi(3/4)},$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

The tuning constants of these three scale parameters for 95% asymptotic efficiency under the standard normal distribution are obtained by Holland and Welsch (1977). Based on these figures, the range of the tuning constant k for Huber weight with σ_{MAD} corresponding to Tukey's c proposed by Bienias *et al.* (1997) are derived as in Table 3.

Table 3: Tuning constants for 95% asymptotic efficiency.

Tuning	95% asymptotic efficiency			Range of tuning constant		
constant	σ_{SD}	σ_{MAD}	σ_{AAD}	for σ_{AAD}		
c for Tukey	4.685	3.160	3.738	4	6	8
k for Huber	1.345	0.907	1.073	1.15	1.72	2.30

* The figures first appeared in Wada (2012), then those of σ_{SD} are corrected in Wada and Noro (2019).

Wada and Noro (2019) made a comparison between Tukey's biweight function and Huber's weight function with σ_{MAD} and σ_{AAD} for a simple regression model by Monte Carlo experiments with random error terms following various t-distributions. The results indicate that Tukey's biweight is more compatible with σ_{AAD}, while Huber's weight function is better with σ_{MAD}. For the tuning constants, a larger value for Tukeys' with σ_{AAD} is recommended and a smaller value for Hubers' with σ_{MAD}. Smaller values of these tuning constants make the estimation more robust but reduce weights and efficiency.

3.3. The algorithm and other settings

Among a few well-known iterative schemes for obtaining M-estimators in regression, which include Newton's method, we adopt the iteratively re-weighted least squares (IRLS) algorithm according to Holland and Welsch (1977). For the weight function and scale parameter, we choose Tukey's biweight function with σ_{AAD} in accordance with Bienias *et al.* (1997), as well as the convergence condition. Our choice of the tuning constant is $c = 8$ to minimize the weight and efficiency reduction.

The modified IRLS algorithm for a robust estimator of the generalised ratio model is as follows. The superscript index in parentheses (j) on each variable shows the iteration number.

i) Compute initial estimator $\hat{\beta}^{(1)}$ by (6).

ii) Obtain quasi-residuals $\breve{r}_i^{(1)}$ by (7), the scale parameter $\sigma^{(1)}$ by (10), and initial weights $w_i^{(1)}$ based on (9) using the predetermined tuning constant $c = 8$.

iii) Compute the ratio estimator $\hat{\beta}^{(2)}$ of $\hat{\beta}_{rob}$ according to (8) using $w_i^{(1)}$.

iv) Obtain new quasi-residuals $\check{r}_i^{(2)}$ by (7) corresponding to $\hat{\beta}^{(2)}$, update the scale parameter $\sigma^{(2)}$ by (10), and the weights $w_i^{(2)}$ by (9).

v) If the scale parameters $\sigma^{(1)}$ and $\sigma^{(2)}$ satisfy the convergence condition

$$\left|1 - \frac{\sigma^{(j)}}{\sigma^{(j-1)}}\right| < 0.001,$$

then make $\hat{\beta}^{(2)}$ the final estimator $\hat{\beta}_{rob}$ and stop iteration. Otherwise increment index j by 1 and go back to iii).

R functions shown in Table 4 are implemented together with a parent function named REGRM, which calls an appropriate child function according to the arguments. An R package containing all those functions is created and stored at the repository, `https://github.com/kazwd2008/REGRM`.

Table 4: List of the functions implemented.

Tukey's biweight function				Huber's weight function			
Name of function	Model	γ	Scale	Name of function	Model	γ	Scale
RrTa.aad	A	1	AAD	RrHa.aad	A	1	
RrTb.aad	B	1/2	AAD	RrHb.aad	B	1/2	
RrTc.aad	C	0	AAD	RrHc.aad	C	0	
RrTa.mad	A	1	MAD	RrHa.mad	A	1	
RrTb.mad	B	1/2	MAD	RrHb.mad	B	1/2	
RrTc.mad	C	0	MAD	RrHc.mad	C	0	

3.4. Illustration

To see the effect of robustification, experiments are conducted with 30% contaminated two-variable datasets as shown in Figure 3 and 4. One dataset is size 1000 and the other, 15. The 70% normal data follow a normal distribution with correlation of 0.6. Outliers, which follow a normal distribution without correlation, are placed with some distance away from normal data. The results of estimator A and B are shown together with the experiments of estimator A' and B' with all data, normal data (excluding outliers) and outliers (excluding normal data).

The experiments reveal estimator B' is affected more than A' by outliers with larger values, while estimator A and B are successfully provide similar results of B' with normal data.

4. Evaluation of the proposed estimators

Monte Carlo simulation with random data is performed based on the models A', B', and C' shown in Table 2. In the simulation, variable x is uniformly distributed random numbers from 1 to 10 and the ratio is $\beta = 5$. The quasi-error term ε_i is also random following a t-distribution with degrees of freedom $1, 2, 3, 5, 10$, and infinity. The objective variable y is calculated based on each model equation of Table 2 using the above-mentioned components. The experiments to estimate $\hat{y}$ by the estimators A, B, and C with σ_{AAD} and σ_{MAD} are performed $k = 100,000$ times with the size $n = 100$ by each degree of freedom of the t-distribution for the quasi-error term. Hereafter, we describe the estimators A, B, and C with σ_{AAD} as $\text{A}_{\text{AAD}}, \text{B}_{\text{AAD}}$, and C_{AAD}, and those with σ_{MAD} as $\text{A}_{\text{MAD}}, \text{B}_{\text{MAD}}$, and C_{MAD}.

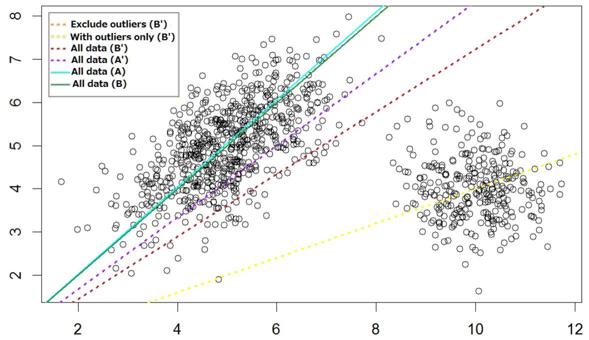

Figure 3: Experiment 1 ($n = 1000$).

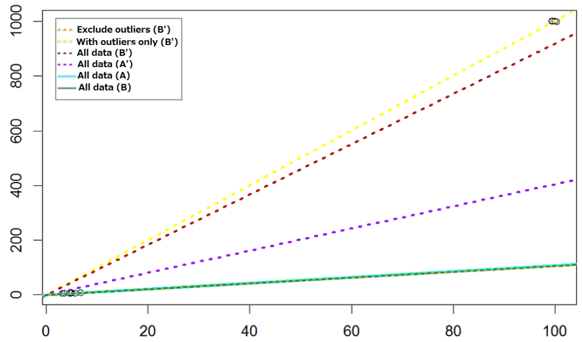

Figure 4: Experiment 2 ($n = 15$).

4.1. Accuracy

For the above-mentioned data, observed values of y contain errors, while true values are known and derived by $y = \beta x$. Comparisons among the three sets of the estimators, i.e., $(\mathrm{A}', \mathrm{A}_{\mathrm{AAD}}, \mathrm{A}_{\mathrm{MAD}})$, $(\mathrm{B}', \mathrm{B}_{\mathrm{AAD}}, \mathrm{B}_{\mathrm{MAD}})$, and $(\mathrm{C}', \mathrm{C}_{\mathrm{AAD}}, \mathrm{C}_{\mathrm{MAD}})$, are made based on $\mathrm{RMSE}_{\mathrm{A}}, \mathrm{RMSE}_{\mathrm{B}}$, and $\mathrm{RMSE}_{\mathrm{C}}$, respectively. They are defined as follows:

$$\begin{aligned}
\mathrm{RMSE}_{\mathrm{A}} &= \frac{1}{k}\sum_{j=1}^{k}\sqrt{\frac{\sum_{i=1}^{n}\left\{\left(\hat{\beta}x_{ij}-y_{ij}\right)/x_{ij}\right\}^2}{n}},\\
\mathrm{RMSE}_{\mathrm{B}} &= \frac{1}{k}\sum_{j=1}^{k}\sqrt{\frac{\sum_{i=1}^{n}\left\{\left(\hat{\beta}x_{ij}-y_{ij}\right)/\sqrt{x_{ij}}\right\}^2}{n}},\\
\mathrm{RMSE}_{\mathrm{C}} &= \frac{1}{k}\sum_{j=1}^{k}\sqrt{\frac{\sum_{i=1}^{n}\left\{\left(\hat{\beta}x_{ij}-y_{ij}\right)\right\}^2}{n}}.
\end{aligned}$$

Then relative efficiency is calculated by dividing the RMSE values of $\mathrm{A}_{\mathrm{AAD}}$ and $\mathrm{A}_{\mathrm{MAD}}$ by that of A', those of $\mathrm{B}_{\mathrm{AAD}}$ and $\mathrm{B}_{\mathrm{MAD}}$ by B', and those of $\mathrm{C}_{\mathrm{AAD}}$ and $\mathrm{C}_{\mathrm{MAD}}$ of C', to see the improvement by the robustification. Results are shown as Table 5.

The set of $100,000$ experiments was repeated four times regarding each degree of freedom for the quasi-error term and the figures in table 5 is one of them. Superior figures are underlined in the table after comparing estimators of a same model with different scale parameter. The pairs with no underline show disagreement in results among the four set of experiments. Those figures may not be stable because of the mismatched models (e.g., data following model A with estimators other than A) or datasets of df=1 which may contain extreme outliers.

Nevertheless, the robust estimators are better alternatives than the non-robust one when the data fit the model of the estimator. In addition, σ_{AAD} is a better choice than σ_{MAD} with Tukey's biweight function, unless the data are highly contaminated with extreme outliers or in the ideal situation that the quasi-error term follows the normal distribution.

4.2. Computational efficiency

The maximum number of iteration needed to compute the estimators and mean number of iteration are shown in Table 6 and 7, respectively. At least two iterations are necessary because of the algorithm shown in the previous section, and in most cases investigated here, less than 10 iterations were necessary. Number of iteration tends to increase as the tails of quasi-error term longer.

Table 5: Relative efficiency of the robustified estimators.

Data	Estimator	df=1	df=2	df=3	df=5	df=10	df=Inf.
A	$\mathrm{A}_{\mathrm{AAD}}$	0.0188	$\underline{0.4395}$	$\underline{0.6291}$	$\underline{0.7264}$	$\underline{0.7693}$	0.7886
	$\mathrm{A}_{\mathrm{MAD}}$	0.0634	0.4560	0.6511	0.7419	0.7754	$\underline{0.7850}$
B	$\mathrm{A}_{\mathrm{AAD}}$	0.0208	$\underline{0.5195}$	$\underline{0.7630}$	$\underline{0.8983}$	$\underline{0.9581}$	$\underline{0.9929}$
	$\mathrm{A}_{\mathrm{MAD}}$	0.0131	0.5240	0.7840	0.9172	0.9687	0.9932
C	$\mathrm{A}_{\mathrm{AAD}}$	0.0239	0.6080	$\underline{0.9181}$	$\underline{1.1069}$	$\underline{1.1935}$	$\underline{1.2524}$
	$\mathrm{A}_{\mathrm{MAD}}$	0.0148	0.6013	0.9296	1.1207	1.2022	1.2533
A	$\mathrm{B}_{\mathrm{AAD}}$	0.0418	$\underline{0.5847}$	$\underline{0.8331}$	$\underline{0.9708}$	$\underline{1.0345}$	1.0734
	$\mathrm{B}_{\mathrm{MAD}}$	0.1466	0.6030	0.8638	1.0012	1.0590	1.0903
B	$\mathrm{B}_{\mathrm{AAD}}$	0.0335	$\underline{0.5397}$	$\underline{0.7859}$	$\underline{0.9219}$	$\underline{0.9759}$	1.0067
	$\mathrm{B}_{\mathrm{MAD}}$	0.0213	0.5495	0.8120	0.9418	0.9837	$\underline{1.0017}$
C	$\mathrm{B}_{\mathrm{AAD}}$	0.0274	$\underline{0.4588}$	$\underline{0.6803}$	$\underline{0.8072}$	$\underline{0.8597}$	0.8920
	$\mathrm{B}_{\mathrm{MAD}}$	0.0173	0.4622	0.6975	0.8224	0.8666	0.8884
A	$\mathrm{C}_{\mathrm{AAD}}$	0.0389	0.7984	$\underline{1.1462}$	$\underline{1.3369}$	$\underline{1.4289}$	$\underline{1.4988}$
	$\mathrm{C}_{\mathrm{MAD}}$	0.1408	$\underline{0.7981}$	1.1651	1.3755	1.4779	1.5576
B	$\mathrm{C}_{\mathrm{AAD}}$	0.0248	$\underline{0.5333}$	$\underline{0.7749}$	$\underline{0.8965}$	$\underline{0.9479}$	0.9733
	$\mathrm{C}_{\mathrm{MAD}}$	$\underline{0.0152}$	0.5372	0.7969	0.9194	0.9640	0.9827
C	$\mathrm{C}_{\mathrm{AAD}}$	0.0146	$\underline{0.2975}$	$\underline{0.4193}$	$\underline{0.4830}$	$\underline{0.5117}$	0.5226
	$\mathrm{C}_{\mathrm{MAD}}$	$\underline{0.0092}$	0.3030	0.4323	0.4932	0.5157	$\underline{0.5201}$

Table 6: Maximum number of iterations.

Data	Scale	Estimator	df=1	df=2	df=3	df=5	df=10	df=Inf
A	AAD	A	4	4	3	3	3	3
		B	4	4	3	3	3	3
		C	5	4	4	4	3	4
	MAD	A	2	2	2	2	2	2
		B	7	5	5	4	4	4
		C	10	8	7	7	6	6
B	AAD	A	4	4	3	3	3	3
		B	4	4	3	3	3	3
		C	4	4	4	3	3	3
	MAD	A	2	2	2	2	2	2
		B	7	5	5	4	3	3
		C	10	8	7	5	4	4
C	AAD	A	4	4	4	4	4	4
		B	4	4	3	3	3	3
		C	4	4	3	3	3	3
	MAD	A	2	2	2	2	2	2
		B	7	5	4	4	3	3
		C	8	6	5	4	4	3

4.3. Other aspects

The problem of nonconvergence of Tukey's biweight function is reported by Wada (2012) and Wada and Noro (2019) in the context of estimating a linear model. We did not encounter the same problem, since the models used in this paper have only one parameter to estimate. Another problem reported by Wada and Noro (2019) is the phenomenon that all the robust weights w_i reach zero during computation of a robust estimator of Tukey's biweight function with the MAD scale. It occurred in our experiments of the estimator A with the MAD scale

Table 7: Mean number of iterations.

Data	Scale	Estimator	df=1	df=2	df=3	df=5	df=10	df=Inf
A	AAD	A	2.8084	2.2946	2.0988	2.0171	2.0017	2.0001
		B	2.8142	2.3248	2.1353	2.0386	2.0099	2.0018
		C	2.8225	2.3717	2.1944	2.0861	2.0387	2.0160
	MAD	A	2.0000	2.0000	2.0000	2.0000	2.0000	2.0000
		B	3.4186	2.6419	2.3840	2.1896	2.0768	2.0231
		C	3.8091	3.0400	2.7608	2.5365	2.3833	2.2635
B	AAD	A	2.8239	2.3357	2.1431	2.0470	2.0151	2.0047
		B	2.7986	2.2853	2.0971	2.0182	2.0019	2.0001
		C	2.7972	2.3143	2.1359	2.0430	2.0119	2.0024
	MAD	A	2.0000	2.0000	2.0000	2.0000	2.0000	2.0000
		B	3.3215	2.5485	2.2880	2.0991	2.0219	2.0022
		C	3.6258	2.8320	2.5375	2.3109	2.1685	2.0788
C	AAD	A	2.8644	2.4659	2.3065	2.2109	2.1588	2.1234
		B	2.7969	2.2910	2.0968	2.0181	2.0028	2.0004
		C	2.7754	2.2707	2.0916	2.0174	2.0018	2.0002
	MAD	A	2.0000	2.0000	2.0000	2.0000	2.0000	2.0000
		B	3.2726	2.5268	2.2928	2.1204	2.0420	2.0110
		C	3.4828	2.6912	2.3977	2.1763	2.0578	2.0123

for the data of df=1 and 2 as shown in Table 8.

Although this study does not focus on the comparison of the two weight functions, resuls of the estimators of Huber's weight function with the same datasets are shown in Appendix.

Table 8: Number of aborts due to all weights being zero (in $100,000$ experiments).

Data	Other conditions	df=1	df=2
A	Estimator A with MAD scale	1622	3
B		1616	8
C		1889	2

5. Application with real data

The robust estimators for the generalized ratio model proposed 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 at the national and regional levels, and to obtain basic information to conduct various statistical surveys by investigating the economic activities 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 was a leading candidate at the beginning, it is well known that the estimation is very sensitive to outliers; therefore, we needed to take appropriate measures for the problem.

After implemented the functions based on the robust estimator of the generalised ratio model, estimator A and B were compared using previous census data by Monte Carlo simulation. Estimator B was selected to estimate missing sales by expenses, salaries by expenses, and expenses by sales.

The estimator B has a problem to be influenced by extremely large observations in spite of the robustification. Such observations are not regarded as outliers even when they have different tendencies compared to the majority of other observations. Figure 5 shows a good example.

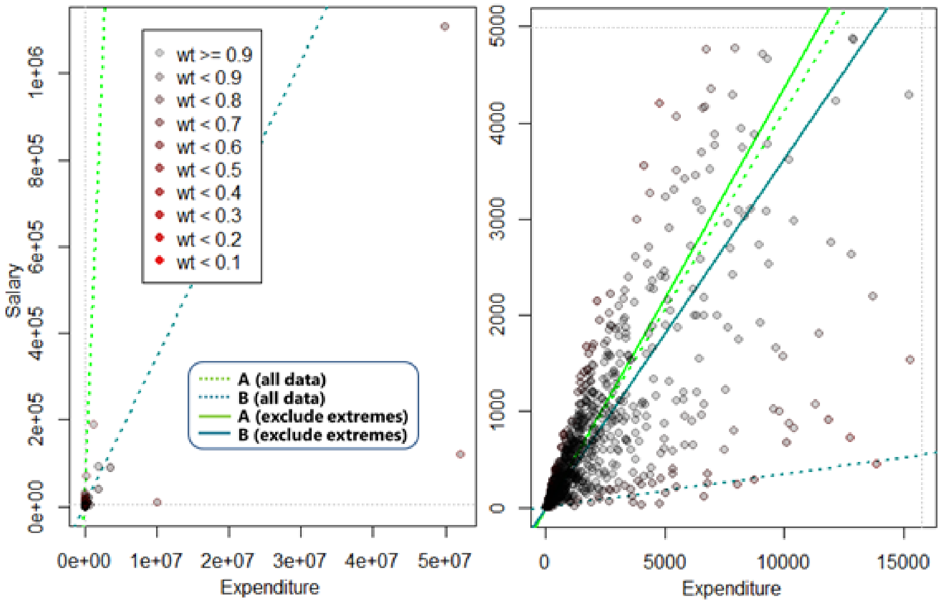

Figure 5: Industry 55A: Agents and brokers.

The left side of the scatter plot shows the whole enterprises in the industry together with the results of estimator A and B. There are a few extremely large values, and estimator B is affected by them seriously. The right side plot closes-up the smaller observations with higher density in the same dataset. To cope with the problem, extremely large outliers if any in each imputation class were removed from estimation in the course of data processing for the 2016 Census.

6. Conclusions

The proposed generalised ratio model broadens the conventional definition of the ratio model with regards to the variance of the error term. Robustified estimators based on the model effectively alleviate the influence of outliers.

Application of the robust estimator may contribute to the accuracy of official statistics, as the survey data tend to have longer tails. The R functions based on the proposed estimator are implemented, evaluated and provided at a public repository.

Users' policy toward outliers may reflect the choices of a weight function, and a suitable scale parameter for Tukey's biweight function is AAD regarding longer tailed datasets.

As for the conventional ratio model, the robustified estimator is highly affected by very large observations; therefore, removing extremely large outliers may necessary before estimation.

7. Future work

Simulations in this paper are with uniformally distributed x values to have fatter tails than normal distribution. However, further simulation may also be needed for economic data with lognormally distributed x, since it would be more realistic.

Another interesting topic is an application for restricted data. Economic surveys gather multiple financial variables for each establishment, for an example. Those variables may contain some restrictions such as a total and its components. Further study is necessary.

Appendix

The results of Monte Carlo simulation for the estimators with Huber's weight function are shown in this appendix. The datasets used is identical with thosed used in section 3. The table 9, 10 and 11 are comparable with Table 5, 6 and 7.

Table 9: Relative efficiency of the robustified estimators with Huber weight.

Data	Estimator	df=1	df=2	df=3	df=5	df=10	df=Inf.
A	A_{AAD}	0.0380	0.4636	0.6424	0.7305	0.7709	0.7895
	A_{MAD}	0.0331	0.4767	0.6657	0.7496	0.7796	0.7845
B	A_{AAD}	0.0716	0.6209	0.8580	0.9832	1.0404	1.0753
	A_{MAD}	0.0679	0.6332	0.8883	1.0170	1.0711	1.0991
C	A_{AAD}	0.0621	0.8704	1.2142	1.3962	1.4792	1.5307
	A_{MAD}	0.0614	0.8580	1.2268	1.4295	1.5252	1.5857
A	B_{AAD}	0.0440	0.5488	0.7808	0.9052	0.9610	0.9956
	B_{MAD}	0.0151	0.5492	0.8019	0.9272	0.9755	0.9975
B	B_{AAD}	0.0561	0.5681	0.8032	0.9273	0.9780	1.0077
	B_{MAD}	0.0246	0.5749	0.8311	0.9518	0.9888	1.0007
C	B_{AAD}	0.0478	0.5635	0.7940	0.9103	0.9536	0.9759
	B_{MAD}	0.0176	0.5631	0.8173	0.9364	0.9746	0.9893
A	C_{AAD}	0.0520	0.6438	0.9424	1.1174	1.1957	1.2530
	C_{MAD}	0.0170	0.6291	0.9495	1.1305	1.2075	1.2586
B	C_{AAD}	0.0447	0.4835	0.6953	0.8125	0.8613	0.8937
	C_{MAD}	0.0199	0.4833	0.7128	0.8299	0.8710	0.8904
C	C_{AAD}	0.0377	0.3120	0.4285	0.4884	0.5125	0.5279
	C_{MAD}	0.0136	0.3154	0.4430	0.5011	0.5184	0.5248

Table 10: Maximum number of iterations.

Data	Scale	Estimator	df=1	df=2	df=3	df=5	df=10	df=Inf
A	AAD	A	4	4	3	3	3	3
		B	4	4	3	3	3	3
		C	5	4	4	3	3	3
	MAD	A	2	2	2	2	2	2
		B	7	5	4	4	4	3
		C	9	7	6	5	5	5
B	AAD	A	4	4	4	3	3	3
		B	4	4	3	3	3	3
		C	4	4	4	3	3	3
	MAD	A	2	2	2	2	2	2
		B	6	5	4	3	3	3
		C	8	6	5	5	4	4
C	AAD	A	4	4	4	3	3	3
		B	4	4	3	3	3	3
		C	4	4	3	3	3	3
	MAD	A	2	2	2	2	2	2
		B	6	5	4	4	3	3
		C	9	5	4	4	4	3

Table 11: Mean number of iterations.

Data	Scale	Estimator	df=1	df=2	df=3	df=5	df=10	df=Inf
A	AAD	A	2.8559	2.2759	2.0833	2.0116	2.0010	2.0000
		B	2.8632	2.3050	2.1171	2.0298	2.0072	2.0015
		C	2.8748	2.3516	2.1719	2.0696	2.0293	2.0129
	MAD	A	2.0000	2.0000	2.0000	2.0000	2.0000	2.0000
		B	3.3292	2.5609	2.3421	2.1795	2.0795	2.0238
		C	3.7026	2.8954	2.6514	2.4817	2.3707	2.2755
B	AAD	A	2.8751	2.3166	2.1236	2.0358	2.0106	2.0030
		B	2.8417	2.2664	2.0818	2.0128	2.0012	2.0001
		C	2.8430	2.2959	2.1172	2.0330	2.0088	2.0021
	MAD	A	2.0000	2.0000	2.0000	2.0000	2.0000	2.0000
		B	3.2442	2.4850	2.2578	2.0927	2.0212	2.0015
		C	3.5348	2.7239	2.4842	2.2987	2.1712	2.0741
C	AAD	A	2.9243	2.4434	2.2821	2.1843	2.1341	2.1009
		B	2.8326	2.2697	2.0820	2.0126	2.0014	2.0002
		C	2.8077	2.2515	2.0779	2.0128	2.0014	2.0001
	MAD	A	2.0000	2.0000	2.0000	2.0000	2.0000	2.0000
		B	3.1959	2.4693	2.2532	2.1042	2.0352	2.0090
		C	3.4034	2.6122	2.3641	2.1690	2.0540	2.0067

When looking at the cases in those models of the data and the estimators are consistent, results of the simulation indicate Tukey's biweight function is slightly more efficient than Hubers', while Hubers' converges a little bit faster, except for the unstable results with the data of $df = 1$. However, those difference may be negligibly small in practical use.

References

Beaton AE, Tukey JW (1974). "The Fitting of Power Series, Meaning Polynomials, Illustrated on Bandspectroscopic Data." *Technometrics*, **16**, 147–185.

Bienias JL, Lassman DM, Scheleur SA, Hogan H (1997). "Improving Outlier Detection in Two Establishment Surveys." In *Statistical Data Editing Volume No.2 Methods and Techniques*, pp. 76–83. UNSC/UNECE.

Cochran WG (1977). *Sampling Techniques, 3rd ed.* John Wiley & Sons.

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

Farrella PJ, Salibian-Barrerab M (2006). "A Comparison of Several Robust Estimators for a Finite Population Mean." *Journal of Statistical Studies*, **26**, 29–43.

Holland PW, Welsch RE (1977). "Robust Regression Using Iteratively Reweighted Least-squares." *Communications in Statistics-theory and Methods*, **6**(9), 813–827.

Huber PJ (1964). "Robust Estimation of a Location Parameter." *The Annals of Mathematical Statistics*, **35**, 73–101.

Huber PJ (1973). "Robust Regression: Asymptotics, Conjectures and Monte Carlo." *The Annals of Statistics*, **1**(5), 799–821.

Huber PJ, Ronchetti EM (2009). *Robust Statistics, 2nd ed.* John Wiley & Sons.

Wada K (2012). "Detection of Multivariate Outliers — Regression Imputation by the Iteratively Reweighted Least Squares — (in Japanese)." *Research Memoir of Official Statistics*, **69**, 23–52. URL https://www.stat.go.jp/training/2kenkyu/ihou/69/pdf/2-2-692.pdf.

Wada K, Noro T (2019). "Consideration on the Influence of Weight Functions and the Scale for Robust Regression Estimator (in Japanese)." *Research Memoir of Official Statistics*, **76**, 101–114. URL https://www.stat.go.jp/training/2kenkyu/ihou/76/pdf/2-2-767.pdf.

Zhang Z (1997). "Parameter Estimation Techniques: A Tutorial with Application to Conic Fitting." *Image and vision Computing*, **15**(1), 59–76. URL https://hal.inria.fr/file/index/docid/74015/filename/RR-2676.pdf.

Affiliation:

Kazumi Wada
National Statistics Center (NSTAC)
19-1, Wakamatsu-cho, Shinjuku-ku, Tokyo 162-8668, Japan
E-mail: kwada@nstac.go.jp

Keiichiro Sakashita
National Statistics Center (NSTAC)
19-1, Wakamatsu-cho, Shinjuku-ku, Tokyo 162-8668, Japan
E-mail: ksakashita@nstac.go.jp

Hiroe Tsubaki
The Institue of Statistical Mathematics (ISM)
10-3, Midori-cho, Tachikawa, Tokyo 190-8562, Japan
E-mail: tsubaki@ism.ac.jp

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 50 *Submitted:* 2019-07-13
January 2021 *Accepted:* 2020-02-11