An Algorithm of Generalized Robust Ratio Model Estimation for Imputation

WADA, Kazumi * TAKATA, Seiji † TSUBAKI, Hiroe ‡

Abstract

This paper proposes an algorithm of simultaneous robust estimation for the generalized ratio model proposed by Wada and Sakashita (2017) with an implemented R function. It helps to abbreviate the model selection process prior to imputation in the course of survey data processing.

Wada and Sakashita (2017) robustify the ratio model by introducing homoscedastic quasi-error term to determine robust weights for each observation based on the idea of M-estimation. They also extended the ratio model so that the errors are proportional to the explanatory variable to a different powers. The algorithm we propose is to estimate the power of the explanatory variable together with the ratio of objective variables robustly.

The estimate of power may not be very accurate as with the weighted two-stage least squares; however, the accuracy of ratio matters for imputation, since the value of the power is not used for estimation of the objective variable. Therefore, the proposed algorithm could be of use as long as the estimation of the ratio has good accuracy regardless of the power.

Key Words: M-estimators, Outlier, IRLS (Iteratively Reweighted Least Squares)

1. Introduction

Imputation is an unavoidable task in survey data processing, and ratio imputation is widely used in the field of official statistics among a variety of methods. One of the reasons is the ratio model has a heteroscedastic error term. Survey variables are often heteroscedastic and need data transformation prior to a linear regression model; however, such transformation makes some important estimations such as mean and total unstable. The ratio model is capable to accommodate heteroscedastic errors proportional to the explanatory variable to the power one-half.

The conventional ratio model is,

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

regarding the explanetory variable x_i which has highly correlated with the objective variable y_i, where β is the ratio of y_i to x_i and a heteroscedastic error term $\epsilon_i \sim N(0, x_i\sigma^2)$ with a constant variance σ^2 (e.g. Cochran (1953)). A simple regression model without intercept,

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

looks identical with the model (1); however, the error term of the regression model is homoscedastic as $\epsilon_i \sim N(0, \sigma^2)$. Figure 1 shows the appearence of datasets following these two models. Their different error terms have the relation, $\epsilon_i = \varepsilon\sqrt{x_i}$.

Wada and Sakashita (2017) proposes generalization of this conventional ratio model using the relation between the above mentioned error terms, and the error term of the generalized model is $\epsilon_i \sim N(0, x_i^\gamma\sigma^2)$. They distribute the corresponding R functions for the

*National Statistics Center (NSTAC), 19-1, Wakamatsu-cho, Shinjuku-ku, Tokyo 162-8668, Japan

†Ministry of Internal Affairs and Communication (MIAC), 19-1, Wakamatsu-cho, Shinjuku-ku, Tokyo 162-8668, Japan

‡The Institute of Statistical Mathematics (ISM), 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan

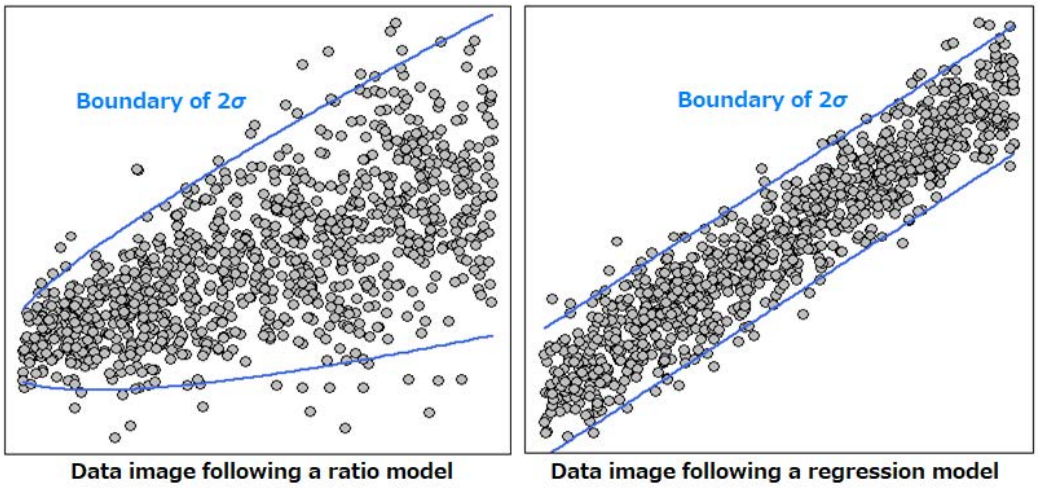

Figure 1: Data images following a ratio model and a regression model.

robustified estimators at the repository `http://github.com/kazwd2008/IRLS` as files named `RrT.r` and `RrH.r` regarding a few prescribed γ values.

The proposed robustified estimator was adopted with $\gamma = 1/2$ for imputing major corporate accounting items of the 2016 Economic Census for Business Activities in Japan. It required a model selection process regarding different γ values preceded the imputation step in the course of statistics production. The research objective of this paper is estimating γ together with β to omit the model selection step.

We propose an algorithm of simultaneous robust estimation for β and γ. The corresponding R function is distributed at the above mentioned repository as a file named `RBreds.r`. The file includes the following two functions: `RBred` is the robust version and `Bred`, non-robust version which is for comparative evaluation. We confirmed `Bred` has better performance regarding accuracy of β compared to the estimation using `optime` function in R, which is a general-purpose optimization based on Nelder-Mead algorithms. The robust version `RBred` naturally outperforms `Bred` with contaminated datasets, and would be of use for the purpose of imputation.

We first summarize the idea of Wada and Sakashita (2017) in section 1.1 and then describe about the computation in section 1.2.

1.1 Generalized Ratio Model and Its Robustified Estimator

Wada and Sakashita (2017) reformulate the ratio model (1) $y_i = \beta x_i + \sqrt{x_i}\varepsilon_i$ with a homoscedastic error term as with a regression model. It is essential to robustify the ratio model by means of M-estimation. A generalized ratio model,

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

is also proposed by them. The new homoscedastic error term ε_i in this model is called "quasi-error term," and has the relation $\epsilon_i = x_i^{\gamma}\varepsilon_i$ with the heteroscedastic error term in the conventional ratio model. The corresponding generalized ratio estimator and its homoscedastic quasi-residuals are as follows:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} y_i x_i^{1-2\gamma}}{\sum_{i=1}^{n} x_i^{2(1-\gamma)}}. \tag{4}$$

$$\check{r}_i = \frac{y_i - \hat{\beta} x_i}{x_i^{\gamma}}. \tag{5}$$

Please note the generalized estimator (4) comes down to the conventional ratio estimator,

$$\hat{\beta} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}, \tag{6}$$

which corresponds to model (1) when $\gamma = 1/2$.

The robustified version of the estimator (4) is,

$$\hat{\beta}_{rob} = \frac{\sum w_i y_i x_i^{1-2\gamma}}{\sum w_i x_i^{2(1-\gamma)}}, \tag{7}$$

where w_i is computed according to a weight function using quasi-residuals $\check{r}$ obtained by the eq. (5). Weights usually have values between 0 to 1, and a selected weight function reduces weights of outliers if any to alleviate their influence on the parameter estimation.

1.2 Computation of the Robustified Estimator

Holland and Welsch (1977) recommend the iteratively reweighted least squares (IRLS) algorithm for M-estimation proposed by Beaton and Tukey (1974), in which the algorithm is called by the name of "biweight regression fitting". Wada (2012) implemented the settings of Bienias *et al.* (1997) in UNSC/UNECE (1997), which aims to share best practices within the national statistical offices. Bienias *et al.* (1997) introduces the IRLS algorithm by the name of resistant fitting and appy average absolute deviation (AAD) for the scale parameter,

$$\sigma_{AAD} = \frac{1}{n} \Sigma_{i=1}^{n} |\check{r}_i|, \tag{8}$$

and Tukey's biweight function (Beaton and Tukey, 1974)

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

for the weight function among others, where e_i is a standardized residual using the scale parameter $\hat{\sigma}$.

Wada and Noro (2019) considers influence of the choice of weight function as well as the scale parameter, and explore the features on the setting of Bienias *et al.* (1997) further. They find the selection of Tukey's biweight function together with the AAD scale is fast to converge and can eliminate the influence of extreme outliers. The implemented R functions by Wada and Noro (2019) is placed at the repository `http://github.com/kazwd2008/IRLS` as files named Tirls.r of Tukey's biweight function and Hirls.r of Huber's weight function (Huber, 1964). The functions for robustified estimators of the generalized ratio model regarding a few prescribed γ values by Wada and Sakashita (2017) are also in the same repository as files named `RrT.r` and `RrH.r`.

The algorithm for the generalized ratio estimator with a given γ value is as follows:

1. Estimate initial parameter based on eq. (4).
2. Obtain quasi-residuals based on eq. (5) based on the latest estimation.
3. Compute scale parameter $\hat{\sigma}$ (e.g., based on eq. (8)) and then standardize the quasi-residuals by $\check{r}_i/\hat{\sigma}$ using the obtained scale parameter.

4. Calculate a robust weight w_i for each observation according to a weight function (e.g., by eq. (9)) and the standardized quasi-residuals.
5. Make robust estimation based on eq. (7) using w_i.
6. If the latest j-th scale parameter $\hat{\sigma}^{(j)}$ and the previous $(j-1)$-th $\hat{\sigma}^{(j-1)}$ satisfies the conversion condition,

$$\left|1 - \frac{\sigma^{(j)}}{\sigma^{(j-1)}}\right| < 0.001, \tag{10}$$

the latest $\hat{\beta}^{(j)}$ is the final estimator β_{rob} and the iteration is terminated. Otherwise increment index j by 1 and go back to 2.

2. Simultaneous Estimation

We first describe how to estimate γ by the two stage least squares (2SLS) and how to robustify the estimation in section (2.1). Then its robustification is discussed section (2.3). Our proposed algorithm of the simultaneous robust estimation of β and γ is shown in section (2.2). The proposed algorithm is based on the robust estimators of the generalized ratio estimator implemented by Wada and Sakashita (2017) and the 2SLS estimation discussed in section (2.2) is incorporated.

2.1 Two-Stage Least Squares Estimation

Following is the procedure to estimate γ called two-stage least squares (2SLS) estimation (e.g. Greene (2002), p.79). First step is to estimate β of the model,

$$y_i = \beta x_i + \epsilon_i, \tag{11}$$

by the ordinary least squares, where $\epsilon_i \sim N(0, \sigma^2 x_i^{\beta})$. Please note eq. (11) is another form of the model (3) expressed with heteroscedastic error $\epsilon_i = x_i^{\gamma}\varepsilon_i$. The estimates of β is not efficient but at least unbiased under heteroscedasiticity.

Next step is estimating γ using an instrumental variable $r_i^2 = (y_i - \hat{\beta}x_i)^2$, where $r_i = y_i - \beta x_i$. Taking the logarithm,

$$\log|r_i| = \gamma \log|x_i| + \log(\sigma) \tag{12}$$

is derived. It shows γ is obtained as the single regression parameter in explaining $\log|r_i|$ by $\log|x_i|$.

Let z' be an $n \times 2$ matrix where we have size n of observations on 2 variables, and u' be an $n \times 1$ vector of observations on the dependent variable. Computation of the matrix,

$$\hat{\gamma} = (z'^{\top} z')^{-1} z'^{\top} u', \tag{13}$$

where

$$z' = \begin{pmatrix} 1 & \log(x_1) \\ \vdots & \vdots \\ 1 & \log(x_n) \end{pmatrix}, \qquad u' = \begin{pmatrix} \log(|y_1 - \hat{\beta}x_1|) \\ \vdots \\ \log(|y_n - \hat{\beta}x_n|) \end{pmatrix}.$$

provide a vector with two elements and estimated γ is the second one.

2.2 Robustification of the Power Estimate

Now we consider robustification of the estimator in section 2.1 by introducing robust weight w_i. The model (11) is identical with (1) , so its robustified estimator is obtained by eq. (7) as discussed in section 1.2.

The robustified form of Eq. (13) is,

$$\gamma_{rob} = (z^\top z)^{-1} z^\top u, \tag{14}$$

where

$$z = \begin{pmatrix} 1 & \log(w_1 x_1) \\ \vdots & \vdots \\ 1 & \log(w_n x_n) \end{pmatrix}, \qquad u = \begin{pmatrix} \log(|y_1 - \hat{\beta} x_1| w_1) \\ \vdots \\ \log(|y_n - \hat{\beta} x_n| w_n) \end{pmatrix}.$$

2.3 Proposed Algorithm

Following is the proposed algorithm to estimate robust β and γ simultaneously. We incorporate robust 2SLS estimation discussed in section 2.2 into the algorithms of the robust estimators of the generalized ratio estimator proposed by Wada and Sakashita (2017) described in section 1.2.

I. Initial estimation

- (i) Initial estimation based on eq. (4) with an appropriate initial value $\gamma^{(0)}$
- (ii) Obtain conventional residuals $r_i = y_i - \hat{\beta} x_i$ as well as quasi-residuals based on eq. (5)
- (iii) Compute scale parameter $\hat{\sigma}$ (e.g., based on eq. 8) and standardize the quasi-residuals obtained by $\check{r}_i / \hat{\sigma}$

II. Iterative non-robust estimation of β and γ

- (i) Estimate γ using $\hat{\beta}$ based on eq. (13)
- (ii) Estimate β based on eq. (4) using newly estimated $\hat{\gamma}$
- (iii) Caliculate quasi-residuals based on eq. (5) and scale parameter $\hat{\sigma}$
- (iv) Go back to II (i) unless the latest σ meets convergence condition (10)

III. Iterative robust estimation of β using weights with fixed γ

- (i) Compute robust weights w_i according to a weight function based on the latest estimation of β and γ.
- (ii) Estimate robust β based on eq. (7)
- (iii) Caliculate conventional residuals $r_i = y_i - \hat{\beta} x_i$, quasi-residuals based on eq. (5) and scale parameter $\hat{\sigma}$
- (iv) Go back to III (i) unless the latest σ meets convergence condition (10)

IV. Simultaneous iterative robust estimation of β and γ

- (i) Compute robust weights w_i according to a weight function based on the latest estimation of β and γ.
- (ii) Estimate γ using latest $\hat{\beta}$ based on eq. (14)
- (iii) Estimate β based on eq. (7) using newly estimated $\hat{\gamma}$
- (iv) Calculate conventional residuals $r_i = y_i - \hat{\beta} x_i$, quasi-residuals based on eq. (5), and scale parameter $\hat{\sigma}$
- (v) Go back to IV (i) unless the latest σ meets convergence condition (10)

3. Evaluation of the Proposed Algorithm

We implemented the algorithm described in the previous section as a function `RBred` together with non-robust version `Bred` in a file named `RBreds.r` at the repository `http://github.com/kazwd2008/IRLS`. The function `Bred` comprises the part I and II of the algorithm for `RBred` and implemented for the evaluation purpose only. Please note that the non-robust estimation for the model (3) is practically useless for the purpose of imputation, because the estimation of β comes down to the results of a single regression model without intercept (2) regardless of the value of γ.

In addition to the function `Bred`, we also prepare a non-robust optimization code using R `optim` function, which is a general-purpose optimization based on Nelder-Mead algorithm. The estimation of β and γ can be defined a minimization problem as follows:

$$\arg\min_{\beta,\gamma} \left(\frac{y_i - \beta x_i}{x_i^{\gamma}} \right)^2 .$$

The R code for estimation is as follows.

```
Op1 <- function(x, y, pm) (sum((y - pm[1] * x) / x^pm[2])^2)
optim(pm, Op1, x=x, y=y)
```

3.1 Random Datasets Without Outliers

A comparison is made for the estimation with `optim`, `Bred`, and `RBred` with the following five different datasets without outliers.

The size of tested datasets is 200. The explanatory variable is $x \sim N(100, 1)$, the quasi error term $\varepsilon \sim N(0, 0, 2)$, and the objective variable y is calculated based on the following models using the values of x and ε.

(1) $y = 2x + \varepsilon$
(2) $y = 2x + \varepsilon x^{0.25}$
(3) $y = 2x + \varepsilon x^{0.5}$
(4) $y = 2x + \varepsilon x^{0.75}$
(5) $y = 2x + \varepsilon x$

The code of `optim` requires initial values of β and γ as two elements in vector `pm`. The initial value of β is $(0, 0.25, 0.5, \ldots, 1.25)$ while initial γ is $(0, 0.5, 1, \ldots, 5)$. As a total, 66 combinations of those initial values are tested for each dataset.

Function $Bred$ and $RBred$ only need an initial value for γ, so the values, $(0, 0.5, 1, \ldots, 5)$ are tested for each dataset.

The results are shown in Figure 2. As for `optim`, estimated values fractuate depending on the initial values, although all the estimation are converged. On the other hand, both `Bred` and `RBred` return same estimation regardless of the initial γ for each dataset, and the estimation of `Bred` is better than those of `RBred`.

3.2 Random Datasets With Outliers

In this section, the same dataset created in section 3.1 is used after the following modification. For each datasets, 10 observations with variable x greater than 100 are selected randomly, and those y values are multipliyed by 10. Figure 3 shows the sample of datasets with and without outliers. Atrificial outlers are shown in red color.

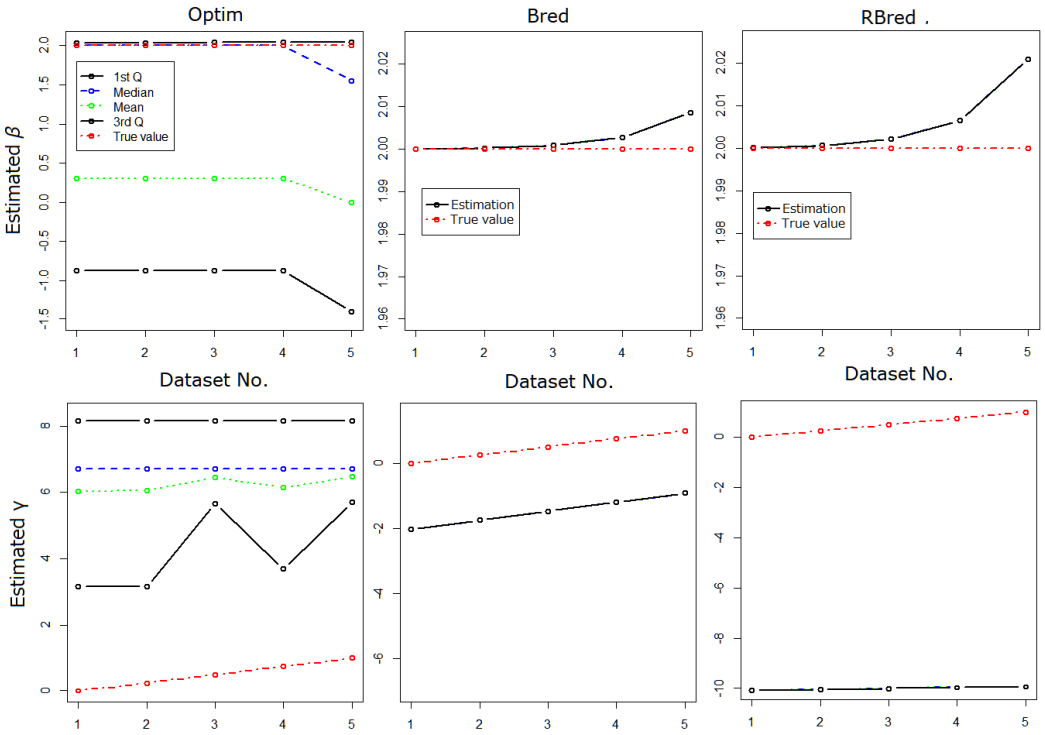

Figure 2: Estimation with different initial vallues.

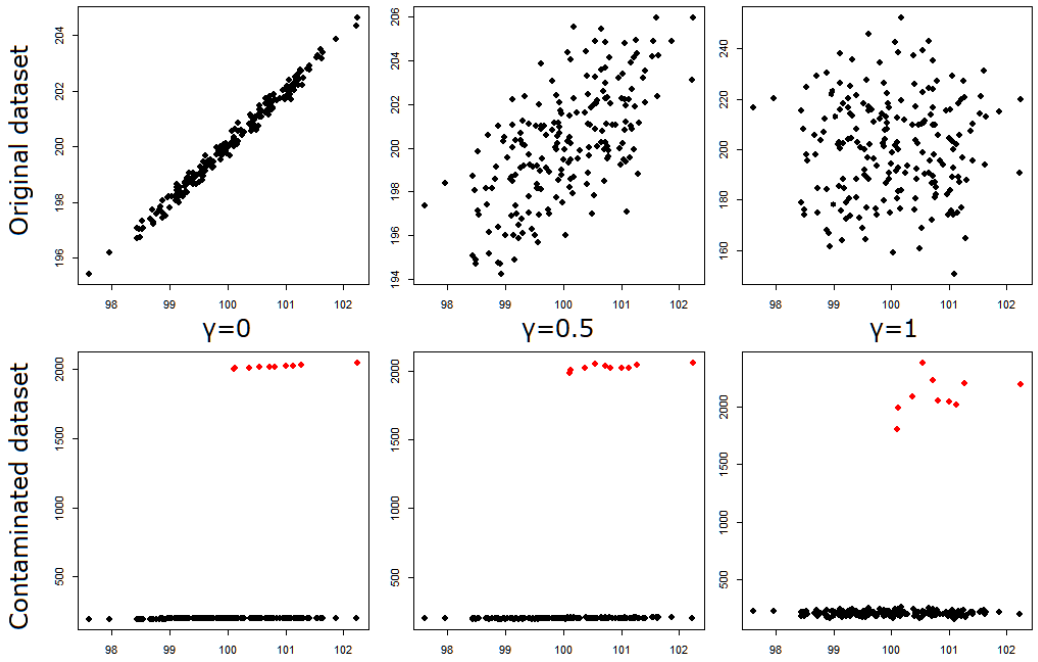

Figure 3: Tested datasets.

The results with the contaminated datasets are shown in Figure 4. The result of `Bred` is the most severely impaired among the three due to the contamination. The result of `optim` also worsen. So, the estimation of β by `RBred` is the best among them, as expected.

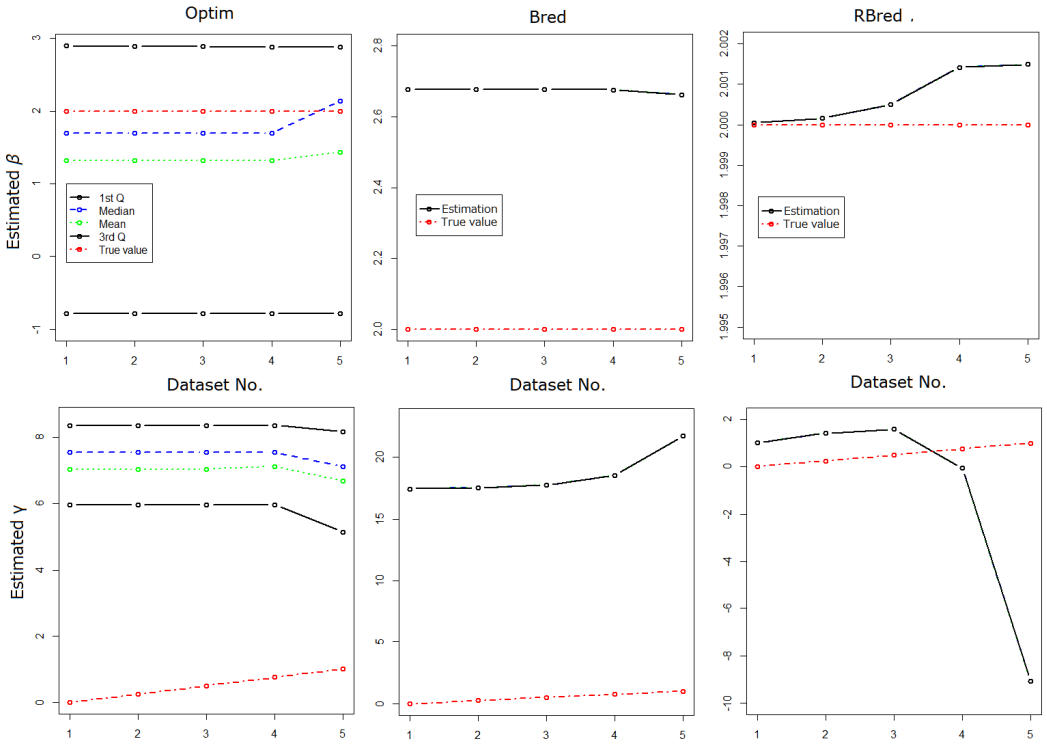

Figure 4: Tested datasets.

4. Conclusion

We proposed an algorithm for robust simultaneous estimation of β and γ regarding the generalized ratio model proposed by Wada and Sakashita (2017). A R function is also implemented and distributed at the github repository. Evaluation is made by datasets with and without contamination. The R function `RBred` has a expected feature so far; however, further evaluation is necessary toward practical use for imputation.

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 (1953). *Sampling Techniques*. New York: Wiley.

Greene WH (2002). *Econometric Analysis, 5th ed.* Prentice Hall.

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.

UNSC/UNECE (1997). *Statistical Data Editing Volume No.2, Methods and Techniques*. United Nations.

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.

Wada K, Sakashita K (2017). "Generalized Robust Ratio Estimator for Imputation." In *Proceedings of New Techniques and Technologies for Statistics (NTTS)*. Brussels, Belgium. Accessed June 2019, URL https://conference-service.com/NTTS2017/documents/agenda/data/x_abstracts/x_abstract_56.docx.