Outliers in official statistics

Kazumi Wada¹

Received: 10 January 2020 / Accepted: 19 September 2020 / Published online: 24 October 2020
© The Author(s) 2020

Abstract

The purpose of this manuscript is to provide a survey on the important methods addressing outliers while producing official statistics. Outliers are often unavoidable in survey statistics. They may reduce the information of survey datasets and distort estimation on each step of the survey statistics production process. This paper defines outliers to be focused on each production step and introduces practical methods to cope with them. The statistical production process is roughly divided into the following three steps. The first step is data cleaning, and outliers to be focused are that may contain mistakes to be corrected. Robust estimators of a mean vector and covariance matrix are introduced for the purpose. The next step is imputation. Among a variety of imputation methods, regression and ratio imputation are the subjects in this paper. Outliers to be focused on in this step are not erroneous but have extreme values that may distort parameter estimation. Robust estimators that are not affected by remaining outliers are introduced. The final step is estimation and formatting. We have to be careful about outliers that have extreme values with large design weights since they have a considerable influence on the final statistics products. Weight calibration methods controlling the influence are discussed based on the robust weights obtained in the previous imputation step. A few examples of practical application are also provided briefly, although multivariate outlier detection methods introduced in this paper are mostly in the research stage in the field of official statistics.

Keywords Outlier detection · Robust estimation · MSD · M-estimators · Weight calibration

✉ Kazumi Wada
k5.wada@soumu.go.jp

¹ Statistical Research and Training Institute, Ministry of Internal Affairs and Communications (MIC), 2-11-16 Izumi-cho, Kokubunji-shi, Tokyo 185-0024, Japan

1 Introduction

1.1 What are outliers

Outliers are extreme or atypical values that can reduce and distort the information in a dataset. The problem of how to deal with outliers has long been a concern. Barnett and Lewis (1994, p. 3), one of the pioneering books in mathematical statistics dealing with outlier detection, reference Pierce (1852) published more than 150 years ago. Eliminating outliers from estimation carries the risk of losing information, however including the risks of contamination. To deal with the problem, Barnett and Lewis (1994, p. 3) devised a principle to accommodate outliers using robust methods of inference, allowing for the use of all the data while alleviating the undue influence of outliers. We follow this principle and focus on the robust statistical methods introduced by Huber (1964) that are the most suitable for survey data processing. Therefore, statistical tests are beyond the scope of our discussion.

Some outliers in survey statistics contain a mistake of some sort that requires correction. Others may not involve a mistake, but represent a trend different from that of the majority while having a large design weight in the dataset. Careful consideration of the influence of outliers on estimation needs to be given, and statisticians compiling official statistics need to determine whether such extreme values deserve their prescribed sampling weights in terms of representativeness, as discussed by Chambers (1986).

While the UNECE Data Editing Group broadly defines outliers as observations in the tails of a distribution (Economic Commission for Europe of the United Nations 2000, p. 10), narrower definitions vary depending on the purpose of the activities in the statistical production process. Outliers require appropriate treatment at each of the processing steps; otherwise, they may negatively impact on estimation efficiency and introduce bias into the resulting statistical product. The objective of this paper is to introduce both practical methods currently in use and experimental methods in research intended for use in statistical production to address the problem of data outliers.

Most conventional outlier detection methods in the field of official statistics are univariate approaches mainly applied to the search for erroneous observations so that they can be corrected, and entirely valid datasets can be established. A range-check to determine upper and lower thresholds for “normal” (i.e., not outlying) data is a typical example, as is the quartile method. However, such univariate methods cannot detect multivariate outliers, that is, outliers involving different relationships among the variables. In multivariate cases, scatter plot matrices or other visualization techniques have been frequently used compared to the multivariate methods because of their computational complexity and processing time or the difficulty associated with the inspection of detected multivariate outliers. Complicating also matters, with multivariate methods, just which outliers are detected can depend on which particular method is being used.

Historically, statistical tables have been the major final product of official statistics, which means that the demand for detecting multivariate outliers having different relationships among the variables has not been high. However, in 2007, the Statistics Act of Japan (Act No. 53) was revised for the first time in 60 years. The new act recognizes official statistics as an information infrastructure and promotes the use of microdata (e.g., Nakamura 2017). Given this change in policy, the need to detect multivariate outliers has increased since outliers tend to be more problematic in microdata, not only for users but also for providers, in terms of privacy protection. Besides, the practical usability of multivariate outlier detection methods is increasing with the continuing improvements being made in computer technologies both in hardware and statistical software.

In the next subsection, a general model of the statistical production process is described. The model consists of three steps: data cleaning, imputation, and estimation and formatting. The outliers to be focused on depend on the purpose of each step.

In Sect. 2, available multivariate outlier detection methods for the data cleaning step are discussed. Section 3 describes robust regression for imputation. M-estimators discussed in Sect. 3 are then extended to the ratio model in Sect. 4. Calibration of design weights to cope with outliers having large design weights is discussed in Sect. 5. Section 6 provides two examples of practical use of the introduced methods. Concluding remarks and discussing future work are in Sect. 7.

1.2 General model of the statistical production process

Figure 1 provides a general model of the statistical production process for surveys, beginning with raw electronic data. The first step is data cleaning. In this step, erroneous data are detected for correction to ensure a clean, valid dataset. The second step is imputation, where missing values are estimated and replaced as necessary to produce complete datasets for the analysis to be conducted in the next step. The final step involves estimation and formatting to produce the final statistical product.

1.2.1 Data cleaning

The objective of the data cleaning step is to find and correct errors and inconsistencies. Consequently, the outliers in this step are those with a high likelihood of having an error or inconsistency. Any detected outlier is checked and may leave unchanged if it is not wrong. Otherwise, it is corrected based on available information when possible, or removed and estimated upon necessity in the imputation step to ensure a clean dataset.

Section 2 focuses on multivariate outlier detection methods, especially those for elliptical distributions, since these types of methods have not been widely used in practice.

1.2.2 Imputation

Missing data are often unavoidable in survey statistics. Discarding missing records may cause biased estimation even when the missing values are MAR (missing at

Fig. 1 General process of statistics production with relation to outliers

random) (Little and Rubin 2002, pp. 117–127). Therefore, essential variables for estimation often require missing data imputation. Since the input for the imputation step is clean data (from Step 1), the outliers here are not erroneous data but rather extreme values that may distort estimations for imputation. An example of this is high leverage points in regression estimation. Such points may have a substantial influence on the resulting estimation for imputed values.

From among the many imputation methods available, this paper focuses on linear regression and ratio imputation. In general, introducing robust estimation improves the efficiency of the imputation compared to ordinary least squares (OLS) when applied to datasets that have longer tails than the normal distribution.

Robust regression imputation is discussed in Sect. 3, followed by robust ratio imputation in Sect. 4.

1.2.3 Estimation and formatting

In the final step of the statistical production process, the outliers in need of attention are those having large design weights. As an illustration, suppose a particular record in a household survey has a design weight of 1000 and a household income of 5 million yen (approximately 46,000 USD) per month (an atypically high-income level). This is very likely to cause a problem in the statistical tables produced from the survey. This one very wealthy household is treated as a representative of 1000 other households in the area that were not surveyed. As a consequence, the population estimate of the household income for the area will reflect that there are 1000

households with a monthly income of 5 million yen. Design weight calibration based on such “outlyingness” is discussed in Sect. 5.

2 Multivariate outlier detection methods for elliptically distributed datasets

We begin with outlier detection methods for unimodal numerical data, first establishing the difference between univariate and multivariate methods, and then introducing several multivariate methods with desirable characteristics. These methods introduced in this section are mainly used for data cleaning purposes.

2.1 Univariate methods versus multivariate methods

Univariate methods for numerical data are conventionally used in the data cleaning step to identify erroneous observations. A common practice is to set the thresholds for valid data (i.e., non-outliers) at a distance of three-sigma (or more depending on its distribution) from the mean of a target dataset. This method is essentially the idea of a control chart in the field of total quality management (TQM); however, this simple method is not robust as the thresholds are supposed to be decided with a dataset in stable condition (i.e., a dataset without outliers) (Teshima et al. 2012, pp. 173–174). It is well known that with the three-sigma rule or any other non-robust method, deciding thresholds with contaminated datasets induces a masking effect, and therefore, thresholds of such methods must be determined with datasets free from outliers. We need robust methods to determine thresholds with contaminated datasets.

Noro and Wada (2015) illustrate the problem and recommend using order statistics such as the interquartile range (IQR). A box-and-whisker plot using the IQR, as proposed by Tukey (1977), is commonly used when the target dataset is slightly asymmetric. If the dataset is highly asymmetric, an appropriate data transformation may be necessary before applying the method. The scatterplot in Fig. 2 highlights the differences between robust methods and their non-robust counterparts, as well as the distinction between univariate and multivariate methods. It displays the Hertzsprung–Russell star dataset (Rousseeuw and Leroy 1987, p. 28), which contains extreme outliers. The yellow-colored rectangular area shows the thresholds according to the three-sigma rule; the green area shows the thresholds identified by the box-and-whisker method. Both are univariate methods. The orange lines in the diagram show probability ellipses drawn with a mean vector and covariance matrix. Although this represents a multivariate approach, it, too, induces the masking effect as well as the three-sigma rule when applied to contaminated datasets. The red probability ellipses are drawn using modified Stahel-Donoho (MSD) estimators produced by robust principal component analysis (PCA) based on Béguin and Hulliger (2003). MSD and other multivariate methods are discussed in the next subsection.

Fig. 2 Differences between robust and non-robust methods both for univariate and multivariate methods. [After Wada (2010), Fig. 1.4.3, p. 98.]

2.2 Multivariate outlier detection methods for elliptical distributions

To evaluate and compare current methods for the editing and imputation of data, Eurostat conducted the EUREDIT project between March 2001 and February 2003. A series of reports were published and made available at <https://www.cs.york.ac.uk/euredit/>, along with five papers published in the *Journal of the Royal Statistical Society*. In one of the papers, Béguin and Hulliger (2004) note that NSOs had not used multivariate methods except for the Annual Wholesale and Retail Trade Survey (AWRTS) in Statistics Canada. Franklin and Brodeur (1997) report that modified Stahel-Donoho (MSD) estimators have been adopted for AWARTS and describe the algorithm. Béguin and Hulliger (2003) suggest several improvements to the estimators. Wada (2010) implemented both the original MSD and improved estimators in R and confirmed that the suggestions by Béguin and Hulliger (2003) do indeed improve performance, while the improved version of MSD estimators suffered from the curse of dimensionality. Since the improved version is incapable of processing more than 11 variables with a 32-bit PC, Wada and Tsubaki (2013) implemented an R function by parallel computing so that the function can be applied to higher-dimensional datasets.

Béguin and Hulliger (2003) suggest guiding principles for outlier detection, including good detection capability, high versatility, and simplicity. They examined several methods to estimate a mean vector and covariance matrix for elliptically

distributed datasets with a high breakdown point compared to M-estimators (Huber 1981), as well as other desirable properties such as affine and orthogonal equivariance. The methods include Fast-MCD (Rousseeuw and van Driessen 1999), which approximates the minimum covariance determinant (MCD) proposed by Rousseeuw (1985) and Rousseeuw and Leroy (1987); BACON by Billor et al. (2000), named for Francis Bacon; and the Epidemic Algorithm (EA) proposed by Hulliger and Béguin (2001), in addition to the MSD estimators used by Statistics Canada. Béguin and Hulliger (2003) compared some of these methods and found that BACON showed better detection capacity than EA for the UK Annual Business Inquiry (ABI) dataset; however, they conclude that this particular dataset does not require a sophisticated robust method.

3 Multivariate outlier detection for regression imputation

After removing or correcting erroneous data in the data cleaning step, the next step is the imputation of missing values of essential variables. From the variety of imputation methods available, the focus here is on regression imputation. Typically, OLS is used to estimate the parameters of a linear regression model; however, it is well known that the existence of outliers makes such parameter estimation unreliable. After going through the data cleaning step, survey datasets may still contain outliers in another sense. These remaining outliers are assumed to be correct; however, any extreme values in the long tails of a data distribution carry the risk of distorting the parameter estimation used for imputation regardless of their correctness. OLS regression requires to remove such outliers manually. Survey observations are divided into (sometimes a large number of) imputation classes so that a uniform response mechanism is assumed within it. Parameter estimation is conducted in each imputation class separately. A robust regression method relieves us of the burden to remove outliers from each imputation classes beforehand.

We examine M-estimation for regression, which is one of the most popular methods in this section. Disadvantages of M-estimation is also introduced together with other methods to cope with the disadvantages.

3.1 M-estimators

3.1.1 Parameter estimation of the location and regression

Generally, an M-estimate is defined as the minimization problem of

$$\sum_{i=1}^n \rho(x_i; T_n),$$

for any estimate T_n with independent random variables $x_1, \dots, x_n$. Suppose an arbitrary function ρ has a derivative $\psi(x; \theta) = (\partial/\partial\theta)\rho(x; \theta)$, T_n satisfies the implicit equation

$$\sum_{i=1}^n \psi(x_i; T_n) = 0.$$

Huber (1964) discusses the robust estimation of a mean vector, proposes M-estimation of a location with $\sum_{i=1}^n \psi(x_i - T_n) = 0$, and proves their consistency as well as asymptotic normality. Huber (1973) then extends the idea to the regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad (1)$$

with an objective variable $\mathbf{y} = (y_1, \dots, y_n)^\top$, where the error term $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^\top \sim N(0, \sigma^2)$, i.i.d. and independent of $(p+1)$ -dimensional explanatory variables $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ip})$, and regression parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^\top$. The M-estimators $\boldsymbol{\beta}$ minimizes

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

on condition that ρ is differentiable, convex, and symmetric around zero. The estimation equation is

$$\sum_{i=1}^n \psi\left(\frac{y_i - \mathbf{x}_i^\top \boldsymbol{\beta}}{\sigma}\right) \mathbf{x}_i = \sum_{i=1}^n \psi(e_i) \mathbf{x}_i = 0.$$

Due to the condition on ρ described above, ψ is supposed to be a bounded and continuous odd function, since $\psi = \rho'$. Residuals $(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})$ are standardized by a measure of scale σ to make the estimation scale equivariant. Then, $\boldsymbol{\beta}$ is estimated by solving

$$\sum_{i=1}^n w_i e_i \mathbf{x}_i = \sum_{i=1}^n w_i \left(\frac{y_i - \mathbf{x}_i^\top \boldsymbol{\beta}}{\sigma}\right) \mathbf{x}_i^\top = 0, \quad (2)$$

with a weight function defined as $w_i = \psi(e_i)/e_i$ and $w_i = w(e_i)$. Then, it can be re-expressed as

$$\sum_{i=1}^n \mathbf{x}_i^\top w_i \mathbf{x}_i \boldsymbol{\beta} = \sum_{i=1}^n \mathbf{x}_i^\top w_i y_i.$$

It can be re-expressed in a matrix form as $(\mathbf{X}^\top \mathbf{W} \mathbf{X}) \boldsymbol{\beta} = \mathbf{X}^\top \mathbf{W} \mathbf{y}$, and consequently, $\boldsymbol{\beta}$ is estimated by

$$\hat{\boldsymbol{\beta}} = [\mathbf{X}^\top \mathbf{W} \mathbf{X}]^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{y}, \quad (3)$$

where $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top$ is a $n \times (p+1)$ matrix of the explanatory variable, $\mathbf{W} = \text{diag}\{w_i\}$ is a $n \times n$ diagonal matrix of weights. After all, M-estimators for regression can be regarded as weighted least squares (WLS) estimators with their weights based on the residuals.

3.1.2 IRLS algorithm for regression

The intercept of M-estimators for regression is location equivariant, and the slope is location invariant; however, they are not scale equivariant when the scale parameter is given. Scale equivariance is achieved by estimating the scale parameter simultaneously and using it to standardize the residuals. Beaton and Tukey (1974) propose the IRLS algorithm to solve (3) with simultaneous estimation of the scale parameter. Holland and Welsch (1977) recommend it rather than Newton's method, which is theoretically desirable but difficult to implement, or Huber's method (Huber 1973; Bickel 1973), which requires more iterations.

The IRLS algorithm requires an appropriate initial estimate $\hat{\beta}^{(0)}$ and use it to obtain better next estimate of $\hat{\beta}^{(1)}$ together with $\hat{\sigma}$ based on the equation,

$$\hat{\beta}^{(j)} = \hat{\beta}^{(j-1)} + \left\{ X^T \left[W \left(\frac{y - X\hat{\beta}^{(j-1)}}{\hat{\sigma}} \right) \right] X \right\}^{-1} X^T \left\{ \left[W \left(\frac{y - X\hat{\beta}^{(j-1)}}{\hat{\sigma}} \right) \right] (y - X\hat{\beta}^{(j-1)}) \right\}.$$

The calculation is repeated until a conversion condition is met. The superscript j represents the iteration number.

There are some choices of measure for $\hat{\sigma}$. It will be discussed with a selection of a weight function since they are closely related.

3.1.3 Weight functions and measures of scale

Robust weights w_i in (2) are computed based on a weight function. Although there are a variety of choices (see, e.g., Antoch and Eklom 1995; Zhang 1997), we discuss the most popular two weight functions here among them. One is called Huber's weight function

$$w_i = w(e_i) = w\left(\frac{y_i - \mathbf{x}_i^T \hat{\beta}}{\hat{\sigma}}\right) = \begin{cases} \left[1 - (e_i/c)^2\right]^2 & |e_i| \leq c \\ 0 & |e_i| > c \end{cases}, \quad (4)$$

proposed by Huber (1964). This weight function is proved to have a unique solution regardless of the initial values (e.g., Maronna et al. 2006, p. 350) and its estimation efficiency is high with normal or nearly normal datasets (e.g., Hampel 2001; Wada and Noro 2019). The other is Tukey's biweight function

$$w_i = w(e_i) = w\left(\frac{y_i - \mathbf{x}_i^T \hat{\beta}}{\hat{\sigma}}\right) = \begin{cases} 1 & |e_i| \leq k \\ k/|e_i| & |e_i| > k \end{cases}, \quad (5)$$

by Beaton and Tukey (1974). This weight function performs well with datasets with longer tails, while it does not promise a global solution unlike Huber's weight function. The difference between these two weight functions is based on the behavior of extreme outliers. Tukey's function gives zero weight to observations very far from others, while Huber's function never gives zero weight and it cannot escape from

the influence of extreme outliers. The tuning constants c in (4) and k in (5) are sometimes called Huber's c and Tukey's k , respectively. The actual values depend on the measure of scale used.

The most popular measure of scale is median absolute deviation (MAD) defined as follows:

$$\hat{\sigma}_{\text{MAD}} = \text{median}\left(\left|r_i - \text{median}(r_i)\right|\right),$$

where residuals $r_i = y_i - \mathbf{x}_i^\top \boldsymbol{\beta}$. Huber's weight function is commonly used with MAD. Tukey's biweight function also used with MAD (e.g., Holland and Welsch 1977; Mosteller and Tukey 1977, 9. 357); however, there are also some cases with average absolute deviation (AAD),

$$\hat{\sigma}_{\text{AAD}} = \text{mean}\left(\left|r_i - \text{mean}(r_i)\right|\right).$$

Andrews et al. (1972), who conducted a large-scale Monte Carlo experiment involving robust estimation of the location parameter, show that the MAD is better than the AAD or IQR for M-estimators; however, it has not been proved that MAD is better than other scale parameters in the case of regression (Huber and Ronchetti 2009, pp. 172–173.). Holland and Welsch (1977) compare some weight functions with MAD as the measure of scale and show Huber weight function has better efficiency than the biweight function by a Monte Carlo experiment, while Bienias et al. (1997) use Tukey's biweight function with an AAD scale and mention its convergence efficiency.

Wada and Noro (2019) made a comparison of the four estimators combined these two weight functions and the measures of scale by conducting a Monte Carlo experiment with long-tailed datasets with asymmetric contamination. It is known that the 95% asymptotic efficiency on the standard normal distribution is obtained with the tuning constant $k = 1.3450$ for Huber's function (e.g., Ray 1983, p. 108), and $c = 4.6851$ for the biweight function (e.g., Ray 1983, p. 112). These figures are based on the standard deviation (SD), and the corresponding figures of MAD and AAD can be obtained by the relations

$$\begin{aligned}\frac{\sigma_{\text{AAD}}}{\sigma_{\text{SD}}} &= \frac{E|e|}{\sqrt{E(e^2)}} = \sqrt{\frac{2}{\pi}} \approx 0.80, \quad \text{and} \\ \sigma_{\text{SD}} &= \frac{1}{\Phi^{-1}\left(\frac{3}{4}\right)} \cdot \sigma_{\text{MAD}} \approx 1.4826 \cdot \sigma_{\text{MAD}},\end{aligned}$$

with cumulative distribution function of the standard normal distribution Φ where σ_{SD} , σ_{MAD} and σ_{AAD} are scale parameters based on SD, MAD and AAD, respectively. Wada and Noro (2019) obtain the results, as shown in Table 1, and compared the four estimators based on the standardized tuning constants shown in Table 2. The range of those constants is for the biweight functions with AAD based on Bienias et al. (1997) of Tukey's k , which is a part of the reports for official statistics called the Euredit Project conducted from 2000 to 2003 (Barcaroli 2002) funded

Table 1 Tuning constants for 95% asymptotic efficiency with different measures of scale. The figures first appeared in Wada (2012); and those of σ_{AAD} for k for Huber are corrected in Wada and Noro (2019)

Tuning constant	95% asymptotic efficiency		
	σ_{SD}	σ_{MAD}	σ_{AAD}
c for Tukey	4.685	3.160	3.738
k for Huber	1.345	0.907	1.073

Table 2 Tuning constants scaled for a comparison. The figures appeared in Wada (2012) and Wada and Noro (2019)

Tuning constant	Range of tuning constant for σ_{SD}			Range of tuning constant for σ_{MAD}			Range of tuning constant for σ_{AAD}		
c for Tukey	5.01	7.52	10.03	7.43	11.15	14.87	4.00	6.00	8.00
k for Huber	1.44	2.16	2.88	2.13	3.20	4.27	1.15	1.72	2.30

by Eurostat. The smaller value of these tuning constants makes the estimation more resistant to outliers, while larger value increases efficiency in estimation. Wada and Noro (2019) conclude that AAD is computationally more efficient than the widely used MAD for both weight functions. Besides, AAD is more suitable than MAD for Tukey's biweight function. Their compared estimators are available at a public repository (see Table B in Appendix).

3.2 Selection of the weight function and breakdown point

Wada and Tsubaki (2018) suggest choosing between these two weight functions based on purpose. They suggest Tukey's biweight function rather than Huber's weight in case of imputation, since the breakdown point of M-estimators for regression is $1/n$. It is the same as in OLS. Rousseeuw and Leroy (1987) report that the oldest definition of the breakdown point was given by Hodges (1967) regarding univariate parameter estimation and that Hampel (1971) generalized it. The definition offered by Donoho and Huber (1983) is for a finite sample:

Given sample size n for any sample, let

$$\mathbf{Z} = [(x_{11}, \dots, x_{1p}, y_1), \dots, (x_{n1}, \dots, x_{np}, y_n)],$$

and let T be the regression estimator applied to $\mathbf{Z}$. A new sample, $\mathbf{Z}'$, is created by replacing m of the observations arbitrarily in $\mathbf{Z}$. Let $\text{bias}(m; T, \mathbf{Z})$ be the maximum bias produced by the contamination of the replacements in the sample. The value of $\text{bias}(m; T, \mathbf{Z})$ is determined as follows:

$$\text{bias}(m; T, \mathbf{Z}) = \sup_{\mathbf{Z}'} \|T(\mathbf{Z}') - T(\mathbf{Z})\|.$$

If $\text{bias}(m; T, Z)$ is infinite, the indication is that contamination of size m breaks down the estimator. In general, the finite-sample breakdown point of estimator T for sample Z is

$$\epsilon_n^*(T, Z) = \min \left[\frac{m}{n}; \text{bias}(m; T, Z) \text{ is infinite} \right].$$

This can be regarded as the ratio of the smallest number of outliers that can make the value for T arbitrarily far from what is obtained. A breakdown point of $1/n$ means that only one extreme observation in a dataset of any size can adversely affect the estimation and that the breakdown point reaches nearly 0% with large samples. Nevertheless, Tukey's biweight function can eliminate the influence of extreme observations by giving zero weight, unlike Huber's weight function. It is the reason recommended for imputation. Those outliers are only ignored in estimating imputed values, while they are used in survey enumeration. On the other hand, if M-estimators for regression are used for population estimation, i.e., directly estimating the figures appeared in final products such as statistical tables, Huber's weight function might be more suitable as it never gives zero weight. Giving zero weight to observations in producing final survey statistics means discarding valid observations. Generally, survey statisticians working for official statistics avoid wasting precious data, since they are obtained from questionnaires filled by respondents who bear the burden to respond with goodwill.

3.3 Robust estimators to cope with outliers in explanatory variables

M-estimators have another weakness in addition to the low breakdown point that the estimators are not robust against outliers in explanatory variables. LMS (Least Median of Squares) proposed by Hampel (1975) and extended by Rousseeuw (1984), LTS (least trimmed squares) by Rousseeuw (1984), S-estimator by Rousseeuw and Yohai (1984) have higher breakdown points than M-estimators and can also cope with outliers in the explanatory variables. Unfortunately, all of them have difficulty with computation. (See, e.g., Rousseeuw and Leroy 1987; and Huber and Ronchetti 2009, pp. 195–198 for more details.)

The use of these estimators may still be in the research stage in the field of official statistics, while the software is available and may widely be used in some other fields. Generalized M (GM)-estimators and MM-estimators are popular methods. GM-estimators are introduced by Schweppe (as given in Hill 1977), and Coakley and Hettmansperger (1993). Their algorithms and software are available in Wilcox 2005. MM-estimators are first presented by Yohai (1987). Wilcox (2005) implemented an R function called *bmreg* for Schweppe-type GM-estimators and *chreg* for the other GM-estimators by Coakley and Hettmansperger (1993). In CRAN package, *robustbase* also have *lmrob* function, which implements both MM-estimators by Yohai (1987) and SMDS-estimators by Koller and Stahel (2011). Koller and Stahel (2011) achieve a 50% breakdown point and 95% asymptotic efficiency by improving MM-estimators.

Bagheri et al. (2010) compare M-estimators, MM-estimator, Schweppe-type GM-estimator, and the GM-estimator proposed by Coakley and Hettmansperger (1993), concluding that the GM-estimators proposed by Coakley and Hettmansperger were the best among the group. Wada and Tsubaki (2018) examine M-estimators and GM-estimators by Coakley and Hettmansperger (1993) for weight calibration, which will be discussed in Sect. 5. They mention that the explanatory variables chosen for imputation are often selected from among the auxiliary variables used for stratification in sample surveys. If this is the case, outliers in explanatory variables are not expected, and M-estimators could be more suitable than GM-estimators. GM-estimators reduce the robust weight of leverage points in addition to the outliers in the objective variable. It provides robustness while reduces estimation efficiency.

4 Robustification of the ratio estimation for imputation

4.1 Difference between regression imputation and ratio imputation

In regression imputation, missing values y_i in the target variable are replaced by estimated values $\hat{y}_i$ based on a regression model with auxiliary x variables using complete observations regarding all those x and y in the target dataset (e.g., De Waal et al. 2011, p. 230).

Ratio imputation is a special case of regression imputation (De Waal et al. 2011, pp. 244–245), where missing y_i are replaced by the ratio of y_i to a single observed auxiliary x_i . Specifically, the ratio model is

$$y_i = \beta x_i + \epsilon_i, \quad (6)$$

where missing y_i are replaced by $\hat{y}_i = \hat{\beta}x_i$ with the estimated ratio

$$\hat{\beta} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i} \quad (7)$$

where data $i = 1, \dots, n$ of (x, y) are observed n elements in the imputation class. (See Cochran 1977, pp. 150–164 for further details.)

The ratio model (6) resembles the single regression model without an intercept,

$$y_i = \beta x_i + \epsilon_i. \quad (8)$$

However, there is a difference in the error terms. The error term for the ratio model (6) is heteroscedastic and expressed as $\epsilon_i \sim N(0, x_i\sigma^2)$ with scale parameter σ , while the error in the regression model (8) is homoscedastic, as described for the regression model (1), and expressed as $\epsilon_i \sim N(0, \sigma^2)$. Because of this error term difference, the ratio model has an advantage in imputation over regression models in its ability to fit heteroscedastic datasets without data transformations that can make the estimation of means and totals unstable. On the other hand, the heteroscedastic error is an obstacle to robustifying the ratio estimator by means of M-estimation.

4.2 Generalization and robustification of the ratio model

For robustification, Wada and Sakashita (2017) and Wada et al. (2021) re-formulate the original ratio model with the heteroscedastic error term ϵ_i as follows:

$$y_i = \beta x_i + \sqrt{x_i} \epsilon_i,$$

since the two error terms discussed above have the relation, $\epsilon_i = \sqrt{x_i} \epsilon_i$.

They then extend the model to

$$y_i = \beta x_i + x_i^\gamma \epsilon_i, \quad (9)$$

with an error term proportional to x_i^γ . The corresponding ratio estimator becomes

$$\hat{\beta} = \frac{\sum_{i=1}^n y_i x_i^{1-2\gamma}}{\sum_{i=1}^n x_i^{2(1-\gamma)}}. \quad (10)$$

When $\gamma = 1/2$, the model (9) and the estimator (10) correspond to the original ratio model (6) and the estimator (7). According to the value of γ , the generalized model has different features. It also corresponds to the single regression model with an intercept when $\gamma = 0$. Takahashi et al. (2017) also discuss the same model regarding the datasets following the log-normal model and proposed estimation of γ .

The robustified generalized ratio estimator by Wada and Sakashita (2017) and Wada et al. (2021) is

$$\hat{\beta}_{\text{rob}} = \frac{\sum w_i y_i x_i^{1-2\gamma}}{\sum w_i x_i^{2(1-\gamma)}}, \quad (11)$$

where w_i is obtained by a weight function with homoscedastic quasi-residuals,

$$\check{r}_i = \frac{y_i - \hat{\beta}_{\text{rob}} x_i}{x_i^\gamma},$$

and a scale parameter σ .

4.3 Further development: simultaneous estimation of γ

Wada and Sakashita (2017) and Wada et al. (2021) considered the generalized ratio model with fixed γ values, which requires model selection before estimation for imputation. Wada et al. (2019) proposed eliminating the model selection step by simultaneously estimating β and γ in (11) by means of two-stage least squares (2SLS) (e.g., Greene 2002, p. 79) with iterations. The initial estimate of β is obtained by OLS under the model (6). This estimation is not efficient; however, unbiased under heteroscedasticity. Using the instrumental variable, $r_i^2 = (y_i - \hat{\beta} x_i)^2$,

$$\log |r_i| = \gamma \log |x_i| + \log(\sigma),$$

is obtained with $r_i = y_i - \hat{\beta}x_i$. It means that γ is obtained as the single regression parameter. Then, new $\hat{\beta}$ will be obtained using $\hat{\gamma}$, and new $\hat{\gamma}$ will also be estimated based on the latest $\hat{\beta}$. See Wada, Takata and Tsubaki (2019) for more detail of the algorithm. The implemented function named *RBred* is available with its non-robust version called *Bred* (see Appendix). An evaluation was made using contaminated random datasets. Estimation by *RBred* was found to have better efficiency than the *Roptim* function; in particular, its estimation of β appears to be useful for imputation. However, further evaluation may be necessary, especially regarding the estimation of γ , although the value of γ is not necessary for imputation.

5 Weight calibration

The outliers focused on in the estimation and formatting step are extreme values with large design weights. The Horvitz–Thompson estimator is widely used to estimate finite population means and totals for conventional statistical surveys. In such cases, design weights, which are the inverse of the sampling rate, are used as multipliers for each observation. The problem lies in deciding whether an extreme observation deserves the corresponding design weight. (A weight of 1000 applied to an observation means that the value of the observation represents 1000 population elements that were not sampled.) Chambers (1986) considers this outlier problem and argues that for “nonrepresentative” outliers or unique data points that have been judged free of any errors or mistakes, the design weight should be one corresponding to a single population element. This implies that these outliers do not represent other population elements that are not sampled, and consequently, they do not influence the estimation process in any substantial way.

Wada and Tsubaki (2018) propose a design weight calibration method utilizing the robust weights obtained by the M-estimators for regression described in Sect. 3. Henry and Valliant (2012) classified estimation methods for population means or totals in sample surveys as model-based approaches, design-based approaches, and model-assisted approaches. The proposed method corresponds to the latter: the model-assisted approach.

To illustrate, consider selecting a sample using random sampling without replacement from finite population U containing N elements $u_l, l = 1, \dots, N$. The extracted sample S , contains n elements $v_i, i = 1, \dots, n$. Let $\pi = n/N$ be the probability that a population element is included in the sample S . The associated design weight for a sampled element i in S is $g_i = 1/\pi$. Therefore, $\sum_{i=1}^n g_i = N$. The Horvitz–Thompson (HT) estimator (Horvitz and Thompson 1952) for population total $T = \sum_{l=1}^N u_l$ in this case is

$$T_{\text{HT}} = \frac{N}{n} \sum_{i=1}^n v_i.$$

This is also called the inverse probability weight (IPW) estimator.

It is known that the efficiency of an HT-estimator decreases with the presence of outliers or when applied to non-normal data distribution, as HT-estimators have characteristics similar to OLS estimators. To improve the efficiency of the HT-estimator, Wada and Tsubaki (2018) use robust regression estimation. The idea is to adjust the conventional design weight g_i by multiplying it by the robust weight, w_i , obtained by (3) after the iterations converge, since w_i determined by residual can be regarded as an indicator of “outlyingness.” Additional adjustment is necessary to make the sum of the adjusted weight $g_i^* = g_i w_i$ meet the necessary condition that $\sum_{i=1}^n g_i^* = N$. The following two adjustments have been proposed:

$$g_i^{**} = \frac{g_i^* \sum_{i=1}^n g_i}{\sum_{i=1}^n g_i^*}, \quad \text{and} \quad (12)$$

$$g_i^{***} = 1 + \frac{g_i^* \sum_{i=1}^n (g_i - 1)}{\sum_{i=1}^n g_i^*}. \quad (13)$$

The adjustment shown in (12) is a natural form; however, the adjusted weight, g_i^{**} , becomes zero when $w_i = 0$. In such cases, the corresponding observation is actually removed from the population estimation process. However, for official statistics, ignoring a sampled observation is not desirable, since the observation exists in the population. For this reason, the adjustment shown in (13), which guarantees a minimum value of 1 for each g_i^{***} , is proposed.

Design weight calibration by robust weight has an advantage in that the reduction of estimation efficiency is less than the reduction in model-based approaches when the data distribution deviates from the applied model. Wada and Tsubaki (2018) confirm the usefulness of the proposed adjustment (13) in Monte Carlo simulations with random and real datasets.

One disadvantage of this weight calibration method may be that the weight is assigned to a variable in observation, while in conventional approaches, the design weight is assigned to the observation (i.e., all variables in observation are assigned the same design weight).

6 Examples of practical applications

6.1 MSD estimators for unincorporated enterprise survey

Unincorporated Enterprise Survey, conducted by the Statistics Bureau of Japan, Ministry of Internal Affairs and Communications (MIC), had major changes in 2019. Industries surveyed are extended, the sample size is increased ten times, i.e., approximately 37 thousand samples from 3.7 thousand, and questionnaires are collected by mail instead of enumerators. A process of a hot deck imputation is added for accounting items such as sales, total expenses, purchases, operating expenses, and inventories together with a cleaning process of the hot deck donor candidates. It had not been necessary since the response rate of the previous survey was almost 100%; however, mail surveys are typically expected to increase non-responses.

Wada et al. (2020) evaluate outlier detection methods using a mean vector and covariance matrix assume symmetric data distributions. The blocked adaptive computationally efficient outlier nominators (BACON) by Billor et al. (2000), improved MSD by Béguin and Hulliger (2003), Fast-MCD by Rousseeuw and Driessen (1999), and NNVE by Wang and Raftery (2002) are compared using skewed and long-tailed random datasets with asymmetrical contamination, and improved MSD is selected. They also examine an appropriate data transformation for the highly skewed target variables based on the number of outliers detected and scatter plot matrices. Their target variables are highly correlated accounting items that do not have values less than zero, and expected outliers have mostly large values. Therefore, the suitable data transformation, in this case, could be slightly loose than the one which makes the data strictly symmetric. They select the data transformation, which detects a minimum number of outliers in larger values and shows that removing outliers from hot deck donor candidates improves estimation for imputation. Log transformation, biquadratic root transformation, and square root transformation are compared, and biquadratic root transformation is selected. The lower triangular matrix of Fig. 3 shows an example of the manufacturing industry. Based on their results, outliers regarding highly correlated four variables (sales, total expenses, purchases, and operating expenses) of Unincorporated Enterprise Survey are detected by improved MSD after biquadratic root transformation. Beginning inventory and Ending inventory are excluded since there are certain amounts of observation with zero values in some industries, and they do not have high covariance with other variables, although these two variables are highly correlated with each other. Those outliers are removed from hot deck donor candidates in each imputation class, while they are used in aggregation for producing statistical tables.

6.2 Application of the robust estimator of the generalized ratio model

The robust estimator of the generalized ratio model (9) is adopted for the imputation of major corporate accounting items in the 2016 Economic Census for Business Activity in Japan (Wada and Sakashita 2017; Wada et al. 2021), conducted by the Ministry of Internal Affairs and Communications (MIC) and the Ministry of Economy, Trade and Industry (METI) jointly. The items to be imputed are sales explained by expenses, expenses by sales, and salaries by expenses. The model (9) with $\gamma = 1/2$ and $\gamma = 1$ are compared, and $\gamma = 1/2$ is adopted for all of the imputed items.

The imputation class is determined by CART (classification and regression trees). The target variable is the ratio for imputation, and the possible explanatory variables are the 3.5-digit industrial classification code, legal organization, number of employees, type of cooperative, number of regular domestic employees, number of domestic branch offices, and number of branch offices. They are the variables available from Statistical Business Register before the 2016 Census since the imputation class has to be determined before collecting questionnaires. Statistical Business Register is a database on business establishments and enterprises across the country made from the previous Census and surveys as well as various administrative

Fig. 3 Outliers detected by MSD estimators in the manufacturing industry with square root in the upper triangular matrix, and biquadratic root transformation in the lower matrix. [After Wada et al. (2020), Fig. 3, p. 10.]

information. Among those explanatory variables, type of cooperative and number of regular domestic employees are adopted. The minimum number of complete observations for parameter estimation in each imputation class is 30. In case if an imputation class does not have enough observations, the class is merged with another class within the same 1-digit industrial classification code. If there are plural choices, the merged class is determined by the Mann–Whitney U test.

7 Concluding remarks

The focus of this paper is controlling the influence of outliers in the survey data processing. In addition to conventional univariate methods, some of the multivariate methods introduced here have come to be used in practice, although examples of their use remain limited for the time being. Other methods are still in the research

stage. For example, the simultaneous estimation of β and γ described in Sect. 4.3 is in the process of improvement, and validation of the weight calibration approach in Sect. 5 may require more time since weighting by item in each observation (rather than by observation) is not yet common practice.

The author hopes that this paper will promote efficiency improvements in official statistics, in terms of both estimation and statistical production.

Acknowledgements This work was supported by JSPS KAKENHI Grant number JP16H2013.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Appendix: Available software discussed in various sections of the paper

Table A. Software in Sect. 2. Tirls.aad is based on Bienias et al. (1997). Used by Wada (2010), Wada and Noro (2019)

Method	Explanation
BACON	S-plus code: in Béguin and Hulliger (2003) Instruction to port S-plus code to R: https://github.com/kazwd2008/BEM
MSD	R single core version both for used in Statistics Canada and improved version: https://github.com/kazwd2008/MSD R paralleled version for high-dimensional data: https://github.com/kazwd2008/MSD.parallel
Fast-MCD	R <i>covMcd</i> function in <i>rrcov</i> package at CRAN (https://cran.r-project.org)
EA	S-plus code is available in Béguin and Hulliger (2003)
NNVE	R <i>cov.nnve</i> function in <i>covRobust</i> package at CRAN

Table B. Functions for M-estimators of Sects. 3 and 4 available at <https://github.com/kazwd2008/IRLS>

File name	R function	Feature	Weight function	Scale parameter
Tirls.r	Tirls.aad	Robust estimation for Regression model	Tukey	AAD
	Tirls.mad			MAD
Hirls.r	Hirls.aad		Huber	AAD
	Hirls.mad			MAD
RrT.r*	RrTa.aad	Robust estimation for generalized ratio model with a fixed γ value	Tukey	AAD
	RrTb.aad			
	RrTc.aad			
	RrTa.mad			MAD
	RrTb.mad			
	RrTc.mad			
RrH.r*	RrHa.aad		Huber	AAD
	RrHb.aad			
	RrHc.aad			
	RrHa.mad			MAD
	RrHb.mad			
	RrHc.mad			
RBreds.r	RBred	(i)	Tukey	AAD
	Bred	(ii)		

*All functions in RrT.r and RrH.r are included in the REGRM package at <https://github.com/kazwd2008/REGRM>.

- (i) Robust estimation for generalized ratio model (γ and β are simultaneously estimated)
- (ii) Non robust estimation for generalized ratio model (γ and β are simultaneously estimated)

Table C. Functions for more advanced estimators for regression introduced in Sect. 3

Package	Function [package]	Location
GM-estimators (Schweppe-type)	Bmreg [WRS]	https://github.com/nicebread/WRS
GM-estimators by Coakley and Hettmansperger (1993)	cmreg [WRS]	
MM-estimators	lmrob [robustbase]	CRAN (https://cran.r-project.org)

References

- Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., & Tukey, J. W. (1972). *Robust estimates of location: Survey and advances*. Princeton: Princeton University Press.
- Antoch, J., & Eklom, H. (1995). Recursive robust regression computational aspects and comparison. *Computational Statistics & Data Analysis*, 19, 115–128.
- Bagheri, A., Midi, H., Ganjali, M., & Eftekhari, S. (2010). A comparison of various influential points diagnostic methods and robust regression approaches: Reanalysis of interstitial lung disease data. *Applied Mathematical Sciences*, 4(28), 1367–1386. <https://www.m-hikari.com/ams/ams-2010/ams-25-28-2010/bagheriAMS25-28-2010.pdf>.

- Barcaroli, G. (2002). The Euredit project: activities and results. *Rivista di statistica ufficiale*.
- Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (3rd ed.). West Sussex: Wiley.
- Beaton, A. E., & Tukey, J. W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, 16, 147–185.
- Béguin, C. & Hulliger, B. (2003). Robust multivariate outlier detection and imputation with incomplete survey data. *EUREDIT Deliverable*, D4/5.2.1/2 Part C. <https://www.cs.york.ac.uk/euredit/results/results.html>. Accessed 19 Oct 2020.
- Béguin, C., & Hulliger, B. (2004). Multivariate outlier detection in incomplete survey data: The epidemic algorithm and transformed rank correlations. *Journal of the Royal Statistical Association, Series A*, 167(Part 2), 275–294.
- Bickel, P. J. (1973). On some analogues to linear combinations of order statistics in the linear model. *The Annals of Statistics*, 1(4), 597–616.
- Bienias, J. L., Lassman, D. M., Scheleur, S. A. & Hogan H. (1997). Improving outlier detection in two establishment surveys. In UNSC and UNECE (Eds.), *Statistical Data Editing 2: Methods and Techniques*, 76–83. <http://www.unece.org/fileadmin/DAM/stats/publications/editing/SDE2.pdf>. Accessed 19 Oct 2020.
- Billor, N., Hadi, A. S., & Velleman, P. F. (2000). BACON: Blocked adaptive computationally efficient outlier nominators. *Computational Statistics & Data Analysis*, 34, 279–298.
- Chambers, R. L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81, 1063–1069.
- Coakley, C. W., & Hettmansperger, T. P. (1993). A bounded influence, high breakdown, efficient regression estimator. *Journal of the American Statistical Association*, 88, 640–644.
- Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.
- De Waal, T., Pannekoek, J., & Scholtus, S. (2011). *Handbook on statistical data editing and imputation*. New York: Wiley.
- Donoho, D. L., & Huber, P. J. (1983). The notion of breakdown point. In P. Bickel, K. Doksum, & J. L. Hodges Jr. (Eds.), *A Festschrift for Erich L. Lehmann*. Belmont: Wadsworth.
- Economic Commission for Europe of the United Nations (UNECE). (2000) *Glossary of terms on statistical data editing*, Conference of European Statisticians Methodological material, Geneva.
- Franklin, S., & Brodeur, M. (1997). A practical application of a robust multivariate outlier detection method. In *Proceedings of the Survey Research Methods Section* (pp. 186–191). American Statistical Association. http://www.asasrms.org/Proceedings/papers/1997_029.pdf. Accessed 19 Oct 2020.
- Greene, W. H. (2002). *Econometric analysis* (5th ed.). Upper Saddle River: Prentice Hall.
- Hampel, F. R. (1971). A general qualitative definition of robustness. *Annals of Mathematical Statistics*, 42, 188–1896.
- Hampel, F. R. (1975). Beyond location parameters: Robust concepts and methods (with Discussion), *Bulletin of the ISI*, 46 (pp. 375–391).
- Hampel, F. (2001). Robust statistics: A brief introduction and overview. Research Report No.94, Seminar für Statistik, Eidgenössische Technische Hochschule (ETH), Zürich. <https://www.research-collection.ethz.ch/bitstream/handle/20.500.11850/145174/1/eth-24068-01.pdf>. Accessed 19 Oct. 2020.
- Henry, K., & Valliant, R. (2012) Comparing alternative weight adjustment methods, section on survey research methods. In *Proceedings of the Joint Statistical Meeting (JSM2012)*, 4696–4710. http://www.asasrms.org/Proceedings/y2012/Files/306157_76012.pdf. Accessed 19 Oct 2020.
- Hill, R. W. (1977). Robust regression when there are outliers in the carriers. unpublished Ph.D. thesis, Harvard University, Dept. of Statistics.
- Hodges, J. L., Jr. (1967) Efficiency in normal samples and tolerance of extreme values for some estimates of location. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1, 163–168. https://digitalassets.lib.berkeley.edu/math/ucb/text/math_s5_v1_article-10.pdf. Accessed 19 Oct 2020.
- Holland, P. W., & Welsch, R. E. (1977). Robust regression using iteratively reweighted least-squares. *Communications in Statistics Theory and Methods*, A6(9), 813–827.
- Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite population. *Journal of the American Statistical Association*, 47, 663–685.
- Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35(1), 73–101.
- Huber, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, 1(5), 799–821.
- Huber, P. J. (1981). *Robust statistics*. New York: Wiley.

- Huber, P. J. (1983). Minimax aspects of bounded-influence regression. *Journal of the American Statistical Association*, 78, 66–80.
- Huber, P. J., & Ronchetti, E. M. (2009). *Robust statistics* (2nd ed.). New York: Wiley.
- Hulliger, B. & Béguin, C. (2001). Detection of multivariate outliers by a simulated epidemic. In *Proceedings of the ETK/NTTS 2001 Conference*, 667–676. Eurostat. <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.519.7282&rep=rep1&type=pdf>. Accessed 19 Oct 2020.
- Koller, M., & Stahel, W. A. (2011). Sharpening wald-type inference in robust regression for small samples. *Computational Statistics & Data Analysis*, 55(8), 2504–2515.
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.
- Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). *Robust statistics: Theory and methods*. Wiley.
- Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression*. Reading: Addison Wesley.
- Nakamura, H. (2017). Microdata access for official statistics in Japan: Focusing mainly on microdata access at onsite facilities. *Sociological Theory and Methods*, 32(2), 310–320.
- Noro, T., & Wada, K. (2015). A univariate outlier detection manual for tabulating statistical survey (in Japanese). *Research Memoir of Official Statistics*, 72, 41–53. URL <https://www.stat.go.jp/training/2kenkyu/ihou/72/pdf/2-2-723.pdf>.
- Pearce, B. (1852). Criterion for the rejection of doubtful observations. *Astronomical Journal II*, 45, 161–163.
- Ray, W. J. J. (1983). *Introduction to robust and quasi-robust statistical method*. Springer-Verlag.
- Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79(388), 871–880.
- Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. In W. Grossmann, G. Pflug, I. Vincze, & W. Wertz (Eds.), *Mathematical statistics and its applications, vol. B* (pp. 283–297). Dordrecht: Reidel.
- Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. New York: Wiley.
- Rousseeuw, P. J., & Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41, 212–223.
- Rousseeuw, P. J., & Yohai, V. J. (1984). Robust regression by means of S-estimators. In J. Franke, W. Härdle, & D. Martin (Eds.), *Robust and nonlinear time series analysis* (pp. 256–272). New York: Springer.
- Takahashi, M., Iwasaki, M., & Tsubaki, H. (2017). Imputing the mean of a heteroskedastic log-normal-missing variable: A unified approach to ratio imputation. *Statistical Journal of the IAOS*, 33, 763–776.
- Teshima, S., Hasegawa, Y., & Tatebayashi, K. (2012). *Quality recognition and prediction: Smarter pattern technology with the Mahalanobis-Taguchi system*. New York: Momentum Press.
- Tukey, J. W. (1977). *Exploratory data analysis*. Reading: Addison-Wesley.
- Wada K. (2010). Detection of multivariate outliers: Modified Stahel-Donoho estimators (in Japanese). *Research Memoir of Official Statistics*, 67, 89–157. <https://www.stat.go.jp/training/2kenkyu/pdf/ihou/67/wada1.pdf>.
- 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. <https://www.stat.go.jp/training/2kenkyu/ihou/69/pdf/2-2-692.pdf>.
- Wang, N., & Raftery, A. E. (2002). Nearest-neighbor variance estimation (NNVE) robust covariance estimation via nearest-neighbor cleaning. *Journal of the American Statistical Association*, 97(260), 994–1019.
- Wada, K., Kawano, M., & Tsubaki, H. (2020). Comparison of multivariate outlier detection methods for nearly elliptical distributions. *Austrian Journal of Statistics*, 49(2), 1–17. <https://doi.org/10.17713/ajs.v49i2.872>.
- 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. <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. https://nt17.pg2.at/data/abstracts/abstract_56.html. Accessed on 14 Dec 2019.
- Wada, K., Sakashita, K., & Tsubaki, H. (2021). Robust estimation for a generalised ratio model. *Austrian Journal of Statistics*, 50, 74–87. <https://doi.org/10.17713/ajs.v50i1.994>.
- Wada, K., Takata, S. & Tsubaki, H. (2019) An algorithm of generalized robust ratio model estimation for imputation. In *JSM Proceedings, Government Statistics Session* (pp. 3120–3128). Denver: American Statistical Association.

- Wada, K., & Tsubaki, H. (2013). Parallel computation of modified Stahel-Donoho estimators for multivariate outlier detection. In *Proceedings of 2013 IEEE International Conference on Cloud Computing and Big Data (CloudCom-Asia)*, 304–311, 16–19, Dec. 2013, Fuzhou, China. <https://ieeexplore.ieee.org/document/6821008>. Accessed 19 Oct 2020.
- Wada, K., & Tsubaki, H. (2018) Model assisted design weight calibration by outlyingness (in Japanese). *Bulletin of the Computational Statistics of Japan*, 31(2), 101–119. https://www.jstage.jst.go.jp/article/jscswabun/31/2/31_101/_pdf/-char/ja. Accessed 19 Oct 2020.
- Wilcox, R. (2005). *Introduction to robust estimation and hypothesis testing* (3rd ed.). New York: Elsevier.
- Yohai, V. (1987). High breakdown-point and high efficiency estimates for regression. *Annals of Statistics*, 15, 642–665.
- Zhang, Z. (1997). Parameter estimation techniques: A tutorial with application to conic fitting. *Image and Vision Computing*, 15(1), 59–76.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.