Japanese Journal of Statistics and Data Science (2022) 5:783–830
https://doi.org/10.1007/s42081-022-00164-0

ORIGINAL PAPER

Data Science: Present and Future

# A new robust ratio estimator by modified Cook's distance for missing data imputation

**Masayoshi Takahashi**[1]

Received: 11 December 2021 / Revised: 15 May 2022 / Accepted: 16 May 2022 /
Published online: 6 July 2022


## Abstract

In survey data, missing values are prevalent. In official economic statistics, where data are obtained through surveys, ratio imputation is often utilized to deal with missing data; however, outliers may have an influence on the imputation model. The objective of this article is to propose a new robust ratio estimator, named the TC-ratio estimator (ratio estimator with trimming based on Cook's distance), which is robust against outliers on the vertical axis (variable $y$), on the horizontal axis (variable $x$), and on both axes ($x$ and $y$), for missing data imputation. Also, a novel way is suggested to automatically determine the number of outliers. To assess the performance of the new method, Monte Carlo simulations are conducted under 160 different data generation processes, each repeated in 10,000 simulation runs. Relative superiority of the new method is shown against the traditional robust ratio imputation methods, such as the ratio of medians, trimmed means, Winsorized means, and means by $M$-estimators. The current study finds that the new method outperforms these traditional methods when outliers are present only in $y$, only in $x$, and both in $x$ and $y$. Furthermore, when outliers are not present, the performance of this new method is approximately equal to the non-robust method.

**Keywords** Ratio imputation · Ratio estimator · Missing · Outlier · Robust

## 1 Introduction

The ratio estimator is commonly used to estimate the mean or the total of a variable of interest in many fields (Royall, 1970; Cochran, 1977, pp.150–188; Lu & Yan, 2014, p.1). Examples can be found in medical research (Wang et al., 2011), marine science (Hoenig et al., 1997; Stock et al., 2019), forest research (Bullock et al., 2020;

✉ Masayoshi Takahashi
m-takahashi@nagasaki-u.ac.jp

[1] School of Information and Data Sciences, Nagasaki University, Nagasaki, Japan

Snowdon, 1992; Zarnoch & Bechtold, 2000), population growth research (Severud et al., 2019), and, last but not least, official statistics (Scheaffer, 2012, p.171).

When data are obtained through survey questionnaires, missing values are prevalent in data. For example, in official economic statistics, some enterprises do not answer sensitive items such as turnover (sales), while the same enterprises answer non-sensitive items such as the number of employees. Under this circumstance, to compute the mean (or the total) of turnover, missing values are often taken care of by imputation, where the missing values in turnover are predicted by the observed values in the number of employees. While multiple imputation may be used for variances and covariances, single imputation can yield reasonable estimates of means and totals (Takahashi & Watanabe, 2017, p.24; Little & Rubin, 2020, p.72). Specifically, ratio imputation is often used for missing values in official economic statistics (de Waal et al., 2011, pp.244–245; Takahashi et al., 2017).

However, official economic statistics deal with a variety of enterprises, such as small-and-medium size enterprises and large enterprises. Thus, it is important to consider the effects of outliers on the imputation model when dealing with missing values. This is of great importance, because the presence of outliers biases the parameter of the imputation model, which leads to biases in imputed data, leading to biased results in statistical analyses. This is also important for data science in general, because data science requires high-quality data. Official statistics is one of the most important sources of data; however, the quality of such data is dependent on how missing values are taken care of. Therefore, this research contributes to data science in general by helping official statistics to deal with missing values that plague the quality of data.

Traditionally, the following robust estimators are suggested: The median (de Waal et al., 2011, p.210), the trimmed mean (de Waal et al., 2011, p.211), the Winsorized mean (Gwet & Rivest, 1992, p.1174; Mulry et al., 2014, pp.724–725), and the mean by $M$-estimators (Gwet & Rivest, 1992, p.1175; Mulry et al., 2014, pp.725–727, Wada & Sakashita, 2017, p.3). However, the median, the trimmed mean, and the Winsorized mean are univariate approaches to outliers. Thus, they are potentially sensitive to bivariate outliers. We focus on bivariate outliers, but not on higher dimensional outliers, because the ratio imputation model is intrinsically bivariate (de Waal et al., 2011, pp.244–245). Also, the mean by $M$-estimators takes only the residuals into account, which makes it potentially sensitive to outliers (high-leverage points) on the horizontal axis in the sense of the scatter plot, where the predictor is on the horizontal axis and the target variable for imputation is on the vertical axis.

This article proposes a new robust ratio estimator named the TC-ratio estimator, an extension of the ratio estimator with trimming based on Cook's distance (1977). Also, this article applies the TC-ratio estimator to the ratio imputation model to make it robust against outliers on the vertical axis (variable $y$), on the horizontal axis (variable $x$), and on both axes ($x$ and $y$). Furthermore, this study proposes a novel method of automatically determining the number of outliers, based on the coefficient of determination $R^2$. Thus, the process can be automated, which is meaningful in the practice of official statistics, because the schedule for producing estimates in official statistics is usually tight; thus, if a new method can detect and treat influential values in an automated manner, it is more preferable (Mulry et al., 2014, p.722).

Monte Carlo simulations based on 1,600,000 datasets reveal that the robust ratio imputation model by the TC-ratio estimator outperforms the traditional robust ratio imputation models when outliers exist only in $y$, only in $x$, and both in $x$ and $y$. When outliers are not present, the performance of the robust ratio imputation model by the TC-ratio estimator is also approximately equivalent to the non-robust ratio imputation method.

## 2 The ratio estimator

Suppose that the population model is $y_i = \beta x_i + \varepsilon_i$, where $y_i$ is the target incomplete variable, $x_i$ is an auxiliary variable (completely observed), and $\varepsilon_i \sim N\left(0, \sigma^2 x_i^{2\xi}\right)$, where $\xi$ is some constant. In other words, the model is regression without an intercept, also known as regression through the origin (Eisenhauer, 2003; de Waal et al., 2011, p. 245), and the error term $\varepsilon_i$ has the expected value of zero, but the variance is proportional to $x_i^{2\xi}$; in other words, it is heteroskedastic.

Takahashi et al., (2017) show that the weighted least squares (WLS) transform the heteroskedastic error term $\varepsilon_i$ into the homoskedastic error term $\gamma_i = \varepsilon_i / x_i^{\xi}$, where $\gamma_i \sim N\left(0, \sigma^2\right)$. Since $x_i^{\xi}$ is a function of $x_i$, not only the expected value of $\varepsilon_i / x_i^{\xi}$ is zero, conditional on $x_i$, but also, the variance of $\varepsilon_i / x_i^{\xi}$ is constant, conditional on $x_i$. Therefore, Eq. (1) corrects for heteroskedasticity. See Takahashi et al. (2017) about how Eq. (1) is obtained. Note that, in this article, the sums are taken from $i = 1, 2, \ldots, n$, where $n$ is the sample size, unless otherwise stated. Furthermore, the homoskedastic error term $\gamma_i$ is shown in Eq. (2).

$$\widehat{\beta}_{\mathrm{WLS}} = \frac{\sum x_i^{1-2\xi} y_i}{\sum x_i^{2(1-\xi)}}, \tag{1}$$

$$\gamma_i = \frac{y_i - \widehat{\beta}_{WLS} x_i}{x_i^{\xi}}. \tag{2}$$

When $\xi = 0.0$, $\widehat{\beta}_{\mathrm{WLS}}$ reduces to the ordinary least squares (OLS) estimator $\widehat{\beta}_{\mathrm{OLS}}$ in Eq. (3), and the corresponding residual $e_i$ is Eq. (4).

$$\widehat{\beta}_{\mathrm{OLS}} = \frac{\sum x_i^{1-2\times 0.0} y_i}{\sum x_i^{2(1-0.0)}} = \frac{\sum x_i y_i}{\sum x_i^2}. \tag{3}$$

$$e_i = y_i - \widehat{\beta}_{OLS} x_i, \tag{4}$$

When $\xi = 0.5$, $\widehat{\beta}_{WLS}$ reduces to the ratio-of-means estimator $\widehat{\beta}_{\mathrm{ratio}}$ in Eq. (5), which is also known as the ratio estimator (Royall, 1970, p.380; Cochran, 1977, p.150), and the corresponding residual $e_{r,i}$ for the ratio estimator ($\xi = 0.5$) is Eq. (6), where

subscript $r$ denotes ratio. Equation (6) will be an important component to robustify the ratio imputationmodel.

$$\widehat{\beta}_{\text{ratio}} = \frac{\sum x_i^{1-2\times 0.5} y_i}{\sum x_i^{2(1-0.5)}} = \frac{\sum y_i}{\sum x_i} = \frac{\sum y_i/n}{\sum x_i/n} = \frac{\overline{y}}{\overline{x}}, \tag{5}$$

$$e_{r,i} = \frac{y_i - \widehat{\beta}_{\text{ratio}} x_i}{\sqrt{x_i}}. \tag{6}$$

Since $\widehat{\beta}_{\text{ratio}}$ is based on arithmetic means, it does not take the rocket scientist to imagine that $\widehat{\beta}_{\text{ratio}}$ is sensitive to outliers. This is the problem that the current study seeks to solve.

## 3 Definition of outliers and influential observations

The definition of outliers is vague, because outliers are only defined in relation to other observations in the remaining data. Suffice it to say that outliers are those observations that appear to be different from the rest of the data (Ghosh-Dastidar & Schafer, 2006, p.487; Wooldridge, 2020, p.317). In statistical analyses, the presence of outliers may imply that the model sufficiently describes the majority of observations, but it does not describe a small number of observations. Under this circumstance, the data may be modeled as a mixture of two types of distributions (Schafer, 1997, p.385).

There are a variety of reasons why outliers exist in data, but two distinctions are important: (1) outliers are incorrect observations (errors); (2) outliers are correct but unusual observations (Gwet & Rivest, 1992 p.1174; Bonate, 2011, p.71). If outliers are incorrect observations, then these outliers should be corrected in the editing process before imputing missing values (de Waal, 2013; Di Zio & Guarnera, 2013; Ghosh-Dastidar & Schafer, 2006). On the other hand, the types of outliers in the current study are correct but unusual observations. If an observation is correct but has an excessive effect on an estimate of a parameter, then the observation is regarded as influential (Mulry et al., 2014, p.721). As is the case with Mulry et al. (2014, p.721), the focus of this study "is on influential values that remain after all the data have been verified or corrected, so these unusual values are true and not the result of reporting or recording errors." This is important to consider, because if outliers are correct but influential observations, these outliers remain in missing data at the imputation stage. Then, $\widehat{\beta}_{\text{ratio}}$ is influenced by outliers and $\widehat{\beta}_{\text{ratio}}$ is biased, which leads to biases in the imputed data, which further leads to biased results in statistical analyses based on imputed data.

Then, a natural question is what are the influential observations? To discuss this issue, let us first consider unconditional (univariate) outliers and conditional (bivariate) outliers (Fox, 2020, p.40). Suppose that heights are normally distributed with the mean of 170 cm and the standard deviation of 6 cm. If someone's

height is 200 cm, then this is an unconditional (univariate) outlier, because it is five standard deviations above the mean. Also, suppose that weights are normally distributed with the mean of 60 kg with the standard deviation of 10 kg. If the same person's weight is 110 kg, then this is again an unconditional (univariate) outlier, because it is five standard deviations above the mean. However, this person is unlikely to be a conditional (bivariate) outlier. Conditional on the value of the person's height (200 cm), this person's weight (110 kg) is a likely value of the weight.

In the context of regression analysis by OLS, Fox (2020, p.41) notes that the combination of high leverage on the horizontal axis and the unusual size of residuals on the vertical axis exerts influence on the regression coefficients. In other words, influence is a function of unusualness with respect to both horizontal and vertical axes in the sense of the scatter plot (McClendon, 1994, p.52; Bonate, 2011, pp.73–74). These influential observations are the kinds of outliers against which the current study proposes a robust ratio estimator. See Sect. 7.3 for concrete examples.

## 4 Traditional robust ratio estimators

This section briefly surveys the traditional methods of robust ratio estimators, against which the performance of the proposed method will be tested in Sects. 7 and 8.

### 4.1 Ratio of medians

$\widehat{\beta}_{\text{ratio}}$ is estimated by the ratio of arithmetic means. As a well-known fact, the arithmetic mean is sensitive to outliers while the median is insensitive to outliers. It is natural that the median is commonly used as an outlier robust measure of location (de Waal et al., 2011, p.210). Therefore, if we replace $\widehat{\beta}_{\text{ratio}}$ by $\widehat{\beta}_{\text{med}}$ in Eq. (7), it is the ratio-of-medians estimator, where med(•) denotes the median.

$$\widehat{\beta}_{\text{med}} = \frac{\text{med}\left(y_i\right)}{\text{med}\left(x_i\right)}. \tag{7}$$

### 4.2 Ratio of trimmed means

The trimmed mean is also one of the commonly used outlier robust measures of location (de Waal et al., 2011, p.211). While the median is insensitive to outliers, the median is inefficient because it utilizes information from very few observations. The trimmed mean can be regarded as a compromise between the arithmetic mean and the median (DeGroot & Schervish, 2002, p.579).

Let $y_1, y_2, \ldots, y_n$ be a random sample of size $n$, which satisfies the following condition: $y_1 < y_2 < \cdots < y_n$. Also, let $k$ be a positive integer such that $k < n/2$. Suppose that we delete from the data the $k$ smallest observations $y_1, y_2, \ldots, y_k$ and the $k$

largest observations $y_{n-k+1}, \ldots, y_{n-1}, y_n$. Then, the average of the remaining $n-2k$ middle observations is the $k$-th level trimmed mean, which is Eq. (8) (DeGroot & Schervish, 2002, p.578).

$$\overline{y}_{\text{trim}} = \frac{1}{n-2k} \sum_{i=k+1}^{n-k} y_i. \tag{8}$$

Therefore, if we replace $\widehat{\beta}_{\text{ratio}}$ by $\widehat{\beta}_{\text{trim}}$ in Eq. (9), it is the ratio-of-trimmed-means estimator, where $\overline{x}_{\text{trim}}$ is defined in a manner similar to $\overline{y}_{\text{trim}}$.

$$\widehat{\beta}_{\text{trim}} = \frac{\overline{y}_{\text{trim}}}{\overline{x}_{\text{trim}}}. \tag{9}$$

### 4.3 Ratio of Winsorized means

The Winsorized mean is also one of the commonly used outlier robust measures of location (de Waal et al., 2011, p.211). In Winsorization, rather than deleting the lowest and largest $k$ values, as done by the $k$th level trimmed mean, they are set equal to the smallest or largest value not trimmed (Mair & Wilcox, 2020, p.465). Again, let $y_1, y_2, \ldots, y_n$ be a random sample of size $n$, which satisfies the following condition: $y_1 < y_2 < \cdots < y_n$. Also, let $k$ be a positive integer such that $k < n/2$. Suppose that we set the $k$ smallest observations $y_1, y_2, \ldots, y_k = y_k, y_k, \ldots, y_k$ and the $k$ largest observations $y_{n-k+1}, \ldots y_{n-1}, y_n = y_{n-k+1}, \ldots y_{n-k+1}, y_{n-k+1}$. Then, the Winsorized mean is Eq. (10).

$$\overline{y}_{\text{winsor}} = \frac{1}{n}\left( k y_k + \sum_{i=k+1}^{n-k} y_i + k y_{n-k+1} \right). \tag{10}$$

Therefore, if we replace $\widehat{\beta}_{\text{ratio}}$ by $\widehat{\beta}_{\text{winsor}}$ in Eq. (11), it is the ratio-of-Winsorized-means estimator, where $\overline{x}_{\text{winsor}}$ is defined in a manner similar to $\overline{y}_{\text{winsor}}$.

$$\widehat{\beta}_{\text{winsor}} = \frac{\overline{y}_{\text{winsor}}}{\overline{x}_{\text{winsor}}}. \tag{11}$$

### 4.4 Ratio of means by *M*-estimators

Rather than deleting a fixed amount of data or setting it to one value, $M$-estimators provide a more flexible method to deal with outliers, where $M$ stands for maximum likelihood type, because $M$-estimators are found by maximizing a function that might not be the likelihood (DeGroot & Schervish, 2002, pp. 579–581; Mair & Wilcox, 2020, pp. 465–466).

In the context of regression analysis based on OLS, the sum of squared errors is a weighted average of errors, where the weights are their own values. The idea behind $M$-estimators is to replace these weights by some weights that do not keep growing

in magnitude as the errors grow. Essentially, robust methods such as $M$-estimators give less weights to observations with larger residuals (Kennedy, 2003, p. 375; Wooldridge, 2020, p. 323).

To find $M$-estimators, we often need to use the method of iteratively reweighted least squares (IRLS) (Mulry et al., 2014, p. 727). In the current study, based on Wada and Sakashita (2017, p. 3), whose method was applied to the 2016 Japanese Economic Census, we replace $\widehat{\beta}_{\text{ratio}}$ by $\widehat{\beta}_{\text{IRLS}}$ in Eq. (12), where $w_i$ is Tukey's biweight function defined in Eq. (13), $e_{r,i}$ is in Eq. (6), and $\psi$ is an arbitrary constant ranging from 4 (more robust) to 8 (less robust). For the choice of $\psi$, see Wada and Tsubaki (2020, p.3). For more information on a robust ratio estimator by $M$-estimators (IRLS), also see Gwet and Rivest (1992), Pannekoek (2018), Wada (2020), and Wada et al. (2021).

$$\widehat{\beta}_{\text{IRLS}} = \frac{\sum w_i y_i}{\sum w_i x_i}, \tag{12}$$

$$w_i = \begin{cases} \left[1 - \left(\frac{e_{r,i}}{\psi}\right)^2\right]^2 & \text{if } |e_{r,i}| \le \psi \\ 0 & \text{if } |e_{r,i}| > \psi \end{cases}. \tag{13}$$

## 5 Cook's distance for ordinary least squares (OLS)

This section briefly reviews the mechanism of Cook's distance for OLS. If we want to know whether an observation is influential or not, then an obvious way is to delete an observation one at a time and to recalculate how parameter estimates change. Apparently, this requires an iterative procedure such as IRLS; however, there is a method that can directly assess the influence of the $i$th observation with no iterations (Bonate, 2011, p.75). This method is Cook's distance (Cook, 1977), which is originally a composite score that evaluates an observation's influence on a set of regression parameters in the context of OLS (McClendon, 1994, p.107; Bonate, 2011, p.76).

Specifically, Cook's distance $C_i$ is shown in Eq. (14) (Cook, 1977, p.16; Fox, 2020, p. 49), where $p$ is the number of parameters in the model, and $e'_i$ is the studentized residual in Eq. (15), which deals with outliers on the vertical axis. Note that $e_i$ is the OLS residual in Eq. (4), and $s$ is the standard error of the regression (Wooldridge, 2020, pp. 49–50) defined in Eq. (16). Note that $s$ is also called an estimate of the error standard deviation, depending on academic fields. Also, $h_i$ is the hat value in Eq. (17), which deals with the leverage on the horizontal axis (Fox, 2020, p.45).

$$C_i = \frac{e'^2_i}{p} \times \frac{h_i}{1 - h_i}, \tag{14}$$

$$e'_i = \frac{e_i}{s\sqrt{1-h_i}}, \tag{15}$$

$$s = \sqrt{\frac{\sum e_i^2}{n-p}}, \tag{16}$$

$$h_i = \frac{1}{n} + \frac{\left(x_i - \bar{x}\right)^2}{\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2}. \tag{17}$$

Therefore, if an observation has a large value of Cook's distance, this means that the observation is influential in the OLS regression model, in terms of the vertical axis (measured by $e'^2_i$), the horizontal axis (measured by $h_i$), or the combination of both.

## 6 Algorithm of the TC-ratio estimator

### 6.1 Extending Cook's distance to the ratio estimator

This section presents how Cook's distance can be extended to the ratio estimator as the TC-ratio estimator. First, we estimate $\widehat{\beta}_{\text{ratio}}$ in Eq. (5) as if there were no outliers. Second, we calculate $e_{r,i}$, the residual of the ratio estimator in Eq. (6). As we saw in Sect. 2, $\widehat{\beta}_{\text{ratio}}$ is a weighted least squares estimate, where the weight is $1/\sqrt{x_i}$. Therefore, Eq. (6) is different from the OLS residual in Eq. (4), because we need to take the weight into account. Equation (6) is the key component of extending Cook's distance to the ratio estimator. Third, we calculate the studentized residual $e'_{r,i}$ in Eq. (18), where $e_{r,i}$ is the residual of the ratio estimator in Eq. (6) and $s_r$ is the standard error of the regression (Wooldridge, 2020, pp.49–50) for the ratio model in Eq. (19), where $p = 1$, because there is only one parameter in the ratio model. Again, note that $s_r$ is also called an estimate of the error standard deviation, depending on academic fields. Also, note that, as long as $p$ is the number of parameters in the model, the formula for the standard error of the regression is the same with or without an intercept (Eisenhauer, 2003, p.78). Note that $e_i$ is the residuals for the OLS regression model defined in Eq. (4), while $e_{r,i}$ is the residuals for the ratio model defined in Eq. (6). These residuals are the key difference between $s$ in Eq. (16) and $s_r$ in Eq. (19).

$$e'_{r,i} = \frac{e_{r,i}}{s_r\sqrt{1-h_i}}, \tag{18}$$

$$s_r = \sqrt{\frac{\sum e_{r,i}^2}{n-p}}. \tag{19}$$

Fourth, we calculate Cook's distance $C_{r,i}$ in Eq. (20), where $p = 1$, $e'_{r,i}$ is the studentized residual in Eq. (18), and $h_i$ is exactly the same as Eq. (17). Therefore, as was the case with Cook's distance for OLS, if an observation has a large value of $C_{r,i}$, this means that the observation is influential in the ratio model, in terms of the vertical axis (measured by $e'^2_{r,i}$), the horizontal axis (measured by $h_i$), or the combination of both.

$$C_{r,i} = \frac{e'^2_{r,i}}{p} \times \frac{h_i}{1 - h_i}. \tag{20}$$

Based on the values of $C_{r,i}$, we trim the identified outliers, where outliers are defined as large values of $C_{r,i}$. Let $D_i = \left(x_i, y_i\right)$ be a random sample of size $n$ $(i = 1, 2, \ldots, n)$. Also, let $\lambda$ and $k$ be positive integers. When $C_{r,j} > \lambda$, trim $D_j$, and when $C_{r,j} \leq \lambda$, do not trim $D_j$, where $j$ means the $j$-th observation. Suppose that we trim $k$ observations from the data. This means that we have $D_{tcr,i} = \left(x_{tcr,i}, y_{tcr,i}\right)$, where $i = 1, 2, \ldots, n - k$ and the subscript $tcr$ stands for TC-ratio. Then, the average of the remaining $n - k$ observations is the $k$th level trimmed mean based on the robust TC-ratio estimator. Thus, $\widehat{\beta}_{tcr}$ is given in Eq. (21), $\overline{y}_{tcr}$ in Eq. (22) and $\overline{x}_{tcr}$ in Eq. (23).

$$\widehat{\beta}_{tcr} = \frac{\overline{y}_{tcr}}{\overline{x}_{tcr}}, \tag{21}$$

$$\overline{y}_{tcr} = \frac{1}{n-k} \sum_{i=1}^{n-k} y_{tcr,i,} \tag{22}$$

$$\overline{x}_{tcr} = \frac{1}{n-k} \sum_{i=1}^{n-k} x_{tcr,i}. \tag{23}$$

Finally, we compute the imputed values based on $\widehat{y}_i = \widehat{\beta}_{tcr} x_i$, which is the robust ratio imputation model based on the TC-ratio estimator. This estimator is expected to work better than the traditional approaches, such as the ratio of medians, trimmed means, and Winsorized means, because the proposed estimator can detect both univariate (unconditional) and bivariate (conditional) outliers, while the traditional approaches (ratio of medians, trimmed means, and Winsorized means) can only detect univariate (unconditional) outliers.

### 6.2 Automatic method to determine the number of outliers

In the previous section, $\lambda$ is defined as a positive integer that is to be used as a cut-off to determine whether an observation is an outlier or not. In general, there is "no clear guidance on the percentage of trimming to be done" (Young & Mathew, 2015, p.78). Therefore, the choice of $\lambda$ is often arbitrary.

Traditionally, a cutoff for Cook's distance is proposed as $\lambda_{\text{cook}} = 4/(n-p)$, where $n$ is the number of observations and $p$ is the number of parameters (Fox, 2020, p.51). However, this cutoff is so simple that it does not take into account the characteristics of data in hand, because it is only a function of the number of observations and parameters. The current study proposes a novel method of automatically determining a cutoff based on the coefficient of determination $R^2$, by exploiting the fact that deleting an outlier is likely to increase $R^2$, because the model will have a better fit to the remaining data. Also, this study suggests a scree-like plot to graphically assess where the cutoff can be found.

For illustration purposes, we will use the following small dataset in Table 1, where the last two observations (id = 41 and 42) are added to the data as outliers. Based on the values of $C_{r,i}$, the data are sorted in an increasing order.

Since we have 42 observations and there is only one parameter, $\lambda_{\text{cook}} = 4/(n-p) = 4/(42-1) = 0.098$. In this case, we detect ID 38, 39, 40, 41, and 42 as outliers. Alternatively, since $C_{r,i}$ is univariate, we may try using the common measure of univariate outliers, i.e., $\text{UL} = Q_3 + 1.5 \times \text{IQR}$, where UL is the upper limit, $Q_3$ is the third quartile, and IQR is the inter-quartile range (Weiss, 2005, p.122). $Q_3$ in $C_{r,i}$ is 0.013 and IQR in $C_{r,i}$ is 0.012; thus, $\text{UL} = 0.013 + 1.5 \times 0.012 = 0.031$. We detect ID 36, 37, 38, 39, 40, 41, and 42 as outliers. Either way, we detect too many observations as outliers. These simple methods do not work, because they do not take the characteristics of data into account.

If we calculate $R^2$ in $\widehat{y}_i = \widehat{\beta} x_i$ among the 42 observations in Table 1, $R^2 = 0.518$. Note that we are not interested in interpreting the model fit per se, but we are interested in how the model fit changes when we delete an observation with large $C_{r,i}$. Let $R^2_k$ be the coefficient of determination when we trim the $k$ largest observations, where "largest" refers to the size of $C_{r,i}$. Since $C_{r,42} = 0.529$ is the largest value of Cook's distance, if we trim observation 42 from the data, $R^2_1 = 0.605$, which is larger than $R^2 = 0.518$ by 0.086. Since $C_{r,41} = 0.527$ is the second largest, if we trim observation 41 from the data, $R^2_2 = 0.798$, which is larger than $R^2_1 = 0.605$ by 0.194. Since $C_{r,40} = 0.239$ is the third largest, if we trim observation 40 from the data, $R^2_3 = 0.809$, which is larger than $R^2_2 = 0.798$ by 0.011. Since $C_{r,39} = 0.143$ is the fourth largest, if we trim observation 39 from the data, $R^2_4 = 0.815$, which is larger than $R^2_3 = 0.809$ by 0.006. We can continue this process until the last two observations are left. See $R^2_k$ in Table 1.

Naturally, $R^2_k$ tends to go up, as we trim more and more outliers. However, the speed of growth in $R^2_k$ decreases as we trim outliers. This can be used as a method of determining where we should stop trimming outliers. Notice that the increase was 0.086, 0.194, 0.011, and 0.006, which means that, after trimming the two largest outliers, the speed of growth in $R^2_k$ dramatically decreased. See $R^2_k - R^2_{k+1}$ in Table 1, which is the difference between the two adjacent $R^2_k$.

Graphically, the left-hand panel in Fig. 1 plots $1/R^2_k$ based on 42 observations in Table 1 against the number of trimmed observations $k$. Figure 1 is analogous to the scree plot in principal component analysis (Bartholomew et al., 2002, pp.124–125). In the left-hand panel of Fig. 1, there is an elbow at two trimmed

**Table 1** Example dataset to illustrate how a cutoff can be automatically found

| Id | $y_i$ | $x_i$ | $C_{r,i}$ | $k$ | $R_k^2$ | $R_k^2 - R_{k+1}^2$ | $1/R_k^2$ |
|---|---|---|---|---|---|---|---|
| 1 | 32.0 | 16.5 | 0.000008 | 0 | 0.518 | | 1.929 |
| 2 | 43.2 | 23.2 | 0.000009 | 1 | 0.605 | *− 0.086* | 1.654 |
| 3 | 28.3 | 15.4 | 0.000022 | 2 | 0.798 | *− 0.194* | 1.253 |
| 4 | 69.9 | 33.9 | 0.000166 | 3 | 0.809 | − 0.011 | 1.236 |
| 5 | 92.2 | 43.2 | 0.000409 | 4 | 0.815 | − 0.006 | 1.227 |
| 6 | 58.6 | 35.5 | 0.000409 | 5 | 0.809 | 0.006 | 1.236 |
| 7 | 47.2 | 21.3 | 0.000599 | 6 | 0.801 | 0.008 | 1.248 |
| 8 | 83.0 | 37.1 | 0.000751 | 7 | 0.812 | − 0.011 | 1.231 |
| 9 | 82.4 | 51.3 | 0.000875 | 8 | 0.830 | − 0.017 | 1.205 |
| 10 | 29.9 | 20.4 | 0.001136 | 9 | 0.827 | 0.003 | 1.209 |
| 11 | 63.9 | 27.0 | 0.001348 | 10 | 0.828 | − 0.001 | 1.208 |
| 12 | 42.0 | 30.7 | 0.001790 | 11 | 0.839 | − 0.012 | 1.192 |
| 13 | 62.7 | 44.7 | 0.001949 | 12 | 0.842 | − 0.003 | 1.188 |
| 14 | 24.8 | 20.5 | 0.002867 | 13 | 0.855 | − 0.013 | 1.169 |
| 15 | 17.9 | 15.4 | 0.003021 | 14 | 0.867 | − 0.012 | 1.153 |
| 16 | 1.6 | 2.9 | 0.003170 | 15 | 0.873 | − 0.006 | 1.145 |
| 17 | 36.8 | 32.6 | 0.003797 | 16 | 0.885 | − 0.012 | 1.130 |
| 18 | 92.0 | 34.2 | 0.004012 | 17 | 0.865 | − 0.009 | 1.118 |
| 19 | 126.3 | 48.6 | 0.004378 | 18 | 0.904 | − 0.009 | 1.106 |
| 20 | 18.9 | 19.4 | 0.015095 | 19 | 0.914 | − 0.010 | 1.094 |
| 21 | 76.6 | 26.9 | 0.015569 | 20 | 0.915 | − 0.001 | 1.093 |
| 22 | 25.8 | 27.8 | 0.015902 | 21 | 0.924 | − 0.010 | 1.082 |
| 23 | 139.0 | 51.8 | 0.016299 | 22 | 0.929 | − 0.005 | 1.077 |
| 24 | 14.2 | 20.1 | 0.018529 | 23 | 0.936 | − 0.007 | 1.069 |
| 25 | 62.8 | 57.3 | 0.018721 | 24 | 0.940 | − 0.004 | 1.064 |
| 26 | 15.9 | 22.8 | 0.018847 | 25 | 0.949 | − 0.010 | 1.053 |
| 27 | 37.6 | 44.8 | 0.018872 | 26 | 0.957 | − 0.007 | 1.045 |
| 28 | 95.8 | 30.8 | 0.019223 | 27 | 0.960 | − 0.003 | 1.042 |
| 29 | 16.9 | 28.5 | 0.010692 | 28 | 0.964 | − 0.004 | 1.038 |
| 30 | 16.9 | 31.3 | 0.011711 | 29 | 0.968 | − 0.005 | 1.033 |
| 31 | 156.3 | 52.1 | 0.012599 | 30 | 0.973 | − 0.005 | 1.027 |
| 32 | 10.1 | 24.1 | 0.013503 | 31 | 0.980 | − 0.006 | 1.021 |
| 33 | 86.8 | 73.2 | 0.026459 | 32 | 0.982 | − 0.002 | 1.019 |
| 34 | 194.2 | 68.7 | 0.021497 | 33 | 0.985 | − 0.003 | 1.015 |
| 35 | 39.4 | 63.5 | 0.030814 | 34 | 0.990 | − 0.005 | 1.010 |
| 36 | 58.3 | 73.4 | 0.039814 | 35 | 0.991 | − 0.001 | 1.009 |
| 37 | 158.3 | 113.2 | 0.058631 | 36 | 0.991 | − 0.001 | 1.009 |
| 38 | 280.4 | 102.7 | 0.098627 | 37 | 0.993 | − 0.002 | 1.007 |
| 39 | 104.1 | 106.2 | 0.143047 | 38 | 0.998 | − 0.005 | 1.002 |
| 40 | 90.6 | 110.4 | 0.238638 | 39 | 1.000 | − 0.002 | 1.000 |
| *41* | *294.5* | *13.0* | *0.526751* | 40 | 1.000 | − 0.000 | 1.000 |
| *42* | *314.3* | *18.6* | *0.529118* | 41 | 1.000 | − 0.000 | 1.000 |

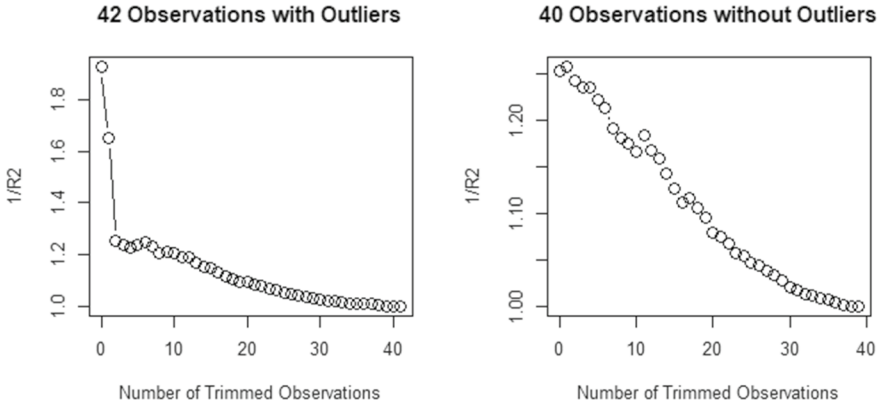


**Fig. 1** Examples of the scree-like plot to detect the number of potential outliers for Table 1

observations. This means that trimming the rest of observations has similar $R_k^2$, which further means that they each explain a similar proportion of the total variance of $y_i$. Therefore, graphically, we can decide that there are two outliers in the data. On the other hand, if Table 1 did not have ID 41 and 42 in the first place, and if we calculate $C_{r,i}$ and $R_k^2$ based on the first 40 observations, the scree-like plot would be the right-hand panel in Fig. 1, which shows no elbows, meaning that there are no outliers in the data.

Next, we locally calculate the vertical distance from one dot to another in Fig. 1, so that we numerically and automatically decide the number of outliers. This is done by calculating $R_k^2 - R_{k+1}^2$. When these vertical distances, $R_k^2 - R_{k+1}^2$, are close enough to zero, then we trimmed enough outliers in the data. Most of these values are close to zero. In fact, the mean is $-0.012$, the median is $-0.005$, the first quartile ($Q_1$) is $-0.010$, and IQR is 0.008. Since $R_k^2 - R_{k+1}^2$ is univariate, we can simply use the common measure of univariate outliers, i.e., $\mathrm{LL} = Q_1 - 1.5 \times \mathrm{IQR}$, where LL is the lower limit and $Q_1$ is the first quartile (Weiss, 2005, p.122). Therefore, $\mathrm{LL} = -0.010 - 1.5 \times 0.008 = -0.021$. Since $R_0^2 - R_1^2 = -0.086$ and $R_1^2 - R_2^2 = -0.194$ are smaller than $\mathrm{LL} = -0.021$, we can numerically and automatically decide that outliers are up to the second largest $C_{r,i}$. On the other hand, if Table 1 did not have ID 41 and 42, and if we calculate $C_{r,i}$ and $R_k^2$ based on the first 40 observations, $\mathrm{LL} = -0.018$. None of $R_k^2 - R_{k+1}^2$ would be smaller than $-0.018$; thus, we numerically and automatically conclude that there are no outliers in the data.

Additionally, in the actual implementation, the moving average of order 3 is used to avoid haphazard idiosyncrasies (large jump from $k$ to $k+1$).

Therefore, it is demonstrated in this subsection that we can determine the number of outliers based on the speed of change in $R_k^2$. This mechanism allows the TC-ratio estimator to be fully automated in the process of outlier detection, because no processes involve human decisions.

## 7 Monte Carlo simulation: settings

Monte Carlo simulation is useful especially when assumptions of a model are violated, but there are no easy analytical solutions available (Mooney, 1997, p.1). Analyses in this study are carried out using R version 4.0.2. In this simulation study, the sample size $n$ is set to 1000, and the number of simulation runs is set to 10,000. Since the means and totals are considered the most important products in official statistics (de Waal et al., 2011, p.245), the parameter of interest in the simulations is set to the mean of a target variable, $\bar{y}$.

### 7.1 Settings of population data

The Monte Carlo simulations are carried out using five different artificially generated populations of values $(x_i, y_i)$, whose values are generated by a gamma distribution, a normal distribution, or a uniform distribution.

A random variable $X$ follows a gamma distribution with parameters $\phi$ and $\omega$, where $x > 0$, $\phi > 0$, $\omega > 0$ if its density function is given by Eq. (24), and $\Gamma(\phi)$ is the gamma function defined in Eq. (25). Also, from Eqs. (26) and (27), the mean is $\phi\omega$ and the variance is $\phi\omega^2$ (DeGroot & Schervish, 2002, p.297; Ross, 2006, pp.237–239). A gamma distribution is one of the commonly used population settings for ratio imputation (Lee et al., 1994, p.236; Rao & Sitter, 1995, p.455; Sitter & Rao, 1997, p.69; Haziza & Valée, 2020).

$$f(X = x) = \frac{1}{\Gamma(\phi)\omega^{\phi}} x^{\phi-1} \exp\left(\frac{-x}{\omega}\right), \tag{24}$$

$$\Gamma(\phi) = \int_0^{\infty} x^{\phi-1} \exp(-x) dx, \tag{25}$$

$$E(X) = \int_0^{\infty} \frac{\phi\omega}{\Gamma(\phi+1)} \left(\frac{x}{\omega}\right)^{\phi} \exp\left(\frac{-x}{\omega}\right) \frac{1}{\omega} dx = \phi\omega, \tag{26}$$

$$\mathrm{var}(X) = \phi\omega^2(\phi+1) - \phi^2\omega^2 = \phi\omega^2. \tag{27}$$

Specifically, a set of 1000 $x$-values are generated by a gamma distribution with mean $\phi\omega = 48$ and variance $\phi\omega^2 = 768$. Then, for each fixed value of $x$, the corresponding value of $y$ is generated by a gamma distribution with mean $\mu_y = bx$ and variance $\sigma_y^2 = d^2x^{2g}$, where the values of $b$, $d$, and $g$ are shown in Table 2. Also, $\rho$ is the correlation between $x$ and $y$, and $\mu_y$ is the true population value of $\bar{y}$. This follows the population settings used in Lee et al., (1994, p.236). Also, the online appendix A reports additional simulation runs based on a gamma distribution with mean $\phi\omega = 24$ and variance $\phi\omega^2 = 768$, where the expected value of $X$ is set to half.

**Table 2** Characteristics of the three populations (gamma distribution)

| Population | $b$ | $d$ | $g$ | $\rho$ | $\mu_y$ |
|---|---|---|---|---|---|
| 1 | 1.5 | 1.84 | 0.75 | 0.75 | 72 |
| 2 | 1.5 | 5.13 | 0.50 | 0.75 | 72 |
| 3 | 1.5 | 13.78 | 0.25 | 0.75 | 72 |

These specific values of $b$, $d$, $g$, ρ, and $\mu_y$ are based on Lee et al., (1994, p.236)

Lee et al., (1994, p.236) show that $\phi$ and $\omega$ can be defined as Eqs. (28) and (29).

$$\phi = \frac{(bx)^2}{d^2 x^{2g}}, \tag{28}$$

$$\omega = \frac{d^2 x^{2g}}{bx}. \tag{29}$$

Therefore, the relation between $x_i$ and $y_i$ can be adequately captured by the ratio estimator model $y_i = \beta x_i + \varepsilon_i$, where $\beta = 1.5$ ($b = 1.5$ in Table 2) and $\varepsilon_i \sim N\left(0, \sigma^2 \sqrt{x_i}\right)$. Also, the online appendix B reports additional simulation runs based on $\beta = 3.0$, where the true ratio is set to double.

Sections 8.1 and 8.2 display the results for population 1. The results for populations 2 and 3 can be found in the Appendix (Sects. 11.1 and 11.2). Furthermore, in discussing the ratio estimator, some authors (Zou et al., 2010, p.871; Wada & Sakashita, 2017, p.3) assume that $x$-values are generated by a uniform distribution, and some authors (Zou et al., 2010, p.871; Lui, 2020, p.140) assume that $x$-values are generated by a normal distribution. Therefore, to make the simulations more general (free of distributional assumptions), the Appendix has extra results for population 4 (uniform distributions) and population 5 (normal distributions).

Under population 4, a set of 1000 $x$-values are generated by a uniform distribution $U(0.1, 2.1)$. Under population 5, a set of 1,000 $x$-values are generated by a normal distribution $N(20, 16)$. Since $x$-values must be positive for the ratio estimator, in case that $x$-values are generated as negative, they are replaced by the minimum value among the positive $x$-values. In both populations 4 and 5, $y_i = 3.9x_i + \sqrt{x_i}\varepsilon_i$, where $\varepsilon_i \sim N(0, 1)$. All of these settings for populations 4 and 5 follow the simulation studies by Zou et al., (2010, p.871), slightly changing $x_i\varepsilon_i$ to $\sqrt{x_i}\varepsilon_i$, because their simulations assume the population for the mean of ratios, not the ratio of means which the current study assumes.

### 7.2 Settings of missing data

Let $y_i$ be the target incomplete variable for imputation, $x_i$ be completely observed in all of the situations to be used as the auxiliary variable, and $u_{1,i}$ and $u_{2,i}$ be two continuous uniform random variables ranging from 0 to 1 for the missingness mechanism. This means that missing occurs in $y_i$, the numerator in the ratio, $\widehat{\beta}_{\text{ratio}} = \overline{y}/\overline{x}$.

Each of the artificially generated datasets is made incomplete using the following two types of missing data generation processes based on missing at random (MAR), where the missingness of $y_i$ depends on the values of $x_i$, $u_{1,i}$, and $u_{2,i}$, i.e., the conditional probability of missing data after controlling for observed data is the same as the probability of observed data (Allison, 2002, p.4; Enders, 2010, p.11; Little & Rubin, 2020, p.14). The average missing rates are set to 30%. It is reported that the family incomes and personal earnings in the National Health Interview Survey (1997–2004) have approximately 30% of missingness (Schenker et al., 2006, p.925). Therefore, 30% is a realistic value as a missing rate. Note that, while any specific real survey may have different rates of missingness, the specific settings on the missing rates should not be much of a concern. The average missing rates are 30% under 10,000 simulation runs, which means that some simulation runs have missingness less than 30%, and other simulation runs have missingness more than 30%. On average across 10,000 runs, it is 30%. Therefore, this setting is supposed to cover a reasonable range of missing rates.

In the first type of missing data generation process under MAR, $y_i$ is missing if $x_i < \text{med}(x_i)$ and $u_{1,i} < 0.5$, and $y_i$ is missing if $x_i > \text{med}(x_i)$ and $u_{2,i} < 0.1$, where med(•) denotes the median. For example, suppose that $y_i$ is turnover (sales) and $x_i$ is the number of employees. The assumption in this setting is that more values are missing among small-and-medium size enterprises than large enterprises, because the missing values of turnover for large enterprises are collected through recontacts in official statistics. Therefore, $\overline{y}$ based on missing data overestimates the true value of $\overline{y}$. Let us call this MAR1.

In the second type of missing data generation process under MAR, $y_i$ is missing if $x_i < \text{med}(x_i)$ and $u_{1,i} < 0.1$, and $y_i$ is missing if $x_i > \text{med}(x_i)$ and $u_{2,i} < 0.5$. Again, for example, suppose that $y_i$ is turnover and $x_i$ is the number of employees. The assumption in this setting is that large enterprises are more likely to refuse to answer turnover than small-and-medium size enterprises, possibly because of some tax-related concerns. Therefore, $\overline{y}$ based on missing data underestimates the true value of $\overline{y}$. Let us call this MAR2.

Both of these two scenarios intuitively sound plausible and we do not expect, a priori, which of the scenarios is more realistic in a given survey of official statistics. Therefore, we use these two types of missing data scenarios. These two types of missing data can be understood as MAR via censoring.

Under the assumption of missing completely at random (MCAR), the probability of missing data does not depend on data, and observed data are a simple random sub-sample of complete data (Allison, 2002, p.3; Enders, 2010, p.7; Little & Rubin, 2020, p.13). Since MAR is a "less restrictive assumption than MCAR" (Little & Rubin, 2020, p.14), in reality, it is safer that we assume MAR rather than MCAR. This takes us back to the case of MAR. Therefore, the current study does not consider the assumption of MCAR.

Under the assumption of not missing at random (NMAR, also known as missing not at random: MNAR), the missingness of $y_i$ depends on the values of $y_i$, $u_{1,i}$, and $u_{2,i}$, even after controlling for $x_i$, i.e., the conditional probability of missing data after controlling for observed data is not the same as the probability of observed data (Allison, 2002, p.5; Enders, 2010, p.11; Little & Rubin, 2020, p.14). Graham

(2009, p.567) states that all missing data are a continuum between pure MAR and pure NMAR. The current study focuses on the case of pure MAR, because the current study is concerned with the influence of outliers on the imputation model under the situation where the imputation model can eliminate the bias due to missing data. In the case of pure NMAR, the literature recommends the use of the selection model and the pattern mixture model (Allison, 2002, pp.77–84; Enders, 2010, pp.290–301; Little & Rubin, 2020, pp.351–355). How the TC-ratio estimator can be extended by way of the selection model or the pattern mixture model is left for future research. Nevertheless, Scheuren (2005, p.317) contends that, in official statistics, about 10–20% are MCAR, about 50% are MAR, and the rest is NMAR. Thus, the assumption of MAR may cover the majority (up to 70%) of the situations that we may encounter in official statistics.

### 7.3 Settings of outliers

Figure 2 in the current study graphically shows the patterns of 5% outlier settings, where white circles represent usual observations and red triangles represent outliers generated by our outlier model, which is described below. Outliers in $y_i$ follow $U\left(0.7\max\left[y_i\right], \max\left[y_i\right]\right)$, where the associated values of $x_i$ are less than $\text{med}\left(x_i\right)$. Outliers in $x_i$ follow $U\left(0.7\max\left[x_i\right], \max\left[x_i\right]\right)$, where the associated values of $y_i$ are less than $\text{med}\left(y_i\right)$.

Furthermore, the cases where outliers exist in both $x$ and $y$ can be divided into three patterns: equal percentage (50:50) in Fig. 2, less outliers in $x$ than in $y$ (25:75) and more outliers in $x$ than in $y$ (75:25) in Fig. 3. In official statistics, ratio imputation is applied to different subpopulations, which is known as group ratio imputation (de Waal et al., 2011, p.245). Some subpopulations may have outliers on the vertical axis, while other subpopulations may have outliers on the horizontal axis, or the combination of both.

Therefore, if a ratio imputation model is robust against outliers anywhere in the scatter plot, it will be beneficial.

The percentage of outliers is set to 1%, 5%, and 10%. This means that we will add 10 outliers to 1000 observations ($n = 1010$ in total for 1% outliers), 50 outliers to 1,000 observations ($n = 1050$ in total for 5% outliers), and 100 outliers to 1,000 observations ($n = 1100$ in total for 10% outliers). Note that these outliers will not be missing in the simulations, because we are interested in the influence of outliers on the parameter of the imputation model when outliers are indeed present in data. Also, the online appendix C reports additional simulation runs, where outliers are also missing.

Therefore, there are 10 types of data without outliers (5 population types and 2 missingness types) and 150 types of data with outliers (5 population types, 2 missingness types, 5 types of outlier locations, and 3 types of outlier percentages). Additionally, in the online appendices, we have 194 types of data. Each of these 354 types of data is repeated 10,000 times. Thus, we have 3,540,000 different types of data in total.

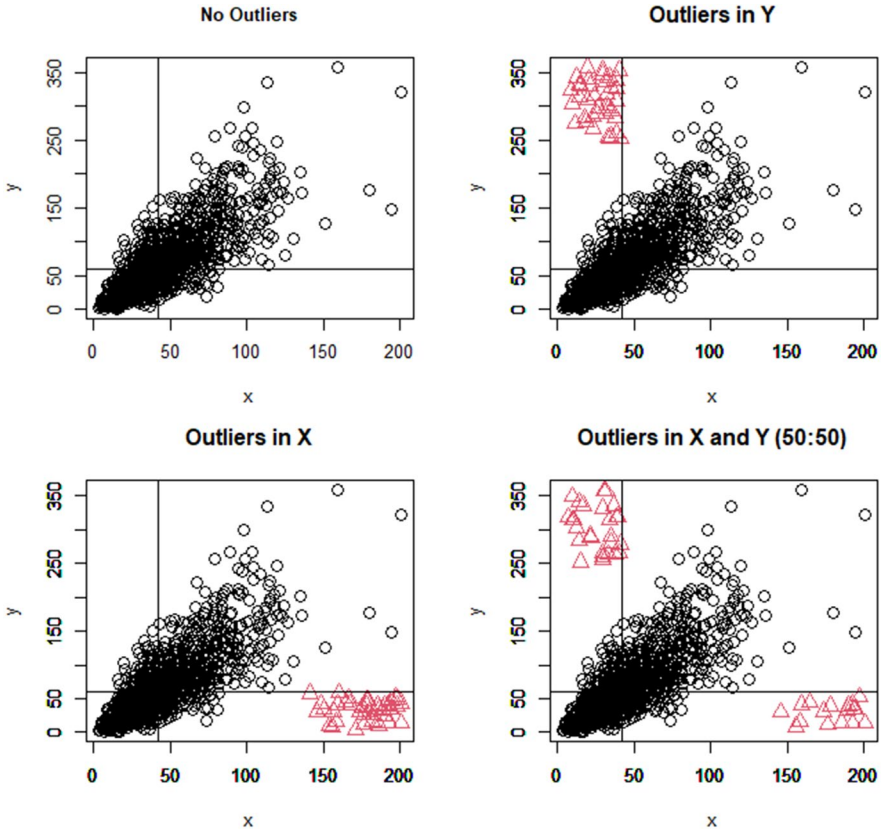


**Fig. 2** Examples of four outlier patterns in the simulations (population 1). White circles represent usual observations, red triangles represent outliers, the vertical line is med ($x$), and the horizontal line is med ($y$)

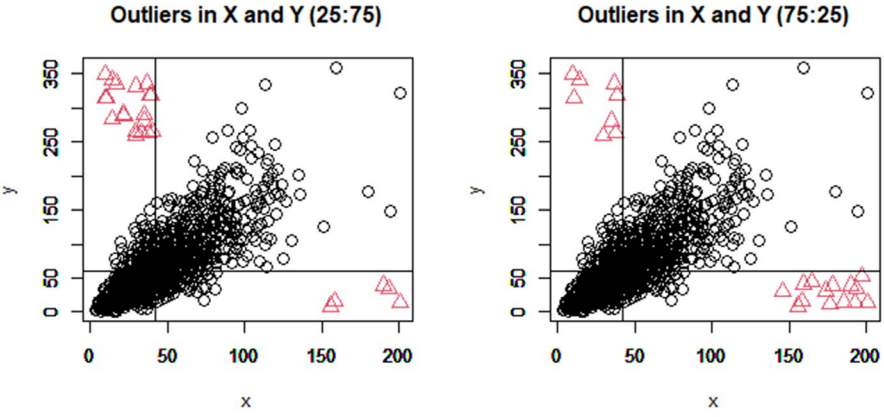


**Fig. 3** Examples of outliers for both axes in the simulations (population 1). White circles represent usual observations, red triangles represent outliers, the vertical line is med ($x$), and the horizontal line is med ($y$)

### 7.4 Evaluation criteria for simulations

Let $\theta$ be the true population parameter and $\widehat{\theta}$ be an estimator of $\theta$. If $\text{Bias}\left(\widehat{\theta}\right) = 0$ in Eq. (30), the expected value of $\widehat{\theta}$ is equal to the true $\theta$. Then, this estimator $\widehat{\theta}$ is an unbiased estimator of true parameter $\theta$ (Mooney, 1997, p.59; Gujarati, 2003, p.899). Therefore, $\text{Bias}\left(\widehat{\theta}\right)$ indicates whether the method is good on average.

$$\text{Bias}\left(\widehat{\theta}\right) = E\left(\widehat{\theta}\right) - \theta. \tag{30}$$

Oftentimes, however, there is a situation where one estimator has smaller bias and larger variance than another estimator. The root mean squared error (RMSE) in Eq. (31) measures the dispersion around the true value of the parameter, taking a balance between bias and efficiency into account (Mooney, 1997, p.59; Gujarati, 2003, p.901–902; Carsey & Harden, 2014, pp.88–89). Therefore, $\text{RMSE}\left(\widehat{\theta}\right)$ indicates whether the method is good across 10,000 runs, taking both bias and efficiency into account.

$$\text{RMSE}\left(\widehat{\theta}\right) = \sqrt{E\left(\widehat{\theta} - \theta\right)^2}. \tag{31}$$

Thus, an estimator $\widehat{\theta}$ in this study is considered good if it has $\text{Bias}\left(\widehat{\theta}\right)$ close to zero and $\text{RMSE}\left(\widehat{\theta}\right)$ close to zero.

### 7.5 Competing methods in the simulations

Table 3 displays the abbreviations about the competing methods used in the simulations. For each of the traditional robust ratio imputation models, see Sect. 4.

Comp is complete data, which are supposed to be ideal, but unavailable in reality. LD is listwise deletion, which throws away all of the rows that contain missing

**Table 3** List of the competing methods in the simulations

| Abbreviations | Methods |
|---|---|
| Comp | Complete data |
| LD | Listwise deletion |
| Ratio | Regular ratio imputation (non-robust) |
| M-1 | Ratio imputation by $M$-estimator IRLS ($\psi = 8$, less robust) |
| M-2 | Ratio imputation by $M$-estimator IRLS ($\psi = 4$, more robust) |
| Med | Ratio-of-medians imputation |
| Trim | Ratio-of-trimmed-means imputation (5% trimming) |
| Wins | Ratio-of-Winsorized-means imputation (5% Winsorizing) |
| C-1 | Ratio imputation by Cook's distance (proposed TC-ratio estimator) |
| C-2 | Ratio imputation by Cook's distance (cutoff $= 4/(n-p)$) |

values. This is the result we obtain if we do not deal with missing values. In the literature of missing data analysis, this method is also known as complete case analysis (Little & Rubin, 2020, pp.47–48). Both Comp and LD are not affected by outliers, because there are no imputation models.

Ratio is the non-robust ratio imputation model. Ratio is expected to work best among imputation methods when outliers are not present, while it is expected not to work well when outliers are present.

M-1 and M-2 are the ratio imputation models by $M$-estimators. These are the methods implemented in the 2016 Economic Census in Japan (Wada & Sakashita, 2017). This study chooses two values for the tuning constant $\psi$ that represent less robust ($\psi = 8$) and more robust ($\psi = 4$), respectively, because $\psi$ cannot be predetermined. For the information on the choice of $\psi$, see Wada and Tsubaki (2020, p.3).

Med is the ratio-of-medians imputation. Trim is the ratio-of-trimmed-means imputation. Wins is the ratio-of-Winsorized-means imputation. In trimming outliers, there are no clear rules about the percentage of trimming. We set 5% as a cutoff for the trimmed and the Winsorized means. Therefore, these two methods are expected to work well when outliers are 5% in the simulations, but work less well under 1% and 10%

C-1 is the proposed TC-ratio estimator, where outliers are detected by modified Cook's distance and the number of outliers is determined by the inverse $R_k^2$. C-2 uses $4/(n-p)$ as a cutoff.

### 7.6 Motivating example for simulation settings

Populations 1, 2, and 3 follow the simulation settings by Lee et al. (1994, p.236). A natural question is to ask whether these settings are realistic. As a real-world example, this subsection uses the anonymized data of the 2004 Japanese National Survey of Family Income and Expenditure, which is based on the actual microdata of the survey and is offered for the purpose of academic analyses. As of this writing, 2004 is the latest version of the anonymized data of the National Survey of Family Income and Expenditure.

Figure 4 displays the distributions of net expenditure and yearly income. Note that yearly income is measured on a yearly basis in the unit of 10,000 Japanese yen, while net expenditure is measured on a monthly basis in the unit of 1 Japanese yen. To make them comparable, net expenditure is divided by 10,000 and multiplied by 12, so that net expenditure is also on a yearly basis in the unit of 10,000 Japanese yen. Since these are sensitive real data from official statistics, for the purpose of disclosure limitation, the axes in Fig. 4 are intentionally hidden. Please pay attention to the shapes of the distributions, not the values of each data point.

Suppose that some values of yearly income are missing and all of the values of net expenditure are observed. Then, we may predict the missing values of $income_i$ by $\text{expenditure}_i$, using $\text{income}_i = \beta \times \text{expenditure}_i + \varepsilon_i$, where $\beta$ is estimated by the ratio of means. Note that the prediction in the imputation model does not require a causal specification (King et al., 2001, p.51), meaning that the imputation model

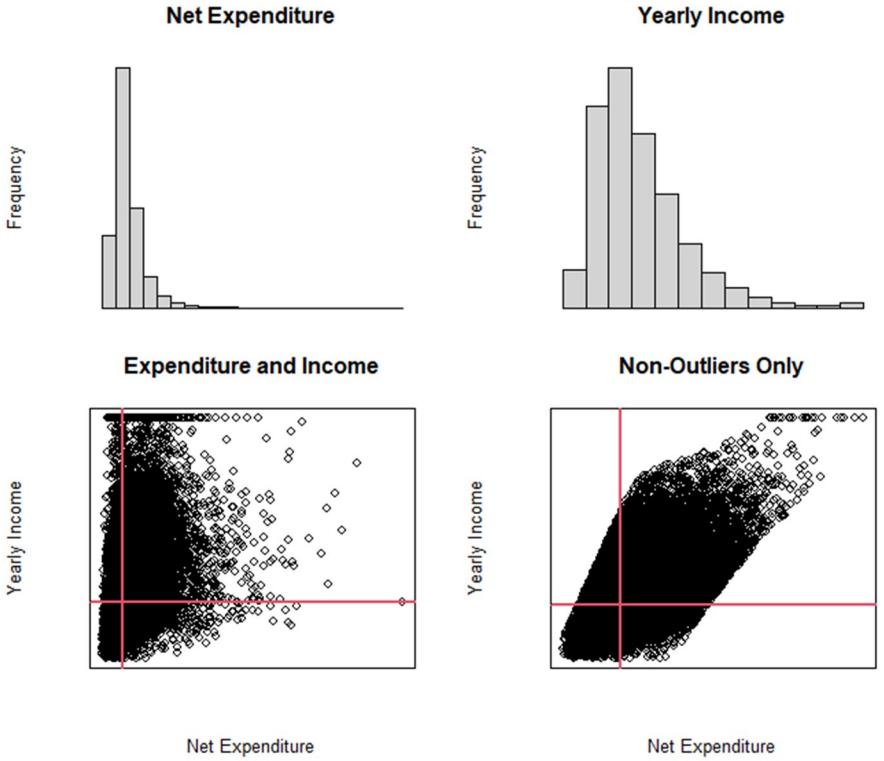


**Fig. 4** Characteristics of expenditure and income. Note that the vertical line is med (net expenditure) and the horizontal line is med (yearly income). Axes are intentionally hidden for disclosure limitation purposes

does not claim that $\text{expenditure}_i$ is the cause of $\text{income}_i$. It simply states that missing values of $\text{income}_i$ may be predicted by $\text{expenditure}_i$.

Figure 4 shows that each variable is skewed to the right, and the bivariate distribution is also heteroskedastic. The mean of income is 669.5 and the mean of expenditure is 379.6. Let $y$ be income and $x$ be expenditure. Then, $\beta = \overline{y}/\overline{x} = 669.5/379.6 = 1.76$, which is slightly higher than the value of $b$, defined in Table 2. The correlation between income and expenditure is 0.46, which is lower than $\rho$, defined in Table 2, but this is still coherent in the sense that the correlation is positive. One of the reasons why the correlation is low is that income is top coded at the value of 2500, which makes correlation lower than it actually must be. Note that the real data already contain some potential outliers. If we trim these potential outliers by the TC-ratio estimator, $\beta = 1.80$ and $\rho = 0.64$, which are quite close to the theoretically defined values in Table 2. Therefore, it is demonstrated in this subsection that the simulation settings above are realistic.

Note that, based on the Statistics Act (Japan), the author obtained the anonymized data of the 2004 National Survey of Family Income and Expenditure

from the National Statistics Center (NSTAC). Also, note that the analyses in this article are the author's own and are different from the officially published results by the Japanese government. For the information on the extended use (secondary use) of official statistics in Japan, see https://www.soumu.go.jp/english/dgpp_ss/seido/2jiriyou.htm.

## 8 Monte Carlo simulation: results

### 8.1 MAR1 in population 1

Table 4 presents the results of the simulations for population 1 (Gamma distribution, $d = 1.84$, $g = 0.75$) under MAR1.

Although there are no absolute criteria to judge the size of bias, Schafer and Graham (2002, p.157) state, "A rule of thumb that we have found useful is that bias becomes problematic if its absolute size is greater than about one half of the estimate's standard error." In Table 4, one standard error of the mean in complete data is about 1.7, which can be found in the column of Comp under RMSE, because RMSE is $\sqrt{\text{variance} + \text{bias}^2}$; thus, for an unbiased estimator, RMSE is the standard error. Therefore, if the absolute value of bias is smaller than $1.7/2 = 0.850$, then we deem the method unbiased and put it in italics.

Also, there are no absolute criteria to judge the size of RMSE, which is meaningful only in comparative terms (Carsey & Harden, 2014, p.89). The smallest RMSE indicates that the estimator is comparatively best among the competing estimators. Thus, the smallest value of RMSE is shown in italics for each outlier setting. Note that Comp is excluded from the comparison of RMSE, because Comp is always the best, but unavailable method.

Under all situations, listwise deletion (LD) is severely biased (bias = 9.005, 9.062). In fact, listwise deletion is always biased under MAR. Therefore, the task is to correct the bias of about 9 points by way of imputation.

When there are no outliers ($\%X = 0.00$, $\%Y = 0.00$), the regular ratio imputation model (ratio) is unbiased and most efficient (bias = − 0.012, RMSE = 1.852). All of the robust ratio imputation models are slightly more biased than the regular ratio imputation model, but most of them, except Med and C-2, can also correct the bias in listwise deletion within half of one standard error. Thus, we consider M-1, M-2, Trim, Wins, and C-1 unbiased, using the rule of thumb by Schafer and Graham (2002, p.157).

In the case of the equal number of outliers in both $x$ and $y$ ($\%X = 0.50$, $\%Y = 0.50$), the bias in the regular ratio imputation model is small (Bias = 0.261). However, as the percentages of outliers increase, the bias in the regular ratio imputation model becomes large (bias = 1.029 for $\%X = 2.50$, $\%Y = 2.50$; bias = 1.746 for $\%X = 5.00$, $\%Y = 5.00$).

In 13 out of 16 cases, the bias of the TC-ratio estimator (C-1) is smaller than half of one standard error. Most importantly, when the bias of the regular ratio imputation model is larger than half of one standard error in 13 cases, the bias of the

**Table 4** Population 1 (Gamma distribution: $d = 1.84$, $g = 0.75$), MAR1

Bias: |value| < 0.850 in italics

| %X | %Y | Comp | LD | Ratio | M-1 | M-2 | Med | Trim | Wins | C-1 | C-2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.00 | 0.00 | *− 0.017* | 9.022 | *− 0.012* | *− 0.303* | *− 0.721* | − 1.425 | *− 0.463* | *− 0.200* | *− 0.326* | − 0.885 |
| 0.00 | 1.00 | *− 0.019* | 9.034 | 1.641 | *− 0.016* | *− 0.152* | *− 0.830* | *0.577* | 1.123 | *0.011* | *0.076* |
| 0.00 | 5.00 | *− 0.009* | 9.029 | 7.888 | 4.001 | 1.222 | 1.859 | 6.823 | 8.087 | *0.386* | 4.014 |
| 0.00 | 10.00 | *− 0.011* | 9.026 | 14.957 | 11.086 | 6.781 | 5.904 | 14.654 | 15.438 | 9.911 | 10.288 |
| 1.00 | 0.00 | *− 0.010* | 9.043 | − 1.066 | − 1.023 | − 1.045 | − 2.008 | − 1.243 | − 1.127 | *− 0.398* | − 1.009 |
| 5.00 | 0.00 | *0.005* | 9.048 | − 3.999 | − 3.935 | − 3.505 | − 4.145 | − 4.254 | − 4.284 | *− 0.838* | − 1.085 |
| 10.00 | 0.00 | *0.017* | 9.051 | − 6.159 | − 6.281 | − 6.441 | − 6.061 | − 6.591 | − 6.447 | − 5.921 | − 3.479 |
| 0.25 | 0.75 | *0.014* | 9.058 | 0.954 | *− 0.265* | *− 0.316* | − 1.097 | *0.123* | *0.537* | *− 0.124* | *− 0.012* |
| 1.25 | 3.75 | *− 0.022* | 9.018 | 4.185 | *0.779* | *− 0.676* | *0.126* | 3.422 | 4.445 | *0.186* | 1.594 |
| 2.50 | 7.50 | *− 0.019* | 9.024 | 7.409 | 3.816 | *0.597* | 1.835 | 7.332 | 7.660 | *0.395* | 4.152 |
| 0.50 | 0.50 | *0.014* | 9.062 | *0.261* | *− 0.520* | *− 0.522* | − 1.396 | *− 0.341* | *− 0.040* | *− 0.242* | *− 0.125* |
| 2.50 | 2.50 | *− 0.027* | 9.012 | 1.029 | − 1.561 | − 1.782 | − 1.439 | *0.219* | 0.901 | *− 0.175* | *− 0.787* |
| 5.00 | 5.00 | *− 0.001* | 9.039 | 1.746 | − 1.359 | − 3.087 | − 1.427 | 1.045 | 1.674 | *− 0.534* | − 1.178 |
| 0.75 | 0.25 | *− 0.001* | 9.046 | *− 0.416* | *− 0.775* | *− 0.761* | − 1.706 | *− 0.801* | *− 0.595* | *− 0.334* | *− 0.320* |
| 3.75 | 1.25 | *− 0.035* | 9.005 | − 1.685 | − 3.032 | − 2.629 | − 2.882 | − 2.419 | − 2.249 | *− 0.374* | − 2.060 |
| 7.50 | 2.50 | *− 0.012* | 9.030 | − 2.652 | − 4.665 | − 4.855 | − 4.163 | − 3.673 | − 2.942 | − 0.907 | − 4.721 |

**Table 4** (continued)

RMSE: Best outcomes in italics for each outlier setting excluding Comp

| %X | %Y | Comp | LD | Ratio | M-1 | M-2 | Med | Trim | Wins | C-1 | C-2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.00 | 0.00 | 1.729 | 9.296 | *1.852* | 1.870 | 1.973 | 2.326 | 1.894 | 1.855 | 1.878 | 2.034 |
| 0.00 | 1.00 | 1.723 | 9.304 | 2.515 | *1.828* | 1.834 | 2.011 | 1.931 | 2.176 | 1.841 | 1.846 |
| 0.00 | 5.00 | 1.739 | 9.304 | 8.302 | 4.529 | 2.283 | 2.685 | 7.205 | 8.493 | *1.903* | 4.665 |
| 0.00 | 10.00 | 1.723 | 9.297 | 15.404 | 11.509 | 7.219 | *6.267* | 15.051 | 15.883 | 10.687 | 10.794 |
| 1.00 | 0.00 | 1.742 | 9.316 | 2.122 | 2.102 | 2.113 | 2.712 | 2.206 | 2.147 | *1.897* | 2.093 |
| 5.00 | 0.00 | 1.749 | 9.322 | 4.395 | 4.335 | 3.949 | 4.515 | 4.616 | 4.652 | *2.044* | 2.143 |
| 10.00 | 0.00 | 1.729 | 9.321 | 6.413 | 6.530 | 6.685 | 6.301 | 6.822 | 6.688 | 6.199 | *3.921* |
| 0.25 | 0.75 | 1.733 | 9.332 | 2.119 | 1.865 | 1.876 | 2.148 | 1.857 | 1.943 | 1.860 | *1.855* |
| 1.25 | 3.75 | 1.730 | 9.290 | 4.733 | 2.058 | 1.958 | 1.881 | 3.979 | 4.969 | *1.869* | 2.576 |
| 2.50 | 7.50 | 1.715 | 9.295 | 7.880 | 4.420 | 2.034 | 2.661 | 7.748 | 8.119 | *1.895* | 4.785 |
| 0.50 | 0.50 | 1.722 | 9.329 | 1.862 | 1.892 | 1.896 | 2.292 | 1.851 | *1.830* | 1.847 | 1.834 |
| 2.50 | 2.50 | 1.731 | 9.286 | 2.239 | 2.406 | 2.543 | 2.329 | 1.873 | 2.158 | *1.853* | 2.050 |
| 5.00 | 5.00 | 1.745 | 9.310 | 2.787 | 2.348 | 3.596 | 2.328 | 2.245 | 2.728 | *1.934* | 2.405 |
| 0.75 | 0.25 | 1.742 | 9.319 | 1.885 | 1.986 | 1.983 | 2.497 | 1.990 | 1.923 | *1.873* | *1.873* |
| 3.75 | 1.25 | 1.745 | 9.284 | 2.519 | 3.537 | 3.199 | 3.402 | 3.022 | 2.909 | *1.895* | 2.873 |
| 7.50 | 2.50 | 1.727 | 9.306 | 3.259 | 5.004 | 5.179 | 4.527 | 4.094 | 3.487 | *2.085* | 5.080 |

*%X* indicates the percentage of outliers in *x*. *%Y* indicates the percentage of outliers in *y*

TC-ratio estimator is smaller than half of one standard error in 10 out of these 13 cases. In the case of ($\%X = 7.50$, $\%Y = 2.50$), the bias of the TC-ratio estimator is larger than half of one standard error, but it is still comparatively smaller than the biases of the other competing methods. The two scenarios ($\%X = 0.00$, $\%Y = 10.00$; $\%X = 10.00$, $\%Y = 0.00$) are found too hard to deal with, because no methods can adequately handle these two scenarios.

Taking both bias and efficiency into account, RMSE shows that the TC-ratio estimator is almost always best among the competing robust ratio imputation methods. In fact, the TC-ratio estimator is judged best in 10 out of 16 patterns. For the remaining six patterns, the differences in RMSE are quite small. The remarkable characteristic of the TC-ratio estimator is that RMSE is quite stable under most situations, ranging from 1.841 to 2.085 in 14 patterns.

Furthermore, when outliers are present, the TC-ratio estimator outperforms the regular ratio imputation model; and when there are no outliers, the TC-ratio estimator (bias = -0.326, RMSE = 1.878) works approximately equally well compared to the regular ratio imputation model (bias = -0.012, RMSE = 1.852). Also, the TC-ratio estimator (C-1) outperforms C-2 in 12 out of 16 patterns with 1 tie in terms of both bias and RMSE. When the proportion of outliers is 1%, the performance of the TC-ratio estimator (C-1) and the usual criterion of $4/(n-p)$ (C-2) is similar; therefore, if we are certain that the proportion of outliers is low, the usual criterion of $4/(n-p)$ might be enough. However, when we want to automate the process of imputation, there is uncertainty as to the proportion of outliers. Therefore, in case that the proportion of outliers is high, the TC-ratio estimator (C-1) is more preferable than the usual criterion of $4/(n-p)$ (C-2).

### 8.2 MAR2 in population 1

Table 5 presents the results of the simulations for population 1 (Gamma distribution, $d = 1.84$, $g = 0.75$) under MAR2.

In Table 5, if the absolute value of bias is smaller than 0.850 (half of one standard error), then it is shown in italics. Also, the smallest value of RMSE is shown in italics. The overall conclusions are similar to the ones in Sect. 8.1.

Remember that, under MAR2, the missing rates of $y_i$ are higher when $x_i > \text{med}(x_i)$. This means that larger values of $y_i$ tend to be missing. Also, $x_i$ and $y_i$ both follow gamma distributions, which are skewed to the right with right long tails. This further means that many missing values are scattered among very large values of $y_i$. Therefore, the situation is more difficult to handle than in Sect. 8.1. In fact, the absolute size of biases of the regular ratio imputation model (ratio) tend to be larger than in MAR1.

When there are no outliers, the biases of M-1, Trim, Wins, and C-1 are smaller than half of the standard error; thus, we consider them unbiased.

When the bias of the regular ratio imputation model is large in 14 cases, the TC-ratio estimator (C-1) corrects the bias within half of one standard error in 8 cases. In the four cases ($\%X = 0.00$, $\%Y = 5.00$; $\%X = 5.00$, $\%Y = 0.00$; $\%X = 2.50$, $\%Y = 7.50$; $\%X = 7.50$, $\%Y = 2.50$), the biases of the TC-ratio estimator are comparatively

**Table 5** Population 1 (gamma distribution: d = 1.84, g = 0.75), MAR2

Bias: |value| < 0.850 in shaded cells

| %X | %Y | Comp | LD | Ratio | M-1 | M-2 | Med | Trim | Wins | C-1 | C-2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.00 | 0.00 | *− 0.012* | − 9.058 | *− 0.006* | *− 0.610* | − 1.449 | − 2.123 | *− 0.838* | *− 0.420* | *− 0.579* | − 1.703 |
| 0.00 | 1.00 | *− 0.030* | − 9.085 | 3.860 | *− 0.006* | *− 0.222* | − 1.070 | 1.244 | 2.317 | *0.247* | *0.332* |
| 0.00 | 5.00 | *0.008* | − 9.030 | 18.336 | 9.595 | 3.182 | 3.578 | 15.756 | 18.919 | 1.189 | 10.548 |
| 0.00 | 10.00 | *− 0.002* | − 9.046 | 34.363 | 25.919 | 16.564 | 10.110 | 33.856 | 35.677 | 25.805 | 25.104 |
| 1.00 | 0.00 | *− 0.015* | − 9.070 | − 2.457 | − 2.233 | − 2.201 | − 3.116 | − 2.476 | − 2.420 | *− 0.847* | − 2.035 |
| 5.00 | 0.00 | *− 0.012* | − 9.037 | − 8.675 | − 8.601 | − 7.814 | − 6.202 | − 8.985 | − 9.295 | − 1.733 | − 2.244 |
| 10.00 | 0.00 | *0.002* | − 9.039 | − 12.828 | − 13.144 | − 13.648 | − 9.190 | − 13.564 | − 13.420 | − 12.551 | − 8.053 |
| 0.25 | 0.75 | *0.007* | − 9.041 | 2.201 | *− 0.583* | *− 0.581* | − 1.575 | *0.293* | 1.053 | *− 0.047* | *0.063* |
| 1.25 | 3.75 | *− 0.023* | − 9.072 | 9.541 | 1.891 | − 1.382 | *0.561* | 7.969 | 10.514 | *0.658* | 5.350 |
| 2.50 | 7.50 | *0.015* | − 9.009 | 16.430 | 8.695 | 1.756 | 3.160 | 16.884 | 17.104 | 1.052 | 11.559 |
| 0.50 | 0.50 | *0.000* | − 9.061 | *0.577* | − 1.157 | − 1.035 | − 2.101 | *− 0.669* | *− 0.165* | *− 0.405* | *− 0.216* |
| 2.50 | 2.50 | *− 0.014* | − 9.054 | 2.313 | − 3.497 | − 3.864 | − 2.002 | *0.529* | 2.039 | *− 0.229* | *− 0.627* |
| 5.00 | 5.00 | *0.010* | − 9.050 | 3.799 | − 2.932 | − 6.656 | − 1.883 | 2.356 | 3.655 | *− 0.635* | − 1.675 |
| 0.75 | 0.25 | *− 0.020* | − 9.048 | − 0.980 | − 1.706 | − 1.566 | − 2.618 | − 1.597 | − 1.324 | *− 0.676* | *− 0.624* |
| 3.75 | 1.25 | *0.021* | − 9.011 | − 3.611 | − 6.615 | − 5.664 | − 4.187 | − 5.207 | − 5.163 | *− 0.661* | − 4.087 |
| 7.50 | 2.50 | *0.013* | − 9.016 | − 5.551 | − 9.917 | − 10.307 | − 5.810 | − 7.875 | − 6.189 | − 1.593 | − 9.642 |

**Table 5** (continued)

RMSE: Best outcomes in italics for each outlier setting excluding Comp

| %X | %Y | Comp | LD | Ratio | M-1 | M-2 | Med | Trim | Wins | C-1 | C-2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.00 | 0.00 | 1.728 | 9.257 | *1.913* | 1.997 | 2.382 | 2.890 | 2.061 | 1.946 | 2.014 | 2.534 |
| 0.00 | 1.00 | 1.720 | 9.283 | 4.414 | *1.923* | 1.927 | 2.246 | 2.311 | 3.061 | 1.995 | 2.074 |
| 0.00 | 5.00 | 1.729 | 9.231 | 18.813 | 10.014 | 3.871 | 4.177 | 16.113 | 19.390 | *2.283* | 11.175 |
| 0.00 | 10.00 | 1.723 | *9.247* | 35.132 | 26.571 | 17.091 | 10.400 | 34.517 | 36.452 | 26.650 | 25.869 |
| 1.00 | 0.00 | 1.725 | 9.268 | 3.092 | 2.911 | 2.884 | 3.660 | 3.083 | 3.046 | *2.116* | 2.757 |
| 5.00 | 0.00 | 1.714 | 9.234 | 8.892 | 8.817 | 8.047 | 6.464 | 9.174 | 9.495 | *2.579* | 2.921 |
| 10.00 | 0.00 | 1.719 | 9.236 | 12.983 | 13.296 | 13.798 | 9.354 | 13.700 | 13.565 | 12.731 | *8.302* |
| 0.25 | 0.75 | 1.735 | 9.239 | 3.004 | 2.002 | 2.005 | 2.524 | *1.952* | 2.228 | 1.977 | 1.956 |
| 1.25 | 3.75 | 1.718 | 9.271 | 10.027 | 2.851 | 2.373 | 2.105 | 8.353 | 10.983 | *2.066* | 5.946 |
| 2.50 | 7.50 | 1.739 | 9.215 | 17.026 | 9.277 | 2.925 | 3.814 | 17.364 | 17.687 | *2.248* | 12.113 |
| 0.50 | 0.50 | 1.732 | 9.259 | 2.038 | 2.215 | 2.159 | 2.867 | 2.004 | *1.922* | 1.984 | 1.924 |
| 2.50 | 2.50 | 1.731 | 9.250 | 3.315 | 3.998 | 4.297 | 2.774 | 2.004 | 3.062 | *1.937* | 2.239 |
| 5.00 | 5.00 | 1.719 | 9.246 | 4.819 | 3.654 | 6.949 | 2.690 | 3.346 | 4.675 | *2.092* | 3.215 |
| 0.75 | 0.25 | 1.744 | 9.251 | 2.158 | 2.551 | 2.462 | 3.263 | 2.468 | 2.313 | 2.071 | *2.026* |
| 3.75 | 1.25 | 1.725 | 9.213 | 4.153 | 6.894 | 5.979 | 4.587 | 5.529 | 5.532 | *2.025* | 4.859 |
| 7.50 | 2.50 | 1.703 | 9.214 | 5.970 | 10.118 | 10.494 | 6.085 | 8.108 | 6.545 | *2.513* | 9.911 |

*%X* indicates the percentage of outliers in *x*. *%Y* indicates the percentage of outliers in *y*

smaller than those of the competing methods. The two scenarios ($\%X = 0.00$, $\%Y = 10.00$; $\%X = 10.00$, $\%Y = 0.00$) are, again, found too hard to deal with, because no methods can adequately handle these two scenarios.

In terms of RMSE, the TC-ratio estimator is judged best in 9 out of 16 patterns. For the remaining seven patterns, the differences in RMSE are quite small. Again, the remarkable characteristic of the TC-ratio estimator is that RMSE is quite stable under most situations ranging from 1.937 to 2.579 in 14 patterns. Also, the TC-ratio estimator (C-1) outperforms C-2 in 12 out of 16 patterns in terms of bias, and in 11 out of 16 patterns in terms of RMSE.

## 9 Summary of the overall results

Table 6 summarizes the results of all the 160 different data patterns. The row, "Unbiased," shows the number of times the bias of each method was less than half of one standard error. The row, "RMSE," shows the number of times the RMSE of each method was smallest among the competing methods.

The TC-ratio estimator is deemed unbiased in 114 out of 160 patterns, and the relative performance of the TC-ratio estimator is best in 106 out of 160 patterns in terms of RMSE. Therefore, the TC-ratio estimator is remarkably robust under a variety of outlier settings, missing data types, and distributional assumptions. Against C-2, in terms of RMSE, C-1 wins 121 times and loses 31 times, with 8 ties, in 160 patterns. For the results of specific data types, see the Appendix in Sect. 11.

## 10 Conclusion

This article proposed a new robust ratio imputation model based on the TC-ratio estimator, which extended Cook's distance to the ratio estimator. Simulation studies showed that the new robust ratio imputation model is robust against many types of outliers under a variety of settings. This method works better than the traditional robust methods (the ratio of medians, trimmed means, Winsorized means, and means by $M$-estimators) when outliers are on the vertical axis. This method works far better than the traditional robust methods when outliers are on the horizontal axis (high-leverage points). Also, this method works approximately equally well compared to the non-robust method when there are no outliers. This is true regardless of

**Table 6** Summary of the overall results in 160 data patterns

| | Comp | LD | Ratio | M-1 | M-2 | Med | Trim | Wins | C-1 | C-2 |
|---|---|---|---|---|---|---|---|---|---|---|
| Unbiased | 160 | 0 | 39 | 55 | 57 | 39 | 55 | 41 | *114* | 60 |
| RMSE | NA | 3 | 11 | 22 | 18 | 4 | 8 | 9 | *106* | 26 |

For RMSE comparisons, Comp is excluded; thus, displayed as NA (not applicable)

the distributional assumptions (gamma, uniform, and normal distributions: also see the Appendix in Sect. 11). Therefore, the TC-ratio estimator is comparatively a more robust ratio estimator than the traditional robust methods.

Furthermore, since $M$-estimators are iterative methods, whether the algorithm converges depends on the choice of parameter settings in $M$-estimators (Mulry et al., 2014, p.733). In case the algorithm does not converge, the literature suggests a need to have a backup strategy (Mulry et al., 2014, pp.744–745). The TC-ratio estimator in the current study is not an iterative method. Therefore, even if $M$-estimators are chosen for a particular survey as a method of imputation, the TC-ratio estimator can be a reliable back-up method of imputation for $M$-estimators. Also, it is reported that developing an automatic data-driven method for $M$-estimators is challenging due to the difficulties in setting the initial value of the tuning constant $\psi$ (Mulry et al., 2018, p.483). The TC-ratio estimator in the current study is a fully automatic data-driven method. In this sense, too, the proposed method is highly useful.

The following is out of scope for this article. The current study proposed an outlier resistant single imputation method, because the goal was to compute the means (or the totals). If the goal is to make an inference about the population parameters based on sample statistics, then we may need to consider either of the following two methods. One is multiple imputation (Carpenter & Kenward, 2013, p.35; van Buuren, 2018, p.25). For this, Takahashi (2017a) and Takahashi (2017b) proposed multiple ratio imputation based on the expectation–maximization with bootstrapping, which is known to be a fast and reliable multiple imputation algorithm (Takahashi, 2017c). How multiple ratio imputation can be robustified by the TC-ratio estimator will be an important future research topic. Another strategy is to use variance estimation procedures for singly imputed data (Deville & Särndal, 1994, p.389, p.392; Haziza & Vallée, 2020). How these variance estimation procedures for singly imputed data can be applied to the TC-ratio estimator will be also another important future research topic.

Let us end this article with a final remark on the potential limitation of the proposed method. Just as Young and Mathew (2015, p.93) note, this article does not necessarily suggest a panacea for outlier treatments in all survey settings. While Cook's distance is known as one of the most well-established methods to detect individually influential observations, Cook's distance may overlook the mutually influential observations or a group of influential observations (Lawrance, 1995, p.181). This problem is known as masking. In cases where observations are jointly influential, Cook's distance can be sequentially applied, but even the sequential approach may not be always successful (Fox, 2020, p.51). There are two ways to deal with this problem. First, Lawrance (1995, p.184) proposed a conditional approach as a measure of the masking in Cook's distance. How the TC-ratio estimator can incorporate the conditional approach by Lawrance (1995) is left for future research. Second, due to the possibility of masking, the literature suggests to complement outlier detection techniques with graphical methods (Fox, 2020, pp.52–53; Filliben & Heckert, 2013). In fact, by examining outliers in detail, we may find "omitted variables, incorrect functional forms, …, or other neglected aspects of a study" (Bollen, 1989, p.31). Whenever possible, subject matter knowledge should be incorporated into statistical analysis in dealing with outliers (de Waal et al., 2011, p.230; Young

& Mathew, 2015, p.77). The current study focused on the side of statistical analysis only. If we incorporate subject matter knowledge in the process of outlier treatments for imputation, the findings of this study will be further strengthened.

## 11 Appendix

This appendix displays the results for the other four populations (gamma with $d = 5.13$, $g = 0.50$; gamma with $d = 13.78$, $g = 0.25$; uniform $[0.1, 2.1]$; and normal with mean = 20, variance = 16). In the following tables, unbiased results (smaller than half of one standard error) are shown in italics. The information on the standard error in each table can be found in the column of Comp under RMSE. Again, see Schafer and Graham (2002, p.157) for this rule of thumb to judge the size of bias. Also, the smallest RMSE value is shown in italics.

### 11.1 Results of the simulations for populations 2–5 in MAR1

See Tables 7, 8, 9, 10.

### 11.2 Results of the simulations for populations 2–5 in MAR2

See Tables 11, 12, 13, 14.

**Table 7** Population 2 (gamma distribution: $d = 5.13$, $g = 0.50$), MAR1

Bias: |value| < 0.850 in italics

| %X | %Y | Comp | LD | Ratio | M-1 | M-2 | Med | Trim | Wins | C-1 | C-2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.00 | 0.00 | *− 0.002* | 9.036 | *− 0.002* | *− 0.297* | *− 0.747* | − 1.111 | *− 0.333* | *− 0.132* | *− 0.339* | − 0.986 |
| 0.00 | 1.00 | *− 0.004* | 9.019 | 1.475 | *− 0.009* | *− 0.218* | *− 0.493* | *0.682* | 1.125 | *− 0.096* | *0.065* |
| 0.00 | 5.00 | *− 0.029* | 8.994 | 7.076 | 3.658 | 1.171 | 2.254 | 6.329 | 7.285 | *0.327* | 3.585 |
| 0.00 | 10.00 | *− 0.016* | 9.023 | 13.471 | 9.989 | 6.146 | 6.432 | 13.415 | 13.939 | 8.081 | 9.217 |
| 1.00 | 0.00 | *− 0.059* | 8.987 | − 1.107 | − 1.090 | − 1.120 | − 1.778 | − 1.175 | − 1.115 | *− 0.441* | − 1.119 |
| 5.00 | 0.00 | *− 0.020* | 9.005 | − 4.020 | − 3.967 | − 3.597 | − 3.994 | − 4.195 | − 4.263 | − 0.909 | − 1.134 |
| 10.00 | 0.00 | *− 0.009* | 9.030 | − 6.175 | − 6.294 | − 6.471 | − 5.950 | − 6.546 | − 6.427 | − 6.010 | − 3.642 |
| 0.25 | 0.75 | *− 0.026* | 9.020 | *0.789* | *− 0.316* | *− 0.425* | *− 0.826* | *0.187* | *0.516* | *− 0.228* | *− 0.065* |
| 1.25 | 3.75 | *0.010* | 9.070 | 3.674 | *0.721* | *− 0.615* | *0.531* | 3.218 | 3.948 | *0.162* | 1.495 |
| 2.50 | 7.50 | *0.003* | 9.051 | 6.502 | 3.322 | *0.489* | 2.279 | 6.647 | 6.749 | *0.381* | 3.691 |
| 0.50 | 0.50 | *0.017* | 9.055 | *0.181* | *− 0.541* | *− 0.588* | − 1.101 | *− 0.236* | *− 0.005* | *− 0.274* | *− 0.155* |
| 2.50 | 2.50 | *− 0.009* | 9.047 | *0.711* | − 1.515 | − 1.773 | − 1.123 | *0.192* | *0.618* | *− 0.172* | *− 0.691* |
| 5.00 | 5.00 | *0.033* | 9.084 | 1.244 | − 1.443 | − 2.983 | − 1.080 | *0.787* | 1.185 | *− 0.364* | − 1.341 |
| 0.75 | 0.25 | *0.004* | 9.047 | *− 0.447* | *− 0.792* | *− 0.810* | − 1.412 | *− 0.684* | *− 0.541* | *− 0.341* | *− 0.380* |
| 3.75 | 1.25 | *− 0.008* | 9.039 | − 1.816 | − 2.989 | − 2.645 | − 2.604 | − 2.320 | − 2.242 | *− 0.358* | − 1.826 |
| 7.50 | 2.50 | *− 0.009* | 9.062 | − 2.885 | − 4.614 | − 4.852 | − 3.961 | − 3.665 | − 3.138 | − 1.107 | − 4.708 |

**Table 7** (continued)

RMSE: Best outcomes in italics for each outlier setting excluding Comp

| %X | %Y | Comp | LD | Ratio | M-1 | M-2 | Med | Trim | Wins | C-1 | C-2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.00 | 0.00 | 1.730 | 9.306 | *1.833* | 1.855 | 1.973 | 2.159 | 1.861 | 1.839 | 1.869 | 2.075 |
| 0.00 | 1.00 | 1.736 | 9.290 | 2.390 | *1.839* | 1.849 | 1.943 | 1.978 | 2.179 | 1.851 | 1.860 |
| 0.00 | 5.00 | 1.737 | 9.270 | 7.453 | 4.185 | 2.237 | 2.996 | 6.695 | 7.655 | *1.887* | 4.229 |
| 0.00 | 10.00 | 1.723 | 9.289 | 13.820 | 10.335 | *6.541* | 6.765 | 13.732 | 14.286 | 8.789 | 9.645 |
| 1.00 | 0.00 | 1.731 | 9.253 | 2.113 | 2.110 | 2.131 | 2.554 | 2.152 | 2.118 | *1.884* | 2.132 |
| 5.00 | 0.00 | 1.745 | 9.275 | 4.405 | 4.357 | 4.025 | 4.375 | 4.556 | 4.626 | *2.047* | 2.147 |
| 10.00 | 0.00 | 1.740 | 9.302 | 6.434 | 6.548 | 6.722 | 6.204 | 6.785 | 6.675 | 6.289 | *4.079* |
| 0.25 | 0.75 | 1.743 | 9.289 | 2.019 | 1.859 | 1.882 | 2.031 | 1.855 | 1.921 | 1.859 | *1.842* |
| 1.25 | 3.75 | 1.731 | 9.336 | 4.200 | 1.991 | 1.920 | 1.960 | 3.762 | 4.451 | *1.843* | 2.461 |
| 2.50 | 7.50 | 1.738 | 9.319 | 6.916 | 3.906 | 1.950 | 2.998 | 7.021 | 7.151 | *1.880* | 4.287 |
| 0.50 | 0.50 | 1.752 | 9.330 | 1.857 | 1.915 | 1.935 | 2.171 | 1.860 | *1.847* | 1.874 | 1.857 |
| 2.50 | 2.50 | 1.711 | 9.313 | 2.015 | 2.358 | 2.522 | 2.152 | 1.831 | 1.968 | *1.824* | 1.984 |
| 5.00 | 5.00 | 1.725 | 9.347 | 2.360 | 2.344 | 3.485 | 2.132 | 2.048 | 2.320 | *1.865* | 2.401 |
| 0.75 | 0.25 | 1.725 | 9.314 | 1.863 | 1.972 | 1.985 | 2.319 | 1.931 | 1.885 | *1.860* | 1.864 |
| 3.75 | 1.25 | 1.706 | 9.302 | 2.557 | 3.477 | 3.186 | 3.166 | 2.916 | 2.869 | *1.840* | 2.676 |
| 7.50 | 2.50 | 1.744 | 9.331 | 3.425 | 4.956 | 5.178 | 4.353 | 4.084 | 3.634 | *2.211* | 5.059 |

*%X* indicates the percentage of outliers in *x*. *%Y* indicates the percentage of outliers in *y*

**Table 8** Population 3 (Gamma distribution, d = 13.78, g = 0.25), MAR1

Bias: |value| < 0.850 in italics

| %X | %Y | Comp | LD | Ratio | M-1 | M-2 | Med | Trim | Wins | C-1 | C-2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.00 | 0.00 | *0.011* | 9.058 | *0.011* | *− 0.317* | *− 0.752* | *− 0.719* | *− 0.200* | *− 0.074* | *− 0.384* | − 0.991 |
| 0.00 | 1.00 | *− 0.018* | 9.043 | 1.338 | *− 0.028* | *− 0.303* | *− 0.082* | *0.778* | 1.101 | *− 0.221* | *0.011* |
| 0.00 | 5.00 | *0.014* | 9.074 | 6.455 | 3.424 | 1.167 | 2.797 | 5.978 | 6.665 | *0.264* | 3.295 |
| 0.00 | 10.00 | *− 0.008* | 9.046 | 12.221 | 9.092 | 5.660 | 7.016 | 12.374 | 12.668 | 6.674 | 8.389 |
| 1.00 | 0.00 | *0.013* | 9.066 | − 1.035 | − 1.055 | − 1.054 | − 1.338 | − 0.986 | − 0.997 | *− 0.411* | − 1.010 |
| 5.00 | 0.00 | *− 0.020* | 9.035 | − 3.989 | − 3.956 | − 3.590 | − 3.704 | − 4.080 | − 4.196 | − 0.924 | − 1.078 |
| 10.00 | 0.00 | *− 0.017* | 9.032 | − 6.132 | − 6.254 | − 6.412 | − 5.734 | − 6.438 | − 6.352 | − 5.970 | − 3.722 |
| 0.25 | 0.75 | *0.001* | 9.037 | *0.721* | *− 0.309* | *− 0.465* | *− 0.386* | *0.322* | *0.551* | *− 0.286* | *− 0.090* |
| 1.25 | 3.75 | *− 0.020* | 9.002 | 3.169 | *0.582* | *− 0.657* | 0.918 | 2.970 | 3.451 | *− 0.017* | 1.348 |
| 2.50 | 7.50 | *− 0.007* | 9.028 | 5.641 | 2.832 | *0.340* | 2.757 | 5.994 | 5.883 | *0.265* | 3.258 |
| 0.50 | 0.50 | *0.018* | 9.059 | *0.115* | *− 0.577* | *− 0.650* | *− 0.713* | *− 0.128* | *0.017* | *− 0.353* | *− 0.219* |
| 2.50 | 2.50 | *0.007* | 9.061 | *0.454* | − 1.485 | − 1.770 | *− 0.711* | *0.184* | *0.389* | *− 0.219* | *− 0.576* |
| 5.00 | 5.00 | *− 0.004* | 9.033 | *0.756* | − 1.612 | − 3.001 | *− 0.732* | *0.503* | *0.707* | *− 0.359* | − 1.562 |
| 0.75 | 0.25 | *0.005* | 9.053 | *− 0.486* | *− 0.838* | − 0.858 | − 1.041 | *− 0.578* | *− 0.514* | *− 0.405* | *− 0.448* |
| 3.75 | 1.25 | *− 0.011* | 9.042 | − 1.928 | − 2.965 | − 2.667 | − 2.269 | − 2.238 | − 2.252 | *− 0.421* | − 1.678 |
| 7.50 | 2.50 | *− 0.022* | 9.011 | − 3.110 | − 4.621 | − 4.888 | − 3.722 | − 3.690 | − 3.334 | − 1.391 | − 4.747 |

**Table 8** (continued)

RMSE: Best outcomes in italics for each outlier setting excluding Comp

| %X | %Y | Comp | LD | Ratio | M-1 | M-2 | Med | Trim | Wins | C-1 | C-2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.00 | 0.00 | 1.747 | 9.321 | *1.817* | 1.841 | 1.961 | 2.000 | 1.837 | 1.823 | 1.861 | 2.071 |
| 0.00 | 1.00 | 1.732 | 9.305 | 2.275 | *1.806* | 1.827 | 1.873 | 1.992 | 2.138 | 1.821 | 1.819 |
| 0.00 | 5.00 | 1.734 | 9.337 | 6.800 | 3.936 | 2.195 | 3.416 | 6.320 | 7.004 | *1.835* | 3.918 |
| 0.00 | 10.00 | 1.733 | 9.306 | 12.521 | 9.404 | *6.041* | 7.330 | 12.652 | 12.967 | 7.328 | 8.780 |
| 1.00 | 0.00 | 1.729 | 9.325 | 2.057 | 2.070 | 2.075 | 2.273 | 2.040 | 2.040 | *1.843* | 2.059 |
| 5.00 | 0.00 | 1.733 | 9.296 | 4.367 | 4.337 | 4.010 | 4.120 | 4.445 | 4.556 | *2.039* | 2.106 |
| 10.00 | 0.00 | 1.753 | 9.297 | 6.390 | 6.509 | 6.664 | *6.003* | 6.681 | 6.602 | 6.252 | 4.146 |
| 0.25 | 0.75 | 1.714 | 9.295 | 1.938 | 1.807 | 1.842 | 1.889 | 1.831 | 1.882 | 1.811 | *1.796* |
| 1.25 | 3.75 | 1.713 | 9.261 | 3.702 | 1.896 | 1.900 | 2.101 | 3.516 | 3.959 | *1.795* | 2.311 |
| 2.50 | 7.50 | 1.712 | 9.288 | 6.035 | 3.433 | 1.864 | 3.371 | 6.356 | 6.266 | *1.823* | 3.855 |
| 0.50 | 0.50 | 1.714 | 9.322 | 1.795 | 1.872 | 1.901 | 1.976 | 1.803 | *1.794* | 1.827 | 1.810 |
| 2.50 | 2.50 | 1.697 | 9.317 | 1.867 | 2.310 | 2.500 | 1.968 | *1.795* | 1.844 | 1.797 | 1.923 |
| 5.00 | 5.00 | 1.742 | 9.298 | 2.046 | 2.426 | 3.497 | 2.008 | 1.910 | 2.023 | *1.861* | 2.457 |
| 0.75 | 0.25 | 1.742 | 9.313 | 1.868 | 1.987 | 2.002 | 2.133 | 1.904 | 1.879 | *1.862* | 1.879 |
| 3.75 | 1.25 | 1.719 | 9.299 | 2.613 | 3.449 | 3.197 | 2.903 | 2.846 | 2.865 | *1.836* | 2.585 |
| 7.50 | 2.50 | 1.730 | 9.269 | 3.586 | 4.950 | 5.200 | 4.128 | 4.090 | 3.779 | *2.356* | 5.079 |

*%X* indicates the percentage of outliers in *x*. *%Y* indicates the percentage of outliers in *y*

**Table 9** Population 4 (uniform distribution: [0.1, 2.1]), MAR1

Bias: |value| < 0.040 in italics

| %X | %Y | Comp | LD | Ratio | M-1 | M-2 | Med | Trim | Wins | C-1 | C-2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.00 | 0.00 | *− 0.002* | 0.556 | *− 0.002* | *− 0.002* | *− 0.002* | *− 0.033* | *− 0.014* | *− 0.006* | *− 0.001* | *− 0.002* |
| 0.00 | 1.00 | *0.000* | 0.558 | 0.040 | *0.006* | *0.007* | *0.001* | *0.025* | *0.034* | *0.005* | *0.005* |
| 0.00 | 5.00 | *0.000* | 0.557 | 0.189 | 0.097 | 0.043 | 0.140 | 0.177 | 0.186 | *0.018* | 0.094 |
| 0.00 | 10.00 | *0.000* | 0.557 | 0.357 | 0.256 | 0.164 | 0.344 | 0.358 | 0.357 | 0.092 | 0.233 |
| 1.00 | 0.00 | *0.000* | 0.556 | *− 0.025* | *− 0.016* | *− 0.010* | − 0.063 | *− 0.039* | *− 0.030* | *− 0.006* | *− 0.009* |
| 5.00 | 0.00 | *0.001* | 0.558 | − 0.106 | − 0.089 | − 0.055 | − 0.178 | − 0.125 | − 0.111 | *− 0.038* | *− 0.036* |
| 10.00 | 0.00 | *0.001* | 0.559 | − 0.182 | − 0.171 | − 0.141 | − 0.287 | − 0.205 | − 0.188 | − 0.119 | − 0.110 |
| 0.25 | 0.75 | *0.000* | 0.557 | *0.023* | *− 0.001* | *0.003* | *− 0.015* | *0.008* | *0.017* | *0.002* | *0.001* |
| 1.25 | 3.75 | *− 0.001* | 0.557 | 0.107 | *0.029* | *0.000* | 0.050 | 0.091 | 0.103 | *0.004* | *0.027* |
| 2.50 | 7.50 | *0.000* | 0.558 | 0.198 | 0.105 | *0.037* | 0.140 | 0.186 | 0.196 | *0.016* | 0.088 |
| 0.50 | 0.50 | *0.001* | 0.558 | *0.008* | *− 0.005* | *0.001* | *− 0.030* | *− 0.006* | *0.003* | *0.001* | *− 0.002* |
| 2.50 | 2.50 | *− 0.002* | 0.555 | *0.030* | *− 0.026* | *− 0.027* | *− 0.033* | *0.011* | *0.024* | *− 0.007* | *− 0.030* |
| 5.00 | 5.00 | *0.000* | 0.558 | 0.058 | *− 0.020* | − 0.053 | *− 0.031* | *0.037* | 0.053 | *− 0.021* | *− 0.032* |
| 0.75 | 0.25 | *− 0.001* | 0.556 | *− 0.010* | *− 0.012* | *− 0.005* | − 0.048 | *− 0.025* | *− 0.016* | *− 0.004* | *− 0.006* |
| 3.75 | 1.25 | *0.000* | 0.558 | *− 0.039* | − 0.064 | − 0.042 | − 0.107 | − 0.060 | − 0.046 | *− 0.018* | − 0.065 |
| 7.50 | 2.50 | *0.001* | 0.557 | − 0.069 | − 0.113 | − 0.107 | − 0.178 | − 0.094 | − 0.075 | − 0.054 | − 0.125 |

**Table 9** (continued)

RMSE: Best outcomes in italics for each outlier setting excluding Comp

| %X | %Y | Comp | LD | Ratio | M-1 | M-2 | Med | Trim | Wins | C-1 | C-2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.00 | 0.00 | 0.078 | 0.564 | *0.080* | *0.080* | *0.080* | 0.086 | 0.081 | *0.080* | *0.080* | 0.080 |
| 0.00 | 1.00 | 0.079 | 0.566 | 0.091 | *0.081* | *0.081* | *0.081* | 0.085 | 0.088 | *0.081* | 0.082 |
| 0.00 | 5.00 | 0.078 | 0.565 | 0.206 | 0.127 | 0.092 | 0.161 | 0.196 | 0.204 | *0.083* | 0.126 |
| 0.00 | 10.00 | 0.078 | 0.565 | 0.368 | 0.270 | 0.184 | 0.354 | 0.369 | 0.368 | *0.123* | 0.250 |
| 1.00 | 0.00 | 0.079 | 0.565 | 0.085 | 0.082 | 0.082 | 0.102 | 0.090 | 0.086 | *0.081* | 0.081 |
| 5.00 | 0.00 | 0.079 | 0.566 | 0.132 | 0.120 | 0.097 | 0.195 | 0.148 | 0.137 | *0.088* | 0.087 |
| 10.00 | 0.00 | 0.077 | 0.567 | 0.198 | 0.189 | 0.162 | 0.299 | 0.220 | 0.204 | 0.144 | *0.136* |
| 0.25 | 0.75 | 0.078 | 0.565 | 0.084 | *0.080* | *0.080* | 0.082 | 0.081 | 0.082 | *0.080* | *0.080* |
| 1.25 | 3.75 | 0.079 | 0.565 | 0.135 | 0.087 | 0.082 | 0.095 | 0.122 | 0.132 | *0.081* | 0.087 |
| 2.50 | 7.50 | 0.079 | 0.566 | 0.216 | 0.133 | 0.090 | 0.162 | 0.204 | 0.213 | *0.083* | 0.123 |
| 0.50 | 0.50 | 0.078 | 0.566 | 0.081 | 0.081 | *0.080* | 0.086 | *0.080* | *0.080* | *0.080* | *0.080* |
| 2.50 | 2.50 | 0.079 | 0.563 | 0.087 | 0.085 | 0.086 | 0.087 | 0.082 | 0.085 | *0.081* | 0.087 |
| 5.00 | 5.00 | 0.078 | 0.566 | 0.100 | *0.083* | 0.097 | 0.086 | 0.089 | 0.098 | *0.083* | 0.089 |
| 0.75 | 0.25 | 0.078 | 0.564 | 0.081 | 0.081 | *0.080* | 0.093 | 0.084 | 0.082 | *0.080* | *0.080* |
| 3.75 | 1.25 | 0.078 | 0.566 | 0.089 | 0.103 | 0.090 | 0.134 | 0.100 | 0.092 | *0.081* | 0.104 |
| 7.50 | 2.50 | 0.078 | 0.566 | 0.105 | 0.138 | 0.134 | 0.196 | 0.123 | 0.110 | *0.096* | 0.149 |

*%X* indicates the percentage of outliers in *x*. *%Y* indicates the percentage of outliers in *y*

**Table 10** Population 5 (normal distribution: mean = 20, variance = 16), MAR1

Bias: |value| < 0.250 in italics

| %X | %Y | Comp | LD | Ratio | M-1 | M-2 | Med | Trim | Wins | C-1 | C-2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.00 | 0.00 | *− 0.003* | 3.554 | *− 0.003* | *− 0.003* | *− 0.003* | *− 0.033* | *− 0.016* | *− 0.007* | *− 0.002* | *− 0.003* |
| 0.00 | 1.00 | *− 0.003* | 3.557 | 0.528 | *0.033* | *0.039* | *0.214* | 0.364 | 0.418 | *0.047* | *0.058* |
| 0.00 | 5.00 | *− 0.003* | 3.552 | 2.496 | 1.109 | 0.500 | 1.207 | 2.031 | 2.376 | *0.106* | 0.930 |
| 0.00 | 10.00 | *− 0.002* | 3.558 | 4.705 | 3.129 | 1.965 | 2.550 | 4.223 | 4.710 | 0.685 | 2.501 |
| 1.00 | 0.00 | *− 0.008* | 3.543 | − 0.505 | *− 0.045* | *− 0.092* | − 0.282 | − 0.383 | − 0.416 | *− 0.055* | *− 0.022* |
| 5.00 | 0.00 | *0.000* | 3.554 | − 2.164 | − 1.222 | − 0.368 | − 1.245 | − 1.824 | − 2.076 | *− 0.120* | *− 0.201* |
| 10.00 | 0.00 | *0.006* | 3.568 | − 3.716 | − 3.181 | − 1.889 | − 2.428 | − 3.420 | − 3.725 | − 0.672 | − 1.055 |
| 0.25 | 0.75 | *0.003* | 3.557 | 0.270 | *0.011* | *0.027* | *0.095* | *0.181* | *0.212* | *0.029* | *0.017* |
| 1.25 | 3.75 | *0.005* | 3.559 | 1.241 | *0.195* | *0.091* | 0.589 | 1.002 | 1.187 | *0.057* | *0.178* |
| 2.50 | 7.50 | *− 0.002* | 3.559 | 2.268 | 0.911 | 0.360 | 1.215 | 2.082 | 2.388 | *0.118* | 0.535 |
| 0.50 | 0.50 | *− 0.008* | 3.552 | *0.000* | *− 0.019* | *− 0.010* | *− 0.036* | *− 0.018* | *− 0.008* | *− 0.011* | *− 0.023* |
| 2.50 | 2.50 | *0.001* | 3.561 | *0.043* | *− 0.425* | *− 0.105* | *− 0.027* | *0.014* | *0.031* | *0.001* | − 0.390 |
| 5.00 | 5.00 | *0.006* | 3.565 | *0.081* | − 0.848 | − 0.614 | *− 0.020* | *0.048* | *0.075* | *0.001* | − 1.062 |
| 0.75 | 0.25 | *0.003* | 3.558 | *− 0.242* | *− 0.020* | *− 0.033* | *− 0.148* | *− 0.189* | *− 0.201* | *− 0.024* | *− 0.017* |
| 3.75 | 1.25 | *− 0.005* | 3.551 | − 1.095 | − 0.849 | *− 0.232* | − 0.641 | − 0.931 | − 1.066 | *− 0.059* | − 0.623 |
| 7.50 | 2.50 | *0.001* | 3.555 | − 1.917 | − 2.164 | − 1.315 | − 1.243 | − 1.817 | − 2.029 | *− 0.131* | − 2.053 |

**Table 10** (continued)

RMSE: Best outcomes in italics for each outlier setting excluding Comp

| %X | %Y | Comp | LD | Ratio | M− 1 | M− 2 | Med | Trim | Wins | C− 1 | C− 2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.00 | 0.00 | 0.515 | 3.608 | *0.522* | *0.522* | 0.523 | 0.531 | 0.523 | 0.523 | *0.522* | 0.523 |
| 0.00 | 1.00 | 0.516 | 3.611 | 0.747 | *0.525* | *0.525* | 0.577 | 0.639 | 0.672 | 0.527 | 0.528 |
| 0.00 | 5.00 | 0.517 | 3.606 | 2.565 | 1.238 | 0.733 | 1.324 | 2.107 | 2.447 | *0.536* | 1.104 |
| 0.00 | 10.00 | 0.519 | 3.613 | 4.757 | 3.195 | 2.054 | 2.613 | 4.275 | 4.763 | *0.873* | 2.612 |
| 1.00 | 0.00 | 0.517 | 3.598 | 0.730 | 0.526 | 0.532 | 0.603 | 0.650 | 0.671 | 0.528 | *0.525* |
| 5.00 | 0.00 | 0.511 | 3.607 | 2.233 | 1.337 | 0.637 | 1.354 | 1.902 | 2.148 | *0.531* | 0.555 |
| 10.00 | 0.00 | 0.517 | 3.623 | 3.764 | 3.237 | 1.977 | 2.491 | 3.470 | 3.773 | *0.858* | 1.182 |
| 0.25 | 0.75 | 0.517 | 3.612 | 0.594 | *0.525* | 0.526 | 0.542 | 0.556 | 0.567 | 0.526 | *0.525* |
| 1.25 | 3.75 | 0.506 | 3.611 | 1.354 | 0.555 | 0.524 | 0.790 | 1.132 | 1.304 | *0.518* | 0.563 |
| 2.50 | 7.50 | 0.513 | 3.613 | 2.343 | 1.064 | 0.639 | 1.329 | 2.159 | 2.461 | *0.534* | 0.804 |
| 0.50 | 0.50 | 0.510 | 3.605 | 0.521 | *0.520* | *0.520* | 0.529 | *0.520* | *0.520* | *0.520* | 0.521 |
| 2.50 | 2.50 | 0.520 | 3.616 | 0.542 | 0.681 | 0.540 | 0.536 | 0.531 | 0.536 | *0.527* | 0.683 |
| 5.00 | 5.00 | 0.517 | 3.619 | 0.547 | 1.001 | 0.821 | 0.533 | 0.532 | 0.547 | *0.524* | 1.202 |
| 0.75 | 0.25 | 0.516 | 3.612 | 0.578 | *0.524* | *0.524* | 0.551 | 0.557 | 0.562 | *0.524* | 0.526 |
| 3.75 | 1.25 | 0.521 | 3.605 | 1.220 | 1.009 | 0.582 | 0.836 | 1.073 | 1.195 | *0.531* | 0.890 |
| 7.50 | 2.50 | 0.518 | 3.609 | 1.995 | 2.236 | 1.438 | 1.354 | 1.898 | 2.104 | *0.542* | 2.163 |

*%X* indicates the percentage of outliers in *x*. *%Y* indicates the percentage of outliers in *y*

**Table 11** Population 2 (gamma distribution: d = 5.13, g = 0.50), MAR2

Bias: |value| < 0.850 in italics

| %X | %Y | Comp | LD | Ratio | M− 1 | M− 2 | Med | Trim | Wins | C− 1 | C− 2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.00 | 0.00 | *0.014* | − 9.034 | *0.001* | *− 0.691* | − 1.708 | − 1.805 | *− 0.649* | *− 0.307* | *− 0.757* | − 2.144 |
| 0.00 | 1.00 | *0.005* | − 9.045 | 3.495 | *0.022* | *− 0.460* | *− 0.633* | 1.468 | 2.383 | *− 0.094* | *0.318* |
| 0.00 | 5.00 | *0.010* | − 9.038 | 16.481 | 8.858 | 3.141 | 4.328 | 14.617 | 17.067 | 1.116 | 9.489 |
| 0.00 | 10.00 | *0.024* | − 9.055 | 30.925 | 23.359 | 15.065 | 11.390 | 30.937 | 32.176 | 22.167 | 22.512 |
| 1.00 | 0.00 | *0.002* | − 9.053 | − 2.441 | − 2.436 | − 2.475 | − 2.818 | − 2.293 | − 2.305 | − 0.954 | − 2.352 |
| 5.00 | 0.00 | *0.011* | − 9.024 | − 8.659 | − 8.715 | − 8.302 | − 6.049 | − 8.854 | − 9.202 | − 2.032 | − 2.506 |
| 10.00 | 0.00 | *− 0.018* | − 9.061 | − 12.794 | − 13.173 | − 13.849 | − 9.109 | − 13.456 | − 13.326 | − 12.791 | − 8.939 |
| 0.25 | 0.75 | *0.014* | − 9.028 | 1.925 | *− 0.654* | − 0.892 | − 1.187 | *0.479* | 1.117 | *− 0.381* | *0.007* |
| 1.25 | 3.75 | *0.007* | − 9.018 | 8.323 | 1.756 | − 1.300 | 1.142 | 7.427 | 9.308 | *0.443* | 4.937 |
| 2.50 | 7.50 | *− 0.033* | − 9.069 | 14.298 | 7.495 | 1.409 | 3.873 | 15.193 | 14.948 | 0.941 | 10.359 |
| 0.50 | 0.50 | *0.009* | − 9.047 | *0.409* | − 1.273 | − 1.364 | − 1.738 | *− 0.473* | *− 0.069* | *− 0.638* | *− 0.314* |
| 2.50 | 2.50 | *− 0.041* | − 9.065 | 1.560 | − 3.388 | − 4.007 | − 1.643 | *0.435* | 1.356 | *− 0.366* | *− 0.467* |
| 5.00 | 5.00 | *− 0.006* | − 9.047 | 2.627 | − 3.182 | − 6.585 | − 1.445 | 1.692 | 2.506 | *− 0.352* | − 2.198 |
| 0.75 | 0.25 | *− 0.002* | − 9.050 | − 1.047 | − 1.862 | − 1.887 | − 2.301 | − 1.396 | − 1.204 | *− 0.830* | − 0.843 |
| 3.75 | 1.25 | *0.024* | − 9.022 | − 3.942 | − 6.601 | − 5.976 | − 3.919 | − 5.059 | − 5.140 | *− 0.789* | − 3.683 |
| 7.50 | 2.50 | *0.009* | − 9.053 | − 6.064 | − 9.829 | − 10.442 | − 5.601 | − 7.876 | − 6.627 | − 2.397 | − 9.689 |

**Table 11** (continued)

RMSE: Best outcomes in italics for each outlier setting excluding Comp

| %X | %Y | Comp | LD | Ratio | M-1 | M-2 | Med | Trim | Wins | C-1 | C-2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.00 | 0.00 | 1.703 | 9.236 | *1.923* | 2.039 | 2.562 | 2.745 | 2.033 | 1.952 | 2.099 | 2.874 |
| 0.00 | 1.00 | 1.740 | 9.253 | 4.080 | *1.967* | 2.013 | 2.222 | 2.492 | 3.134 | 2.009 | 2.098 |
| 0.00 | 5.00 | 1.725 | 9.239 | 16.860 | 9.232 | 3.822 | 4.889 | 14.922 | 17.440 | *2.277* | 10.047 |
| 0.00 | 10.00 | 1.728 | *9.258* | 31.467 | 23.828 | 15.471 | 11.685 | 31.408 | 32.722 | 22.842 | 23.094 |
| 1.00 | 0.00 | 1.729 | 9.256 | 3.098 | 3.093 | 3.131 | 3.482 | 2.979 | 2.984 | *2.212* | 3.037 |
| 5.00 | 0.00 | 1.727 | 9.227 | 8.882 | 8.935 | 8.531 | 6.349 | 9.054 | 9.410 | *2.835* | 3.162 |
| 10.00 | 0.00 | 1.737 | 9.269 | 12.955 | 13.330 | 14.003 | 9.296 | 13.601 | 13.479 | 12.965 | *9.184* |
| 0.25 | 0.75 | 1.752 | 9.239 | 2.794 | 2.064 | 2.154 | 2.420 | 2.049 | 2.297 | 2.050 | *2.009* |
| 1.25 | 3.75 | 1.739 | 9.225 | 8.744 | 2.730 | 2.356 | 2.438 | 7.792 | 9.716 | *2.036* | 5.503 |
| 2.50 | 7.50 | 1.730 | 9.275 | 14.759 | 7.994 | 2.627 | 4.478 | 15.573 | 15.397 | *2.211* | 10.831 |
| 0.50 | 0.50 | 1.721 | 9.250 | 1.999 | 2.305 | 2.365 | 2.700 | 1.997 | *1.948* | 2.085 | 1.987 |
| 2.50 | 2.50 | 1.730 | 9.273 | 2.711 | 3.931 | 4.451 | 2.643 | *2.020* | 2.551 | 2.023 | 2.237 |
| 5.00 | 5.00 | 1.734 | 9.253 | 3.692 | 3.837 | 6.893 | 2.510 | 2.784 | 3.583 | *2.044* | 3.353 |
| 0.75 | 0.25 | 1.722 | 9.253 | 2.188 | 2.669 | 2.695 | 3.086 | 2.372 | 2.265 | *2.154* | *2.154* |
| 3.75 | 1.25 | 1.734 | 9.230 | 4.423 | 6.890 | 6.287 | 4.403 | 5.408 | 5.516 | *2.128* | 4.561 |
| 7.50 | 2.50 | 1.754 | 9.261 | 6.431 | 10.043 | 10.638 | 5.932 | 8.124 | 6.952 | *3.300* | 9.953 |

*%X* indicates the percentage of outliers in *x*. *%Y* indicates the percentage of outliers in *y*

**Table 12** Population 3 (Gamma distribution, $d = 13.78$, $g = 0.25$), MAR2

Bias: |value| < 0.850 in italics

| %X | %Y | Comp | LD | Ratio | M-1 | M-2 | Med | Trim | Wins | C-1 | C-2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.00 | 0.00 | *− 0.001* | − 9.050 | *− 0.008* | − 0.882 | − 1.963 | − 1.422 | *− 0.491* | *− 0.229* | − 1.069 | − 2.381 |
| 0.00 | 1.00 | *− 0.004* | − 9.052 | 3.146 | *− 0.068* | *− 0.845* | *− 0.148* | 1.605 | 2.357 | *− 0.569* | *0.123* |
| 0.00 | 5.00 | *− 0.019* | − 9.049 | 14.968 | 8.255 | 3.028 | 5.217 | 13.677 | 15.534 | *0.821* | 8.567 |
| 0.00 | 10.00 | *0.001* | − 9.063 | 28.059 | 21.264 | 13.845 | 12.698 | 28.479 | 29.237 | 19.041 | 20.408 |
| 1.00 | 0.00 | *− 0.012* | − 9.054 | − 2.433 | − 2.689 | − 2.781 | − 2.477 | − 2.124 | − 2.207 | − 1.223 | − 2.457 |
| 5.00 | 0.00 | *− 0.011* | − 9.047 | − 8.643 | − 8.865 | − 8.677 | − 5.826 | − 8.730 | − 9.121 | − 2.430 | − 2.691 |
| 10.00 | 0.00 | *− 0.032* | − 9.066 | − 12.759 | − 13.227 | − 13.983 | − 8.965 | − 13.342 | − 13.237 | − 12.909 | − 9.771 |
| 0.25 | 0.75 | *− 0.034* | − 9.077 | 1.656 | *− 0.818* | − 1.306 | *− 0.784* | *0.611* | 1.118 | *− 0.839* | *− 0.239* |
| 1.25 | 3.75 | *− 0.012* | − 9.041 | 7.263 | 1.565 | − 1.380 | 1.784 | 6.909 | 8.229 | *− 0.074* | 4.530 |
| 2.50 | 7.50 | *− 0.008* | − 9.035 | 12.468 | 6.502 | 1.128 | 4.768 | 13.765 | 13.086 | *0.685* | 9.372 |
| 0.50 | 0.50 | *− 0.006* | − 9.068 | *0.233* | − 1.483 | − 1.764 | − 1.353 | *− 0.329* | *− 0.031* | − 1.028 | − 0.592 |
| 2.50 | 2.50 | *− 0.008* | − 9.021 | 0.998 | − 3.316 | − 4.185 | − 1.104 | *0.400* | *0.840* | *− 0.701* | *− 0.318* |
| 5.00 | 5.00 | *0.015* | − 9.029 | 1.625 | − 3.412 | − 6.554 | − 0.891 | 1.131 | 1.516 | *− 0.342* | − 2.580 |
| 0.75 | 0.25 | *− 0.003* | − 9.048 | − 1.121 | − 2.085 | − 2.240 | − 1.907 | − 1.225 | − 1.120 | − 1.140 | − 1.136 |
| 3.75 | 1.25 | *− 0.009* | − 9.046 | − 4.240 | − 6.664 | − 6.364 | − 3.628 | − 4.959 | − 5.171 | − 1.138 | − 3.455 |
| 7.50 | 2.50 | *− 0.015* | − 9.067 | − 6.509 | − 9.804 | − 10.636 | − 5.297 | − 7.889 | − 7.016 | − 3.907 | − 9.749 |

**Table 12** (continued)

RMSE: Best outcomes in italics for each outlier setting excluding Comp

| %X | %Y | Comp | LD | Ratio | M-1 | M-2 | Med | Trim | Wins | C-1 | C-2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.00 | 0.00 | 1.741 | 9.260 | *1.986* | 2.163 | 2.779 | 2.647 | 2.073 | 2.011 | 2.286 | 3.103 |
| 0.00 | 1.00 | 1.726 | 9.259 | 3.770 | *1.986* | 2.139 | 2.242 | 2.603 | 3.123 | 2.082 | 2.094 |
| 0.00 | 5.00 | 1.725 | 9.261 | 15.310 | 8.624 | 3.735 | 5.760 | 13.968 | 15.872 | *2.199* | 9.090 |
| 0.00 | 10.00 | 1.730 | *9.269* | 28.498 | 21.658 | 14.215 | 12.992 | 28.869 | 29.681 | 19.631 | 20.926 |
| 1.00 | 0.00 | 1.739 | 9.268 | 3.130 | 3.329 | 3.413 | 3.309 | 2.909 | 2.955 | *2.374* | 3.168 |
| 5.00 | 0.00 | 1.719 | 9.255 | 8.870 | 9.084 | 8.902 | 6.173 | 8.941 | 9.336 | *3.157* | 3.344 |
| 10.00 | 0.00 | 1.724 | 9.275 | 12.919 | 13.383 | 14.136 | *9.171* | 13.489 | 13.390 | 13.079 | 9.997 |
| 0.25 | 0.75 | 1.730 | 9.293 | 2.626 | 2.151 | 2.376 | 2.394 | 2.138 | 2.325 | 2.198 | *2.071* |
| 1.25 | 3.75 | 1.738 | 9.256 | 7.654 | 2.605 | 2.441 | 2.927 | 7.275 | 8.608 | *2.017* | 5.109 |
| 2.50 | 7.50 | 1.719 | 9.242 | 12.841 | 6.957 | 2.417 | 5.321 | 14.093 | 13.451 | *2.133* | 9.817 |
| 0.50 | 0.50 | 1.743 | 9.284 | *2.007* | 2.471 | 2.653 | 2.616 | 2.054 | 2.008 | 2.273 | 2.113 |
| 2.50 | 2.50 | 1.734 | 9.237 | 2.310 | 3.876 | 4.621 | 2.460 | *2.020* | 2.217 | 2.114 | 2.231 |
| 5.00 | 5.00 | 1.734 | 9.242 | 2.834 | 4.008 | 6.875 | 2.346 | 2.384 | 2.754 | *2.047* | 3.497 |
| 0.75 | 0.25 | 1.717 | 9.254 | *2.234* | 2.838 | 2.960 | 2.890 | 2.312 | 2.240 | 2.289 | 2.303 |
| 3.75 | 1.25 | 1.743 | 9.260 | 4.698 | 6.965 | 6.670 | 4.207 | 5.337 | 5.563 | *2.305* | 4.370 |
| 7.50 | 2.50 | 1.739 | 9.278 | 6.833 | 10.022 | 10.832 | 5.677 | 8.142 | 7.312 | *4.756* | 10.003 |

*%X* indicates the percentage of outliers in *x*. *%Y* indicates the percentage of outliers in *y*

**Table 13** Population 4 (Uniform distribution: [0.1, 2.1]), MAR2

Bias: |value| < 0.040 in italics

| %X | %Y | Comp | LD | Ratio | M-1 | M-2 | Med | Trim | Wins | C-1 | C-2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.00 | 0.00 | *0.000* | − 0.556 | *0.000* | *0.000* | *0.000* | *− 0.032* | *− 0.020* | *− 0.011* | *0.000* | *0.000* |
| 0.00 | 1.00 | *0.000* | − 0.557 | 0.097 | *0.014* | *0.016* | 0.040 | 0.069 | 0.082 | *0.014* | *0.018* |
| 0.00 | 5.00 | *0.001* | − 0.557 | 0.453 | 0.239 | 0.112 | 0.340 | 0.438 | 0.449 | 0.066 | 0.257 |
| 0.00 | 10.00 | *0.001* | − 0.556 | 0.841 | 0.614 | 0.409 | 0.746 | 0.866 | 0.846 | 0.377 | 0.590 |
| 1.00 | 0.00 | *0.000* | − 0.558 | − 0.060 | *− 0.037* | *− 0.022* | − 0.100 | − 0.085 | − 0.071 | *− 0.012* | *− 0.021* |
| 5.00 | 0.00 | *− 0.001* | − 0.557 | − 0.248 | − 0.215 | − 0.142 | − 0.321 | − 0.288 | − 0.261 | − 0.086 | − 0.082 |
| 10.00 | 0.00 | *− 0.001* | − 0.558 | − 0.409 | − 0.391 | − 0.340 | − 0.525 | − 0.460 | − 0.424 | − 0.263 | − 0.239 |
| 0.25 | 0.75 | *0.001* | − 0.556 | 0.057 | *− 0.001* | *0.009* | *0.004* | *0.029* | 0.043 | *0.008* | *0.006* |
| 1.25 | 3.75 | *− 0.001* | − 0.558 | 0.255 | 0.073 | *0.006* | 0.142 | 0.223 | 0.246 | *0.025* | 0.097 |
| 2.50 | 7.50 | *0.000* | − 0.557 | 0.460 | 0.250 | 0.100 | 0.306 | 0.443 | 0.457 | 0.100 | 0.244 |
| 0.50 | 0.50 | *0.000* | − 0.557 | *0.017* | *− 0.015* | *0.001* | *− 0.032* | *− 0.011* | *0.003* | *0.001* | *− 0.003* |
| 2.50 | 2.50 | *0.000* | − 0.557 | 0.076 | − 0.056 | − 0.060 | *− 0.030* | *0.033* | 0.062 | *− 0.004* | *− 0.037* |
| 5.00 | 5.00 | *− 0.001* | − 0.560 | 0.129 | − 0.047 | − 0.123 | *− 0.030* | 0.084 | 0.121 | *− 0.019* | − 0.046 |
| 0.75 | 0.25 | *0.000* | − 0.558 | *− 0.022* | *− 0.027* | *− 0.010* | − 0.067 | − 0.049 | − 0.035 | *− 0.006* | *− 0.008* |
| 3.75 | 1.25 | *0.000* | − 0.558 | − 0.092 | − 0.152 | − 0.102 | − 0.183 | − 0.139 | − 0.109 | *− 0.035* | − 0.124 |
| 7.50 | 2.50 | *0.000* | − 0.557 | − 0.157 | − 0.260 | − 0.252 | − 0.297 | − 0.217 | − 0.173 | − 0.108 | − 0.266 |

**Table 13** (continued)

RMSE: Best outcomes in italics for each outlier setting excluding Comp

| %X | %Y | Comp | LD | Ratio | M-1 | M-2 | Med | Trim | Wins | C-1 | C-2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.00 | 0.00 | 0.078 | 0.564 | *0.082* | *0.082* | *0.082* | 0.094 | 0.084 | 0.083 | *0.082* | 0.083 |
| 0.00 | 1.00 | 0.079 | 0.565 | 0.127 | *0.084* | 0.085 | 0.099 | 0.107 | 0.116 | 0.085 | 0.086 |
| 0.00 | 5.00 | 0.079 | 0.565 | 0.462 | 0.254 | 0.141 | 0.353 | 0.446 | 0.458 | *0.108* | 0.275 |
| 0.00 | 10.00 | 0.079 | 0.564 | 0.849 | 0.622 | 0.420 | 0.754 | 0.873 | 0.853 | *0.389* | 0.602 |
| 1.00 | 0.00 | 0.079 | 0.565 | 0.103 | 0.091 | 0.086 | 0.134 | 0.119 | 0.110 | *0.084* | 0.086 |
| 5.00 | 0.00 | 0.078 | 0.565 | 0.262 | 0.231 | 0.166 | 0.333 | 0.301 | 0.275 | 0.119 | *0.116* |
| 10.00 | 0.00 | 0.077 | 0.565 | 0.418 | 0.401 | 0.352 | 0.532 | 0.468 | 0.432 | 0.278 | *0.254* |
| 0.25 | 0.75 | 0.079 | 0.564 | 0.101 | *0.084* | *0.084* | 0.090 | 0.088 | 0.094 | *0.084* | *0.084* |
| 1.25 | 3.75 | 0.078 | 0.566 | 0.269 | 0.111 | *0.084* | 0.169 | 0.238 | 0.261 | 0.086 | 0.131 |
| 2.50 | 7.50 | 0.080 | 0.565 | 0.469 | 0.265 | *0.133* | 0.321 | 0.452 | 0.467 | 0.134 | 0.265 |
| 0.50 | 0.50 | 0.078 | 0.565 | 0.085 | 0.084 | *0.083* | 0.095 | *0.083* | *0.083* | *0.083* | *0.083* |
| 2.50 | 2.50 | 0.079 | 0.565 | 0.113 | 0.102 | 0.104 | 0.094 | 0.090 | 0.104 | *0.084* | 0.094 |
| 5.00 | 5.00 | 0.078 | 0.568 | 0.156 | 0.098 | 0.151 | 0.094 | 0.119 | 0.149 | *0.087* | 0.105 |
| 0.75 | 0.25 | 0.078 | 0.566 | 0.086 | 0.087 | *0.083* | 0.111 | 0.096 | 0.090 | *0.083* | *0.083* |
| 3.75 | 1.25 | 0.078 | 0.565 | 0.125 | 0.174 | 0.133 | 0.203 | 0.162 | 0.138 | *0.090* | 0.153 |
| 7.50 | 2.50 | 0.079 | 0.565 | 0.179 | 0.275 | 0.268 | 0.309 | 0.233 | 0.193 | *0.138* | 0.281 |

*%X* indicates the percentage of outliers in *x*. *%Y* indicates the percentage of outliers in *y*

**Table 14** Population 5 (normal distribution: mean = 20, variance = 16), MAR2

Bias: |value| < 0.250 in italics

| *%X* | *%Y* | Comp | LD | Ratio | M-1 | M-2 | Med | Trim | Wins | C-1 | C-2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.00 | 0.00 | *0.002* | − 3.553 | *0.002* | *0.002* | *0.002* | *− 0.035* | *− 0.011* | *− 0.006* | *0.002* | *0.001* |
| 0.00 | 1.00 | *0.008* | − 3.547 | 0.723 | *0.054* | *0.062* | 0.293 | 0.509 | 0.579 | *0.076* | *0.106* |
| 0.00 | 5.00 | *− 0.001* | − 3.559 | 3.379 | 1.526 | 0.703 | 1.599 | 2.772 | 3.213 | *0.174* | 1.340 |
| 0.00 | 10.00 | *0.002* | − 3.557 | 6.321 | 4.231 | 2.691 | 3.367 | 5.700 | 6.328 | 1.187 | 3.446 |
| 1.00 | 0.00 | *0.006* | − 3.555 | − 0.666 | *− 0.043* | *− 0.105* | − 0.343 | − 0.499 | − 0.552 | *− 0.061* | *− 0.011* |
| 5.00 | 0.00 | *0.004* | − 3.553 | − 2.888 | − 1.683 | − 0.525 | − 1.512 | − 2.450 | − 2.770 | *− 0.138* | *− 0.214* |
| 10.00 | 0.00 | *− 0.002* | − 3.554 | − 4.922 | − 4.255 | − 2.617 | − 2.845 | − 4.550 | − 4.935 | − 0.847 | − 1.294 |
| 0.25 | 0.75 | *− 0.003* | − 3.558 | 0.356 | *0.005* | *0.028* | *0.119* | *0.241* | 0.279 | *0.034* | *0.032* |
| 1.25 | 3.75 | *− 0.005* | − 3.568 | 1.664 | 0.265 | *0.122* | 0.755 | 1.358 | 1.591 | *0.071* | 0.402 |
| 2.50 | 7.50 | *0.001* | − 3.555 | 3.032 | 1.238 | 0.512 | 1.550 | 2.799 | 3.178 | *0.198* | 0.944 |
| 0.50 | 0.50 | *0.001* | − 3.554 | *0.016* | *− 0.011* | *0.002* | *− 0.032* | *− 0.005* | *0.003* | *0.004* | *0.002* |
| 2.50 | 2.50 | *− 0.002* | − 3.558 | *0.055* | − 0.575 | *− 0.145* | *− 0.032* | *0.016* | *0.037* | *0.009* | *− 0.318* |
| 5.00 | 5.00 | *0.000* | − 3.559 | *0.098* | − 1.140 | − 0.832 | *− 0.027* | *0.057* | *0.089* | *0.025* | − 1.165 |
| 0.75 | 0.25 | *0.005* | − 3.555 | − 0.326 | *− 0.026* | *− 0.042* | *− 0.184* | − 0.252 | − 0.275 | *− 0.029* | *− 0.010* |
| 3.75 | 1.25 | *− 0.010* | − 3.568 | − 1.466 | − 1.162 | − 0.327 | − 0.796 | − 1.261 | − 1.429 | *− 0.072* | − 0.662 |
| 7.50 | 2.50 | *− 0.004* | − 3.554 | − 2.536 | − 2.889 | − 1.794 | − 1.469 | − 2.418 | − 2.674 | *− 0.162* | − 2.480 |

**Table 14** (continued)

RMSE: Best outcomes in italics for each outlier setting excluding Comp

| %X | %Y | Comp | LD | Ratio | M-1 | M-2 | Med | Trim | Wins | C-1 | C-2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.00 | 0.00 | 0.509 | 3.605 | 0.519 | 0.519 | 0.519 | 0.535 | 0.519 | 0.519 | 0.519 | *0.512* |
| 0.00 | 1.00 | 0.514 | 3.600 | 0.897 | *0.526* | *0.526* | 0.615 | 0.730 | 0.782 | 0.531 | 0.537 |
| 0.00 | 5.00 | 0.512 | 3.612 | 3.437 | 1.629 | 0.889 | 1.691 | 2.833 | 3.274 | *0.552* | 1.487 |
| 0.00 | 10.00 | 0.515 | 3.611 | 6.374 | 4.294 | 2.770 | 3.421 | 5.751 | 6.382 | *1.318* | 3.559 |
| 1.00 | 0.00 | 0.516 | 3.609 | 0.852 | 0.527 | 0.536 | 0.643 | 0.727 | 0.765 | 0.530 | *0.526* |
| 5.00 | 0.00 | 0.510 | 3.606 | 2.946 | 1.775 | 0.741 | 1.606 | 2.513 | 2.831 | *0.538* | 0.563 |
| 10.00 | 0.00 | 0.513 | 3.607 | 4.965 | 4.306 | 2.692 | 2.901 | 4.594 | 4.978 | *1.004* | 1.401 |
| 0.25 | 0.75 | 0.508 | 3.610 | 0.632 | *0.518* | 0.519 | 0.549 | 0.572 | 0.590 | 0.520 | 0.521 |
| 1.25 | 3.75 | 0.516 | 3.621 | 1.759 | 0.599 | 0.545 | 0.934 | 1.464 | 1.689 | *0.534* | 0.687 |
| 2.50 | 7.50 | 0.517 | 3.609 | 3.098 | 1.367 | 0.746 | 1.648 | 2.865 | 3.244 | *0.568* | 1.145 |
| 0.50 | 0.50 | 0.508 | 3.607 | 0.522 | *0.519* | *0.519* | 0.538 | *0.519* | 0.520 | *0.519* | 0.520 |
| 2.50 | 2.50 | 0.514 | 3.612 | 0.543 | 0.787 | 0.551 | 0.540 | 0.528 | 0.536 | *0.524* | 0.681 |
| 5.00 | 5.00 | 0.514 | 3.613 | 0.567 | 1.264 | 1.006 | 0.542 | 0.543 | 0.568 | *0.526* | 1.319 |
| 0.75 | 0.25 | 0.512 | 3.609 | 0.619 | *0.523* | 0.524 | 0.570 | 0.581 | 0.592 | 0.524 | 0.525 |
| 3.75 | 1.25 | 0.512 | 3.622 | 1.565 | 1.289 | 0.626 | 0.963 | 1.370 | 1.530 | *0.528* | 0.968 |
| 7.50 | 2.50 | 0.513 | 3.607 | 2.602 | 2.950 | 1.899 | 1.568 | 2.485 | 2.738 | *0.550* | 2.615 |

*%X* indicates the percentage of outliers in *x*. *%Y* indicates the percentage of outliers in *y*

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s42081-022-00164-0.

**Acknowledgements** Based on the Statistics Act (Japan), the author obtained the anonymized data of the 2004 National Survey of Family Income and Expenditure from the National Statistics Center (NSTAC). Note that the analyses in this article are the author's own and are different from the officially published results by the Japanese government. For the information on the extended use (secondary use) of official statistics in Japan, see https://www.soumu.go.jp/english/dgpp_ss/seido/2jiriyou.htm.

## Declarations

**Conflict of interest** No competing interest.

## References

Allison, P.D. (2002). *Missing data*. Sage Publications.
Bartholomew, D.J., Steele, F., Moustaki, I., & Galbraith, J.I. (2002). *The analysis and interpretation of multivariate data for social scientists*. Chapman & Hall/CRC.
Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons.
Bonate, P.L. (2011). *Pharmacokinetic-pharmacodynamic modeling and simulation* (2nd ed.). Springer.
Bullock, E. L., Nolte, C., Reboredo, S., Ana, L., & Woodcock, C. E. (2020). Ongoing forest disturbance in Guatemala's protected areas. *Remote Sensing in Ecology and Conservation, 6*(2), 141–152. https://doi.org/10.1002/rse2.130
Carpenter, J. R., & Kenward, M. G. (2013). *Multiple imputation and its application*. John Wiley & Sons.
Carsey, T.M. & Harden, J.J. (2014). *Monte Carlo simulation and resampling methods for social science*. Sage Publications.
Cochran, W.G. (1977). *Sampling techniques* (3rd ed.). John Wiley & Sons.
Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics, 19*(1), 15–18. https://doi.org/10.2307/1268249
de Waal, T. (2013). Selective editing: A quest for efficiency and data quality. *Journal of Official Statistics, 29*(4), 473–488. https://doi.org/10.2478/jos-2013-0036
de Waal, T., Pannekoek, J., & Scholtus, S. (2011). *Handbook of statistical data editing and imputation*. John Wiley & Sons.
DeGroot, M.H. & Schervish, M.J. (2002). *Probability and statistics* (3rd ed.). Addison-Wesley.
Deville, J.-C., & Särndal, C.-E. (1994). Variance estimation for the regression imputed Horvitz-Thompson estimator. *Journal of Official Statistics, 10*(4), 381–394.
Di Zio, M., & Guarnera, U. (2013). A contamination model for selective editing. *Journal of Official Statistics, 29*(4), 539–555. https://doi.org/10.2478/jos-2013-0039
Eisenhauer, J. G. (2003). Regression through the origin. *Teaching Statistics, 25*(3), 76–80. https://doi.org/10.1111/1467-9639.00136
Enders, C. K. (2010). *Applied missing data analysis*. The Guilford Press.
Filliben, J.J. & Heckert, A. (2013). Exploratory data analysis. In C. Croarkin, P. Tobias, & J.J. Filliben (Eds.), *NIST/SEMATECH e-Handbook of Statistical Methods*. https://doi.org/10.18434/M32189.
Fox, J. (2020). *Regression diagnostics: An introduction* (2nd Ed.). Sage Publications.
Ghosh-Dastidar, B., & Schafer, J. L. (2006). Outlier detection and editing procedures for continuous multivariate data. *Journal of Official Statistics, 22*(3), 487–506.
Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549–576. https://doi.org/10.1146/annurev.psych.58.110405.085530
Gujarati, D.N. (2003). *Basic econometrics* (4th ed.). McGraw-Hill.
Gwet, J. P., & Rivest, L. P. (1992). Outlier resistant alternatives to the ratio estimator. *Journal of the American Statistical Association, 87*(420), 1174–1182. https://doi.org/10.1080/01621459.1992.10476275

Haziza, D., & Vallée, A. (2020). Variance estimation procedures in the presence of singly imputed survey data: A critical review. *Japanese Journal of Statistics and Data Science, 3*(2), 583–623. https://doi.org/10.1007/s42081-020-00083-y

Hoenig, J. M., Jones, C. M., Pollock, K. H., Robson, D. S., & Wade, D. L. (1997). Calculation of catch rate and total catch in roving surveys of anglers. *Biometrics, 53*(1), 306–317. https://doi.org/10.2307/2533116

Kennedy, P. (2003). *A guide to eonometrics* (5th ed.). Blackwell Publishing.

King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review, 95*(1), 49–69. https://doi.org/10.1017/S0003055401000235

Lawrance, A. J. (1995). Deletion influence and masking in regression. *Journal of the Royal Society, Series B, 57*(1), 181–189. https://doi.org/10.1111/j.2517-6161.1995.tb02023.x

Lee, H., Rancourt, R., & Särndal, C. E. (1994). Experiments with variance estimation from survey data with imputed values. *Journal of Official Statistics, 10*(3), 231–243.

Little, R.J.A. & Rubin, D.B. (2020). *Statistical analysis with missing data* (3rd ed.). John Wiley & Sons.

Lu, J., & Yan, Z. (2014). A class of ratio estimators of a finite population mean using two auxiliary variables. *PLoS ONE, 9*, e89538. https://doi.org/10.1371/journal.pone.0089538

Lui, K. J. (2020). Notes on use of the composite estimator: An improvement of the ratio estimator. *Journal of Official Statistics, 36*(1), 137–149. https://doi.org/10.2478/jos-2020-0007

Mair, P., & Wilcox, R. (2020). Robust statistical methods in R using the WRS2 package. *Behavior Research Methods, 52*, 464–488. https://doi.org/10.3758/s13428-019-01246-w

McCLendon, M. J. (1994). *Multiple regression and causal analysis*. Waveland Press.

Mooney, C.Z. (1997). *Monte Carlo simulation*. Sage Publications.

Mulry, M. H., Kaputa, S. J., & Thompson, K. J. (2018). Setting M-estimation parameters for detection and treatment of influential values. *Journal of Official Statistics, 34*(2), 483–501. https://doi.org/10.2478/jos-2018-0022

Mulry, M. H., Oliver, B. E., & Kaputa, S. J. (2014). Detecting and treating verified influential values in a monthly retail trade survey. *Journal of Official Statistics, 30*(4), 721–747. https://doi.org/10.2478/jos-2014-0045

Pannekoek, J. (2018). Improvements of ratio-imputation using robust statistics and machine learning-techniques, paper presented at the United Nations Economic Commission for Europe workshop on statistical data editing. Retrieved October 8, 2021, from https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2018/T6_Netherlands_PANNEKOEK_Paper.pdf.

Rao, J. N. K., & Sitter, R. R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika, 82*(2), 453–460. https://doi.org/10.1093/biomet/82.2.453

Ross, S. (2006). *A first course in probability* (7th ed.) Pearson/Prentice Hall.

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika, 57*(2), 377–387. https://doi.org/10.1093/biomet/57.2.377

Schafer, J.L. (1997). *Analysis of incomplete multivariate data*. Chapman & Hall/CRC.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147–177. https://doi.org/10.1037/1082-989X.7.2.147

Scheaffer, R.L., Mendenhall III, W., Ott, R.L., & Gerow, K.G. (2012). *Elementary survey sampling* (7th ed.). Broks/Cole.

Schenker, N., Raghunathan, T. E., Chiu, P. E., Makuc, D. M., Zhang, G., & Cohen, A. J. (2006). Multiple imputation of missing income data in the national health interview survey. *Journal of the American Statistical Association, 101*(475), 924–933. https://doi.org/10.1198/016214505000001375

Scheuren, F. (2005). Multiple imputation: How it began and continues. *The American Statistician, 59*(4), 315–319. https://doi.org/10.1198/000313005X74016

Severud, W. J., Delgiudice, G. D., & Bump, J. K. (2019). Comparing survey and multiple recruitment-mortality models to assess growth rates and population projections. *Ecology and Evolution, 9*(22), 12613–12622. https://doi.org/10.1002/ece3.5725

Sitter, R. R., & Rao, J. N. K. (1997). Imputation for missing values and corresponding variance estimation. *The Canadian Journal of Statistics, 25*(1), 61–73. https://doi.org/10.2307/3315357

Snowdon, P. (1992). Ratio methods for estimating forest biomass. *New Zealand Journal of Forestry Science, 22*(1), 54–62.

Stock, B. C., Ward, E. J., Thorson, J. T., Jannot, J. E., & Semmens, B. X. (2019). The utility of spatial model-based estimators of unobserved bycatch. *ICES Journal of Marine Science, 76*(1), 255–267. https://doi.org/10.1093/icesjms/fsy153

Takahashi, M. (2017a). Multiple ratio imputation by the EMB algorithm: Theory and simulation. *Journal of Modern Applied Statistical Methods, 16*(1), 630–656. https://doi.org/10.22237/jmasm/1493598840

Takahashi, M. (2017b). Implementing multiple ratio imputation by the EMB algorithm (R). *Journal of Modern Applied Statistical Methods, 16*(1), 657–673. https://doi.org/10.22237/jmasm/1493598900

Takahashi, M. (2017c). Statistical inference in missing data by MCMC and Non-MCMC multiple imputation algorithms: Assessing the effects of between-imputation iterations. *Data Science Journal, 16*(37), 1–17. https://doi.org/10.5334/dsj-2017-037

Takahashi, M. & Watanabe, M. (2017). *Missing data analysis: Single imputation and multiple imputation in R*. Kyoritsu Shuppan.

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*(3), 763–776. https://doi.org/10.3233/SJI-160306

van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). CRC Press.

Wada, K. (2020). Outliers in official statistics. *Japanese Journal of Statistics and Data Science, 3*(2), 669–691. https://doi.org/10.1007/s42081-020-00091-y

Wada, K. & Tsubaki, H. (2020). Robust tools for statistical data editing and imputation. In *Paper presented at the 2020 UNECE Workshop on Statistical Data Editing*. Retrieved October 8, 2021, from https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.58/2020/mtg1/SDE2020_Poster_Japan_Wada_Paper.pdf .

Wada, K. & Sakashita, K. (2017). Generalized robust ratio estimator for imputation. In *Paper presented at New Techniques and Technologies for Statistics 2017*. Retrieved October 8, 2021, from https://www.nstac.go.jp/services/society_paper/29_02_02.pdf.

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

Wang, J. F., Reis, B. Y., Hu, M. G., Christakos, G., Yang, W. Z., Sun, Q., Li, Z. J., Li, X. Z., Lai, S. J., Chen, H. Y., & Wang, D. C. (2011). Area disease estimation based on sentinel hospital records. *PLoS ONE, 6*, e23428. https://doi.org/10.1371/journal.pone.0023428

Weiss, N.A. (2005). *Introductory statistics* (7th ed.). Pearson/Addison Wesley.

Wooldridge, J.M. (2020). *Introductory econometrics: A modern approach* (7th ed.). Cengage Learning.

Young, D. S., & Mathew, T. (2015). Ratio edits based on statistical tolerance intervals. *Journal of Official Statistics, 31*(1), 77–100. https://doi.org/10.1515/jos-2015-0004

Zarnoch, S. J., & Bechtold, W. A. (2000). Estimating mapped-plot forest attributes with ratios of means. *Canadian Journal of Forest Research, 30*, 688–697.

Zou, G. H., Li, Y. F., Zhu, R., & Guan, Z. (2010). Imputation of mean of ratios for missing data and its application to PPSWR sampling. *Acta Mathematica Sinica, 26*(5), 863–874. https://doi.org/10.1007/s10114-010-6271-3