Sohil, Fariha; Sohail, Muhammad Umair; Shabbir, Javid; Gupta, Sat

Article

Jackknife winsorized variance estimator under imputed data

Statistics in Transition new series (SiTns)

Provided in Cooperation with:
Polish Statistical Association

Suggested Citation: Sohil, Fariha; Sohail, Muhammad Umair; Shabbir, Javid; Gupta, Sat (2022) : Jackknife winsorized variance estimator under imputed data, Statistics in Transition new series (SiTns), ISSN 2450-0291, Sciendo, Warsaw, Vol. 23, Iss. 2, pp. 17-32, https://doi.org/10.2478/stattrans-2022-0014

This Version is available at:
https://hdl.handle.net/10419/266305

STATISTICS IN TRANSITION new series, June 2022
Vol. 23 No. 2, pp. 17–32, DOI 10.2478/stattrans-2022-0014
Received – 12.12.2018; accepted – 25.06.2021

sciendo

Jackknife winsorized variance estimator under imputed data

Fariha Sohil[1], Muhammad Umair Sohail[2], Javid Shabbir[3], Sat Gupta[4]

ABSTRACT

In the present study, we consider the problem of missing and extreme values for the estimation of population variance. The presence of extreme values either in the study variable, or the auxiliary variable, or in both of them, can adversely affect the performance of the estimation procedure. We consider three different situations for the presence of extreme values and also consider jackknife variance estimators for the population variance by handling these extreme values under stratified random sampling. Bootstrap technique ABB is carried out to understand the relative relationship more precisely.

Key words: adjusted imputation, jackknife variance estimators, linearized jackknife, missing values, winsorized variance

2000 AMS Classification: 62D05

1. Introduction

In most social science studies, researchers often face the problem of non-response due to sensitive or embarrassing issues. For example, in the case of student grade point surveys, the students may be reluctant to provide the information about grade point average. Basically, non-response is classified into two basic complete categories: (1) Unit non-response, which occurs when either the interviewee refuses to provide the response regarding the variable of interest or the interviewee is not available. (2) Item none-response, which occurs mainly due to the sensitive or embarrassing nature of the study variable (Y). Muhamad (2016) studied the imputation of missing responses by using the higher order moments of the auxiliary variable.

[1] Department of Education, The Women University, Multan, Pakistan. E-mail: s.fariha@gmail.com.

[2] Department of Statistics, University of Narowal, Narowal, Pakistan. E-mail: umair.sohail@uon.edu.pk. ORCID: https://orcid.org/0000-0002-5440-126X.

[3] Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan. E-mail: js@qau.edu.pk. ORCID: https://orcid.org/0000-0002-0035-7072.

[4] Department of Mathematics and Statistics, University of North Carolina, Greensboro, USA.

The main goal of our current work is to consider the problem of missing at random (MAR) values in the estimation of finite population variance. When the item non-response occurs the missing values of the non-respondent class can be imputed by utilizing the available information from the respondent class. Many methods use the auxiliary information for imputing the missing value.

Rubin (1976) gave a comprehensive concept of missing values by defining terms such as missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR) values. Rubin (1978) considered the problem of inflation in estimated variance by discussing the idea of multiple imputation (MI). The suggested procedure obtains $\lambda\ (\geq 2)$ data sets by imputing the missing values under the same imputation procedure of λ times. To define the Multiple Imputation (MI) methodology by $\overline{y}_{I1}, \overline{y}_{I2}, \cdots, \overline{y}_{I\lambda}$, the λ imputed estimators for the population mean. The final imputed estimator for population mean is given by $\overline{y}_{I.} = \frac{1}{\lambda}\sum_{l=1}^{\lambda} \overline{y}_{Il}$, with estimated variance

$$v\left(\overline{y}_{I.}\right) = \frac{1}{\lambda}\sum_{l=1}^{\lambda}\left(\frac{1}{n}-\frac{1}{N}\right)s_{Il}^{2} + \frac{\lambda+1}{\lambda}\left\{\frac{1}{\lambda-1}\sum_{l=1}^{\lambda}\left(\overline{y}_{Il}-\overline{y}_{I.}\right)^{2}\right\}, \tag{1.1}$$

where s_{Il}^{2} is the sample variance for the $l-th$ imputed data set having n sample and N population size respectively. The variance estimator leads to valid inference about the parameter of interest, when the number of imputation is large, provided the imputation is proper in the sense that imputed values for the non-respondent group are obtained from the posterior distribution of the study variable (Rubin and Schenker, 1986; Mujtaba et al. 2014). The traditional imputation methods like hot deck (HD) may give the underestimate variance of $\overline{y}_{I.}$. Rubin and Schenker (1986) provided the Approximate Bayesian Bootstrapping (ABB) approach for proper variance estimation. For $l = 1, 2, 3, \cdots, \lambda$; we draw r values randomly with replacement from the r observed values and then obtain $(n-r)$ missing values from the r bootstrap donors. The resultant estimates based on the g reference distribution performed well in terms of large sample selection probability, even for λ as small as 2 or 3.

MI is a proper tool to handle the missing data but some of the major limitations are: (1) Cost for handling the multiple data sets is high as compared to single imputation, especially in complex surveys. (2) The general ABB approach for imputing the non-response, that has some issues regarding the clustering, stratification, unequal probabilities of selection, is not currently taken into account. (3) Sometimes the imputation is deterministic, missing values are obtained by the

sample of donor set and the auxiliary data. (4) For smaller values of λ, we may attain a low level of precision for the multiple imputation variance estimator (MIVE), because the last term in (1.1) approaches to zero for a small value of λ.

The main focus of this investigation is to consider the univariate statistics such as mean and total under imputation and provide some recent work on jackknife variance estimation to adjust the imputed values in the presence of extreme observations. We consider the problem of extreme values in the study variable, the auxiliary variable, or in both of them, before imputing the missing values. We consider the stratified random sampling design with commonly used imputation methods such as traditional ratio, classical linear regression method and hot deck within imputation. These imputation strategies are not proper in the sense of (Rubin and Schenker, 1986), but all of them would have the valid design based on inference about the suggested variance estimator. Recently, Chen et al. (2107) suggested an approaches to improving survey-weighted estimates by precisely weighting the survey estimates.
The aim of the study is to consider the problem of extreme values either at the upper or lower end for the precise imputation of missing responses. In present study, we proposed a jacknified Winsorized variance estimator under imputed data by discussing three different cases for the occurrence of extreme values in the field of survey sampling. These are given below:

Case I: Extreme values in study variable

Let the extreme values occur only in the study variable (Y) but not in the auxiliary variable (X). These extreme values should lead to the low correlation with the auxiliary variable which will affect the performance of the estimation procedure.

Case II: Extreme values in the auxiliary variable

Let $X_1, X_2, \cdots, X_N$ be the values of the auxiliary variable having a population mean $(\bar{X})$. Suppose the characteristics of the auxiliary information are not available but we have some relevant information. We want to utilize the auxiliary information in a significant manner for better inference. The one of the possible ways to handle this situation is to use the idea of winsorization for the valid inference about the population parameters.

Case III: Extreme values in both variables

Suppose that, extreme values occur both in the study and the auxiliary variable due to some natural or unnatural disturbance in the experiment. This irregular behaviour of the study and the auxiliary variables may lead to underestimation or overestimation of population parameters. Under such circumstances, we need to use some standard procedures for the valid inference. So, we truncate both variables by

using some specified standard procedures and the results by using a truncated set of values are more reliable as compared to the irregular observed values.

We follow the standard truncation process, where low extreme values are truncated at the first quartile and high extreme values are truncated at the third quartile.

2. Proposed procedure

Motivated by Rao (1996), we consider the problem of extreme values in both the study and the auxiliary variables under stratified random sampling for the estimation of winsorized variance.

2.1. Complete response case

Most of the daily life surveys based on well established frames are often used in stratified sampling. Let n_h be the number of sampled units selected from the $h-th$ stratum $(h=1,\ 2,\ 3,\ \cdots,L)$ such that $\sum_{h=1}^{L} n_h = n.$ For the complete response case, the usual unbiased estimator of $\overline{Y}$ is given by $\overline{\Upsilon} = \sum_{h=1}^{L} W_h \overline{\Upsilon}_h$, where W_h is the stratum weight and $\overline{\Upsilon}_h$ is the sample mean of the $h-th$ stratum after truncation. The variance of the winsorized set of values is

$$v\left(\overline{\Upsilon}\right) = \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{n_h N_h}\right) s_{\Upsilon h}^2, \tag{2.1}$$

where $s_{\Upsilon h}^2 = \frac{1}{n_h - 1}\sum_{j=1}^{n_h}\left(\Upsilon_{j_h} - \overline{\Upsilon}_h\right)^2$ and Υ_{j_h} is the truncated response of the $j-th$ respondent in the $h-th$ stratum.

The jackknife variance estimator of $\overline{\Upsilon}$ after deleting the extreme values, is given by

$$v_J\left(\overline{\Upsilon}\right) = \sum_{h=1}^{L} (n_h - 1)\left(\frac{N_h - n_h}{n_h N_h}\right) s_{\Upsilon h}^{2J}, \tag{2.2}$$

where $s_{\Upsilon h}^{2J} = \frac{1}{n_h}\sum_{j=1}^{n_h}\left\{\overline{\Upsilon}(hj) - \overline{\Upsilon}\right\}^2$ and $\overline{\Upsilon}(hj)$ is sample mean obtained after deleting the $j-th$ response from the $h-th$ stratum.

2.2. Adjusted imputed value

In the case of missing values in Υ, suppose s_h be the sample of size n_h is selected from the $h-th$ stratum having Ω_h sampled units, let r_h be the respondents and r_h' be the non-respondents units who refuse to provide the response regarding the variable of interest. So, $s_h = s_{r_h} \cup s_{r_h}'$. Let $\overline{\Upsilon}_{r_h}$ be the winsorized sample mean of s_{r_h} in $h-th$ stratum. Suppose $\Upsilon^{\#}$ is the imputed value for the $j-th$ unit in s_{r_h}'.

The estimator for the population mean is then given by

$$\overline{\Upsilon}_I = \sum_{h=1}^{L} \frac{W_h}{n_h} \left\{ \sum_{j_h \in s_{r_h}} \Upsilon_{j_h} + \sum_{j_h \in s_{r_h}'} \Upsilon_{j_h}^{\#} \right\}, \tag{2.3}$$

With deterministic approach, the jackknife variance estimator of $\overline{\Upsilon}_I$ is obtained in the usual way by deleting the respondents s_{r_h}, each of the imputed value in the $h-th$ stratum is adjusted in magnitude as $\left\{ \Upsilon_{j_h}^{\#}(hJ) - \Upsilon_{j_h}^{\#} \right\}$, where $\Upsilon_{j_h}^{\#}(hJ)$ is the imputed value for the $j-th$ non-respondent unit in $h-th$ stratum, when hJ respondent is deleted from s_h. Then, the adjusted imputed missing value is equal the "*correct*" value $\Upsilon_{j_h}^{\#}(hJ)$ if $hJ \in s_{r_h}$ and remaining values are unchanged, if the non-respondent hJ is deleted. In the case of stochastic imputation, each of the imputed value is adjusted by $E_{hJ}^{\#}\Upsilon_{j_h}^{\#} - E^{\#}\Upsilon_{j_h}^{\#}$, when $hJ \in s_{r_h}$, where $E^{\#}$ denotes the expectation with respect to the imputation procedure given the donor class and $E_{hJ}^{\#}$ is the expectation when the donor values are adjusted by removing hJ units. Note that the adjusted imputation values reflect that the donor set is changed, when the respondent set is deleted from the sample.

The imputed estimator based on the original and imputed values of the $j-th$ sample units in the $h-th$ stratum is expressed as $\overline{\Upsilon}_I^{@}(hJ)$, after deleting the hJ units. Then, the jackknife winsorized variance estimator, after ignoring the finite population correction factor, is given by (2.4)

$$\begin{aligned} v_J\left(\overline{\Upsilon}_I\right) &= \sum_{h=1}^{L} (n_h - 1) s_{\Upsilon h}^{2@}, \\ s_{\Upsilon h}^{2@} &= \frac{1}{n_h} \sum\nolimits_{j=1}^{n_h} \left\{ \overline{\Upsilon}_I^{@}(hj) - \overline{\Upsilon}_I \right\}^2. \end{aligned} \tag{2.4}$$

where

2.3. Ratio imputation

Suppose the auxiliary information is available on all the sampled units in s_h. The traditional ratio imputation procedure with winsorized data is defined as:

$$\Upsilon_{j_h}^{\#} = \left(\frac{\overline{\Upsilon}_{r_h}}{\overline{\eta}_{r_h}} \right) \eta_{j_h}, \tag{2.5}$$

for $j_h \in s_{r_h'}$. Where $\overline{\Upsilon}_{r_h}$ and $\overline{\eta}_{r_h}$ are the sample mean from the s_{r_h} respondent class in the $h-th$ stratum respectively. This imputation procedure is motivated by the fact that $\Upsilon_{j_h}^{\#}$ is the best predictor of the units which are in $s_{r_h'}$ group, under the following ratio super population model, which is given by:

$$E\left(\Upsilon_{j_h}\right) = b_h \eta_{j_h}, \quad V\left(\Upsilon_{j_h}\right) = \sigma_h^2 \eta_{j_h} \quad \text{and} \quad \mathrm{Cov}\left(\Upsilon_{j_h} \Upsilon_{k_h}\right) = 0, \tag{2.6}$$

This model also holds for s_{r_h}, if there is no selection bias. The probability of response depends upon the η_{j_h}. Sarndal (1992) has shown (2.6) as an imputation model.

Under such an approach, (2.3) will be written as:

$$\overline{\Upsilon}_{Ih} = \left(\frac{\overline{\Upsilon}_{r_h}}{\overline{\eta}_{r_h}} \right) \overline{\eta}_h \tag{2.7}$$

and hence

$$\overline{\Upsilon}_I = \sum_{h=1}^{L} W_h \left(\frac{\overline{\Upsilon}_{r_h}}{\overline{\eta}_{r_h}} \right) \overline{\eta}_h, \tag{2.8}$$

Under (2.6), the estimator $\overline{\Upsilon}_I$ is a design based model, $E\left(\overline{\Upsilon}_I\right) = \overline{Y}$, provide that the model also holds for the respondent units. For the uniform response from all strata, the (2.8) has the same properties as the two-phase separate ratio estimators.

It is readily seen that, $\Upsilon_{j_h}^{\#}(hJ) = \frac{\overline{\Upsilon}_{r_h}(hJ)}{\overline{\eta}_{r_h}(hJ)} \eta_{j_h}$, with the ratio imputation method hJ respondent units are deleted, where $\overline{\Upsilon}_{r_h}(hJ) = \frac{r\overline{\Upsilon}_{r_h} - \Upsilon_{j_h}}{r_h - 1}$ and similar for η as $\overline{\eta}_{r_h}(hJ) = \frac{r\overline{\eta}_{r_h} - \eta_{j_h}}{r_h - 1}$. Using the values, the jackknife variance estimator for $\overline{\Upsilon}_I$ is obtained from (2.4). The linearized version of the jackknife variance estimator is obtained under the model (2.6).

The linearized version of the jackknife variance estimator is helpful in estimating the variance through a computer program. This suggested method is helpful in obtaining the valid jackknife estimator under the uniform response from all strata. Let $\overline{\Upsilon}_I^{@}(hJ)-\overline{\Upsilon}_I = W_h\left\{\overline{\Upsilon}_{I_h}^{@}(hJ)-\overline{\Upsilon}_{I_h}\right\}$ be the adjusted imputed estimator for $\overline{Y}_h$. If hJ units are deleted from h^{th} stratum, we have

$$v_j\left(\overline{\Upsilon}_I\right)=\sum_{h=1}^{L}W_h^2 v_j\left(\overline{\Upsilon}_{Ih}\right) \qquad (2.9)$$

$$v_j\left(\overline{\Upsilon}_{Ih}\right)=\sum\nolimits_{j=1}^{n_h}\frac{n_h-1}{n_h}\left\{\overline{\Upsilon}_{Ih}^{@}(hJ)-\overline{\Upsilon}_{Ih}\right\}^2$$

where .

A linearized version of $v_J(\overline{\Upsilon}_I)$ is given by $v_{L.rat.}\left(\overline{\Upsilon}_I\right)=\sum_h W_h^2 v_J\left(\overline{\Upsilon}_{Ih}\right)$ with

$$v_J\left(\overline{\Upsilon}_{Ih}\right)=\left(\frac{\overline{\eta}_h}{\overline{\eta}_{r_h}}\right)^2\frac{\vartheta_h}{r_h}+2\left(\frac{\overline{\eta}_h}{\overline{\eta}_{r_h}}\right)\frac{\delta_h}{n_h}+\frac{\Delta_h}{n_h}, \qquad (2.10)$$

where $\vartheta_h=\sum_{j_h\in s_{r_h}}\frac{\zeta_{j_h}^2}{(r_h-1)}$, $\delta_h=\left(\frac{\overline{\Upsilon}_{r_h}}{\overline{\eta}_{r_h}}\right)\sum_{j_h\in s_{r_h}}\frac{\zeta_{j_h}\eta_{j_h}}{(r_h-1)}$ and

$\Delta_h=\left(\frac{\overline{\Upsilon}_{r_h}}{\overline{\eta}_{r_h}}\right)^2\sum_{j_h\in s_{r_h}}\frac{(\eta_{j_h}-\overline{\eta}_h)}{(n_h-1)}$ with $\varsigma_{j_h}=\Upsilon_{j_h}-\frac{\overline{\Upsilon}_{r_h}}{\overline{\eta}_{r_h}}\eta_{j_h}$.

Formula in (2.10) is obtained by using the following expression

$$\overline{\Upsilon}_{Ih}^{@}-\overline{\Upsilon}_{Ih}=-\frac{\overline{\Upsilon}_{r_h}(\eta_{j_h}-\overline{\eta}_h)}{\overline{\eta}_{r_h}(n_h-1)}-\left\{\frac{\overline{\eta}_h(hJ)}{\overline{\eta}_{r_h}(hJ)}\right\}\frac{\zeta_{j_h}}{(r_h-1)}$$

if $hJ\in s_{r_h}$ and equals $-\frac{\overline{\Upsilon}_{r_h}(\eta_{j_h}-\overline{\eta}_h)}{\overline{\eta}_{r_h}(n_h-1)}$ if $hJ\in s_{r_h'}$, and noting that $\frac{\overline{\eta}_h(hJ)}{\overline{\eta}_{r_h}(hJ)}=\frac{\overline{\eta}_h}{\overline{\eta}_{r_h}}$ for large r_h, where $\overline{\eta}_h(hJ)=\frac{n_h\overline{\eta}-\eta_{j_h}}{n_h-1}$.

Rao and Sitter (1992) have shown that (2.10) is a design consistent variance estimator under two phase sampling design. It shows that the jackknife variance estimator in (2.4) and linearized version of (2.4) are effective under the uniform response within each stratum.

Moreover Sarndal (1992) provided the following approximation under the model (2.4)

$$v_s(\bar{\Upsilon}_{Ih}) = \left(\frac{\bar{\eta}_h}{\eta_{r_h}}\right)^2 \frac{\vartheta_h}{r_h} + 2\left(\frac{r_h}{n_h}\right)\frac{\delta_h}{n_h} + \frac{\Delta_h}{n_h}. \tag{2.11}$$

After the relative comparison of (2.10) and (2.11), it is noted that $E(\delta_h) = 0$ for the large value of r_h, which could lead the estimators in (2.4) and (2.10) to be unbiased under the model of (2.6). Moreover, it is observed that

$$v_s\left(\bar{\Upsilon}_I\right) = \sum_h W_h^2 v_s\left(\bar{\Upsilon}_{I_h}\right) \tag{2.12}$$

is not the consistent estimator under the uniform response within each stratum. The adjustment of the finite population correction (fpc) is shown by Rao (1996) comprehensively. There is no simple relation to adjust the finite correction in (2.4), but Rao and Sitter (1995) use some internal relationships to recover the finite population correction (fpc). The modified estimator in (2.4) can be used in two-phase sampling within strata, when imputation is used; that is, when the response of non-sampled units of Υ is imputed using the first phase auxiliary information. Whitridge and Kovar (1990) considered the importance of mass imputation by utilizing it on business data. Kovar and Chen (1994) discussed the finite sample properties of (2.4) by the real life application to business survey data.

Here, we discuss the stochastic counterpart of the traditional ratio imputation. In this approach the first donor is selected from $h_0 i_0$ by using the simple random sample with replacement for s_h. Then, the ratio residual $\left(\zeta^*_{j_h}\right)$ is added to (2.5) to get the random imputation value as

$$\Upsilon^{\#}_{j_h} = \left(\frac{\bar{\Upsilon}_{r_h}}{\bar{\eta}_{r_h}}\right)\eta_{j_h} + \zeta^*_{j_h} \tag{2.13}$$

Noting that $E^{\#}\left(\zeta^*_{j_h}\right) = 0$, the resultant ratio estimator is unbiased for $\bar{Y}$ under the model (2.6) and uniform response from all the strata.

With this ratio imputation procedure, we have

$$E^{\#}_{hJ}(\Upsilon^{\#}_{j_h}) = \left\{\frac{\bar{\Upsilon}_{r_h}(hJ)}{\bar{\eta}_{r_h}(hJ)}\right\}\eta_{j_h}, \quad \text{if } hJ \text{ respondents are deleted.} \tag{2.14}$$

and

$$E^{\#}\left(\Upsilon^{\#}_{j_h}\right)=\left(\frac{\overline{\Upsilon}_{r_h}}{\overline{\eta}_{r_h}}\right)\eta_{j_h} \tag{2.15}$$

Thus, the adjusted imputed values under the model (2.4) are, as:

$$y^{\#}_{j_h}+\left\{\frac{\overline{\Upsilon}_{r_h}(hJ)}{\overline{\eta}_{r_h}(hJ)}\right\}\eta_{j_h}-\left\{\frac{\overline{\Upsilon}_{r_h}}{\overline{\eta}_{r_h}}\right\}\eta_{j_h},$$

If hJ units are deleted and remaining units are unchanged. It is easy to express the linear version of jackknife variance estimator, is given by

$$v_L\left(\overline{\Upsilon}_I\right)=\sum_h W_h^2 v_L\left(\overline{\Upsilon}_{I_h}\right), \tag{2.16}$$

where $v_L\left(\overline{\Upsilon}_{I_h}\right)$ is simply obtained by adding a term which is obtained due the random selection from a donor set to (2.10) under the ratio estimator. The extra terms are given by

$$\left(\frac{m_h}{n_h^2}\right)\left\{2\left(\frac{\overline{\Upsilon}_{r_h}}{\overline{\eta}_{r_h}}\right)s^{*}_{\zeta\eta_h}+s^{*2}_{\zeta h}+\left(\frac{r_h}{n_h}\right)\overline{\zeta}^{*2}_h\right\},$$

where $s^{*}_{\zeta\eta_h}=\sum_{j_h\in s_h}\zeta^{*}_{j_h}\left(\eta_{j_h}-\overline{\eta}_h\right)\Big/n_h$, $s^{*2}_{\zeta_h}=\sum_{j_h\in s_h}\left(\zeta^{*}_{j_h}-\overline{\zeta}^{*}_h\right)\Big/n_h$ and $\overline{\zeta}^{*}_h=\sum_{j_h\in s_h}\zeta^{*}_{j_h}\Big/n_h$.

If auxiliary information regarding the variable of interest is unavailable then the traditional ratio imputation is reduced to the simple random imputation within each stratum. For these type of situations, Little and Rubin (1987) considered the approximate Bayesian Bootstrapping for handling it and discussed it in detail in their text in the chapter MI.

2.4. Regression imputation

Let η be observed on all the sampled units in s_h. The classical linear regression estimator is defined as:

$$\Upsilon^{\#}_{j_h}=\overline{\Upsilon}_{r_h}+\hat{\beta}_{r_h}\left(\eta_{j_h}-\overline{\eta}_{r_h}\right)\quad \text{for } j_h\in s_h \tag{2.17}$$

where $\hat{\beta}_{r_h}$ is a linear simple regression coefficient based on the respondent units in the h^{th} stratum. The imputed values $\Upsilon^{\#}_{j_h}$ are the best predictor for the unobserved units of Υ_{j_h} under the following super population model:

$$E\left(\Upsilon_{j_h}\right)=\alpha_h+\beta_h\eta_{j_h},\quad V\left(\Upsilon_{jh}\right)=\sigma_h^2,\ \text{and}\quad \operatorname{cov}\left(\Upsilon_{j_h}\Upsilon_{k_h}\right)=0, \tag{2.18}$$

provided that the model holds s_{r_h}, if there is no selection bias.

Under regression imputation, (2.3) can be written as:

$$\overline{\Upsilon}_I=\sum_{h=1}^{L}W_h\left\{\overline{\Upsilon}_{r_h}+\hat{\beta}_{r_h}\left(\overline{\eta}_h-\overline{\eta}_{r_h}\right)\right\} \tag{2.19}$$

The given estimator in (2.19) is $E(\overline{\Upsilon}_I)=\overline{Y}$ with a uniform response from strata. When the hJ item is deleted, then, the estimator $\Upsilon^{\#}_{j_h}$ is written as:

$$\Upsilon^{\#}_{j_h}(hJ)=\overline{\Upsilon}_{r_h}(hJ)+\hat{\beta}_{r_h}(hJ)\left\{\eta_{j_h}-\overline{\eta}_{r_h}(hJ)\right\} \tag{2.20}$$

where $\hat{\beta}_{r_h}(hJ)$ is the linear regression coefficient, obtained after deleting the hJ units.

Using (2.4), the linear version of $v_J(\overline{\Upsilon}_I)$ is given by $v_{L.reg.}\left(\overline{\Upsilon}_I\right)=\sum_{h=1}^{L}W_h^2v_L\left(\overline{\Upsilon}_h\right)$ with

$$v_L(\overline{\Upsilon}_h)=\frac{1}{r_h^2}\sum_{j_h\in s_{r_h}}\zeta_{j_h}^2\left\{1+\frac{\left(\eta_{j_h}-\overline{\eta}_{r_h}\right)\left(\overline{\eta}_h-\overline{\eta}_{r_h}\right)}{\tilde{s}^2_{\eta r_h}}\right\}^2+2\hat{\beta}_{r_h}\frac{1}{n_hr_h}\sum_{j_h\in s_{r_h}}\zeta_{j_h}$$

$$\left(\eta_{j_h}-\overline{\eta}_h\right)\left\{1+\frac{\left(\eta_{j_h}-\overline{\eta}_{r_h}\right)\left(\overline{\eta}_h-\overline{\eta}_{r_h}\right)}{\tilde{s}^2_{\eta r_h}}\right\}+\frac{\hat{\beta}^2_{r_h}}{n_h}\tilde{s}^2_{\eta_h}, \tag{2.21}$$

where $\zeta_{j_h}=\left(\Upsilon_{j_h}-\overline{\Upsilon}_{r_h}\right)-\hat{\beta}_{r_h}\left(\eta_{j_h}-\overline{\eta}_{r_h}\right)$, $\tilde{s}^2_{\eta r_h}=\frac{1}{r_h}\sum_{j_h\in s_{r_h}}\left(\eta_{j_h}-\overline{\eta}_{r_h}\right)^2$ and $\tilde{s}^2_{\eta_h}=\frac{1}{n_h}\sum_{j_h\in s_{n_h}}\left(\eta_{j_h}-\overline{\eta}_h\right)^2$. Following Rao and Shao (1992), we can say that both

the jackknife and its linear version for variance are approximately unbiased under the model (2.18). Sitter (1997) extended the results under multiple linear regression imputation.

Now, we have to consider the stochastic counterpart of the regression imputation for the winsorized variance estimator. Under this approach, the first donor set $h_0 j_0$ is selected by simple random sample with replacement from the s_{r_h}, independently from each stratum. Then, the regression residual $\zeta^*_{j_h} = \zeta_{h_0 j_0}$ are added to $\hat{\Upsilon}_{j_h} = \overline{\Upsilon}_{r_h} + \hat{\beta}_{r_h r_h}\left(\eta_{j_h} - \overline{\eta}_{r_h}\right)$ to get random imputed missing value $\Upsilon^{\#} = \hat{\Upsilon}_{j_h} + \zeta^*_{j_h}$, $hJ \in s_{n_h}$. Noting that $E^{\#}\left(\zeta^*_{j_h}\right) = 0$, the resultant imputed estimator would be $\overline{\Upsilon}_I$ is unbiased for $\overline{Y}$ under model (2.18), as well as it is also assumed that the probability of response is the same in all strata. So, we have

$$E^{\#}_{hJ}\Upsilon^{\#}_{j_h} = \hat{\Upsilon}_{jh}(hJ) = \overline{\Upsilon}_{r_h}(hJ) + \hat{\beta}_{r_h}(hJ)\left\{\eta_{j_h} - \overline{\eta}_{r_h}(hJ)\right\} \tag{2.22}$$

and $E^{\#}\Upsilon^{\#}_{j_h} = \hat{\Upsilon}_{j_h}$. Thus the adjusted imputed values used (2.4) for variance estimator by $\Upsilon^{\#}_{j_h} + \hat{\Upsilon}_{j_h}(hJ) - \hat{\Upsilon}_{j_h}$. If the hJ respondent units are deleted, $\hat{\Upsilon}_{j_h}(hJ)$ is given by (2.22).

A linear version of (2.4) under stochastic regression imputation is defined as:

$$v_L\left(\overline{\Upsilon}_I\right) = \sum_h W_h^2 v_L\left(\overline{\Upsilon}_{Ih}\right), \tag{2.23}$$

where $v_L(\overline{\Upsilon}_{Ih})$ is obtained by adding a term due to hot-deck imputation from the given formula in (2.21) under linear regression imputation. The extra term is given by $\frac{m_h}{n_h^2}\left\{2\hat{\beta}_{r_h} s^*_{\zeta\eta_h} + s^{*2}_{\zeta} + \frac{r_h}{n_h}\overline{\zeta}^{*2}_h\right\}$, where $s^*_{\zeta\eta_h}, s^{*2}_{\zeta_h}$ and $\overline{\zeta}^{*2}_h$ are the regression residuals.

If $V\left(\Upsilon_{j_h}\right)$ is not the same in each stratum, say $V\left(\Upsilon_{j_h}\right) = \sigma_h^2 \eta_{j_h}$ as in the ratio model (2.6), then the weighted linear regression is appropriate as compared to others. The resultant imputation estimator is unbiased for $\overline{Y}$ but it is not consistent under the uniform response within strata.

3. Numerical study

In addition to our study, here, we discuss numerical results by using bootstrap technique ABB on a real life data set. We obtained the data set from Rudolf et al. (2006), and modified the data using ABB technique and then applied the jackknife technique on the modified data set.

Data Set: In FEV.DAT.csv, the strata are created using the age group of the patients. There are three strata, which are used as imputation classes. Two variables FEV status (Y) and height (X) in inches of the patients are considered. The summary statistics of Y and X are given in Table 1.

Table 1. Summary statistics of the sample data set

Stratum (i)	N_h	$\overline{Y}_h$	$\overline{X}_h$	C_{yh}^2	C_{xh}^2	C_{yxh}^2	ρ_{yxh}^2
Stratum 1	300	2.0335	56.9610	0.0642	0.0060	0.8280	0.8280
Stratum 2	300	3.0530	64.2746	0.0536	0.0037	0.0108	0.7556
Stratum 3	54	3.6667	69.0909	0.0587	0.0026	0.0004	0.7795

For the truncation of the available data set, the procedure is defined as follow:

$$\Upsilon_{jh} = \begin{cases} Q_{1h} & \text{if } y_{jh} < Q_{1h} \\ y_{jh} & \text{if } Q_{1h} < y_{jh} < Q_{3h}, \\ Q_{3h} & \text{if } y_{jh} > Q_{3h} \end{cases} \quad \eta_{jh} = \begin{cases} Q_{1h} & \text{if } x_{jh} < Q_{1h} \\ x_{jh} & \text{if } Q_{1h} < x_{jh} < Q_{3h} \\ Q_{3h} & \text{if } x_{jh} > Q_{jh} \end{cases} \tag{3.1}$$

In Figure 1 we illustrate the original (O) behaviour of the study and the auxiliary variables respectively within each stratum. In the second row, the truncated (T) behaviour of the target study variable w.r.t the auxiliary variable is expressed. After applying the above mentioned truncation procedure, we observed that the correlation coefficient in the first two strata is decreased but in the third stratum it improved significantly.

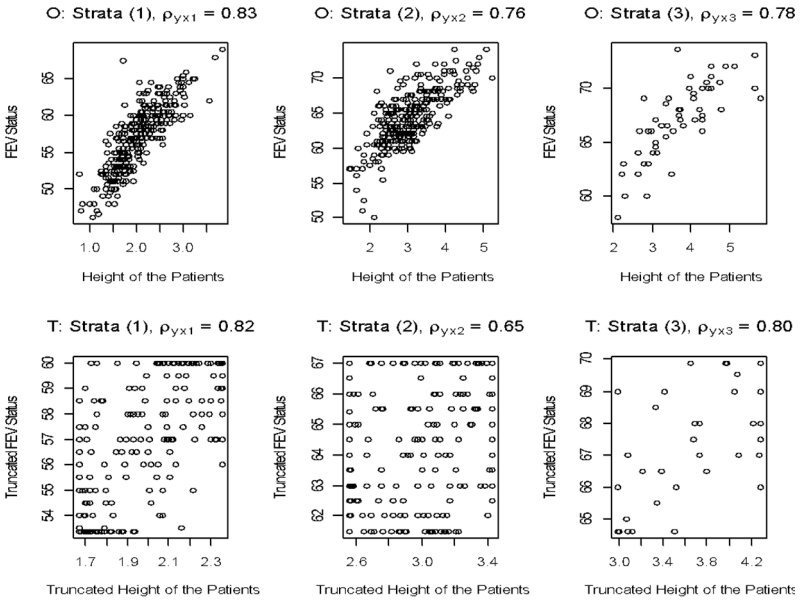

Figure 1. Data illustration within strata

Table 2. Variance of the suggested estimators

Response Rate								Variances				
Strata (i)												
1			2			3						
n_1	r_1		n_2	r_2		n_3	r_3	$v\left(\bar{y}_{l.}\right)$	$v_j\left(\bar{\Upsilon}_I\right)$	$v_{L.rat.}\left(\bar{\Upsilon}_I\right)$	$v_s\left(\bar{\Upsilon}_I\right)$	$v_{L.reg.}\left(\bar{\Upsilon}_I\right)$
80	20		80	20		20	5	0.000000049	0.015270	0.002447	0.000271	0.000026
	40			40			10	0.000000046	0.015039	0.000245	0.000063	0.000022
	60			60			15	0.000000026	0.015019	0.000066	0.000037	0.000018
160	40		160	40		30	10	0.000000180	0.010799	0.000246	0.000031	0.000017
	80			80			15	0.000000159	0.010702	0.000063	0.000016	0.000011
	120			120			20	0.000000145	0.010324	0.000024	0.000011	0.000008
240	60		240	60		40	15	0.000000051	0.009220	0.000069	0.000010	0.000007
	120			120			20	0.000000046	0.008899	0.000025	0.000006	0.000004
	180			180			25	0.000000025	0.008031	0.000012	0.000005	0.000002

Based on the sampled information with hot deck imputation for non-respondent, we consider jackknife winsorized variance estimation with imputed data by adjusting the imputed values. The winsorized variance of a given data set is 0.045130.

A different jackknife and its linearized version of estimators is considered under the ABB approach. In Table 2, the variance of the different versions of the jackknife estimator is given under a different response rate. It is clearly noticed that, the linearized version of the regression estimator under the jackknife technique outperforms as compared to others.

4. Discussion

In our present study, we discussed jackknife winsorized variance estimation based on the single imputed value. We also modified the linearized version of the jackknife variance estimator suggested by Rao (1996) for the precise estimation of winsorized variance, which is helpful for computer programs that use linearized methods for the estimation of variance. We discussed the stratified sampling scheme, because it is commonly used in large scale socio-economic surveys. We used the traditional ratio, classical linear regression and weighted hot deck imputation procedure within the classes. As we know, these imputation procedures are not reliable under multiple imputation but they could provide the valid design-based inference about the stated problem. For the practical application of this procedure, the available complete data set has information of the response status for each item and for the imputation group. The current existing computer algorithms are modified to implement these variance estimators without the permanent retention of the supplementary data.

For the stratified random sampling the imputed estimator for the population characteristics (say mean) is unbiased, under the ratio estimators with the same probability of response from all strata and the design model is also unbiased under the ratio super population model. Similarly, our modified procedure under the guideline of Rao (1996) is also consistent for the estimation of winsorized variance under uniform response from all the strata as well as the unbiased estimator under the ratio model. In this study, we estimate the approximate unbiasedness of the jackknife estimator, when the weighted or hot deck imputation is used to impute the missing values.

Our study is concentrated around the univariate estimation of the population parameters like mean and total under marginal imputation. For some complex population parameters like regression and correlation coefficient, marginal imputation is considered the association between variables. For the common donor hot deck imputation, we have the same donor set for this joint imputing of the non-response values by handling those problems which are being faced in the marginal imputation.

The current work can be extended under some modern methods like the Gibbs sampling for drawing the imputation values from the posterior distribution of the

non-observed values instead of common donor hot deck imputation, but we have mentioned earlier the modern methods for obtaining the significant imputation that take into account the design features which are currently under consideration.

Acknowledgement

Thanks to the editorial team and reviewers for their the expert opinion, which improve the quality of the manuscript.

References

Chen, Q., Elliott, M. R., Haziza, D., Yang, Y., Ghosh, M., Little, R. J., and Thompson, M., (2017). Approaches to improving survey-weighted estimates. *Statistical Science*, 32(2), pp. 227–248.

Fay, R. E., (1993). Valid inferences from imputed survey data, "in proceedings of the survey research methods". *Journal of the American Statistical Association*, 1, pp. 227–232.

Korn, E. L., Graubard, B. I., (2011). *Analysis of health surveys* (Vol. 323), John Wiley & Sons.

Kovar, J. G., Chen, E. J., (1994). Jackknife variance estimation of imputed survey data. *Survey Methodology*, 20, pp. 45–52.

Little, R., Rubin, D. B., (1987). *Statistical Analysis With Missing Data*, New York: Wiley.

Mujtaba, A., Ali, M. and Kohli, K., (2014). Statistical optimization and characterization of pH-independent extended-release drug delivery of cefpodoxime proxetil using Box–Behnken design. *Chemical Engineering Research and Design*, 92(1), pp. 156–165.

Mohamed, C., Sedory, S. A. and Singh, S., (2016). *Imputation using higher order moments of an auxiliary variable. Communications in Statistics-Simulation and Computation*, 46(8), pp. 6588–6617.

Rao, J., (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association*, 91(434), pp. 499–506.

Rao, J., Sitter, R. R., (1992). Jackknife variance estimation under imputation for missing survey data. *Technical Report 214 Carleton University*, Laboratory for Research in Statistics and Probability.

Rao, J. N. and Shao, J., (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79(4), pp. 811–822.

Rao, J. N., Sitter, R., (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82(2), pp. 453–460.

Rao, J. N. K., (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association*, 91(434), pp. 499–506.

Little, R. J., Rubin, D. B., (2019). *Statistical analysis with missing data,* Vol. 793, John Wiley & Sons.

Rudolf, F. J., William, W. J. and Ping, S., (2006). Regression analysis: statistical modeling of a response variable. Elsevier.

Rubin, D. B., (1976). Inference and Missing Data. *Biometrika*, 63(3), pp. 581–592.

Rubin, D. B., (1978). Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse. *Journal of the American Statistical Association*, 1, pp. 20–34.

Rubin, D. B., Schenker, N., (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81(394), pp. 366–374.

Sarndal, C.-E., (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, 18(2), pp. 241–252.

Sitter, R., (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association*, 92(438), pp. 780–787.

Whitridge, P., Kovar, J., (1990). *Use of mass imputation to estimate for subsample variables*, 1, pp. 132–137.

Wolter, K., (1985). *Introduction to Variance Estimation*, New York: Springer-Verlag.