調查研究—方法與應用/第 46 期
2021 年 4 月，頁 7-54

Research Articles

Comparative Analysis and Application of Imputed Estimators for Population Mean under Stratified Unequal Probability Sampling*

Esher Hsu**

ABSTRACT

With continuously increasing demand for accurate data, the sampling design of surveys has become more and more complex. Unequal probability sampling methods are therefore increasingly used in sample surveys. Item nonresponse is inevitable in survey practice. How to obtain unbiased estimation with data imputation for a complex survey is thus an important issue for research. Previous studies have presented some imputed estimators for equal probability sampling with uniform response. It would be worthwhile to explore the performance of

* Financial support by the National Science Council of Taiwan with project number NSC 94-2118-M-305-005 is very much appreciated. The author would like to thank Mr. Yuan-Yi Huang at National Taipei University for his assistance in the simulation programming work and to thank the referees for their thoughtful and valuable comments and suggestions.

** Associate Professor, Department of Statistics, National Taipei University, Taipei, Taiwan. 67 Section 3, Min-Sheng East Rd., Taipei 104, Taiwan, R.O.C. Tel.: +886 2 25024654. E-mail address: hsu@mail.ntpu.edu.tw

imputed estimators applied to complex surveys, such as unequal probability sampling or different missing data mechanisms. This study aims to present imputed estimators of the population mean for survey data imputed with an auxiliary variable under a stratified unequal probability sampling design, and to compare their performance in terms of different missing data mechanisms and different levels of the correlation coefficient between the auxiliary variable and the variable of interest.

By taking nonresponse and imputation into account, this study derives three imputed estimators (weighted, unweighted, and bias-adjusted imputed estimators) and their corresponding variance estimators with stratified unequal probability sampling, where missing data are imputed by ratio imputation. Six cases under different conditions (missing data mechanisms, population distribution, and sample allocation) are selected for a simulation study to compare the performance of the proposed imputed estimators in terms of relative bias and coefficient of variation. The relative bias of the variance estimators is also studied to compare the performance of the corresponding variance estimators. A practical application is performed to show how to apply the imputed estimators derived in this study to real survey data.

As expected, simulation results show that the performance of the estimators varies depending on the missing data mechanisms, population distributions, and methods of sample allocation. Simulation results indicate that the estimation precision of the imputed estimator increases as the correlation between the auxiliary variable and the variable of interest increases for all three imputed estimators. The imputed estimators perform with greater stability in cases of missing completely at random (MCAR) than in cases of missing at random (MAR).

Comparing the performance among the three imputed estimators, this study shows that in cases of high correlation between the auxiliary variable and the variable of interest, the proposed bias-adjusted estimator works well with stratified unequal probability sampling in reducing the estimation bias and the underestimation of mean square error

(MSE) due to unweighted imputation. Moreover, the variance estimator of the bias-adjusted estimator has the smallest relative bias for estimating MSE compared with the two others. The unadjusted imputed estimator with unweighted imputation may cause estimation bias, while its corresponding variance estimators may also underestimate the MSE of the estimator. However, simulation results do not reveal that the bias-adjusted estimator performs better than the imputed estimator with weighted imputation except at a high level of correlation between the auxiliary variable and the variable of interest. In practice, an auxiliary variable which has high correlation with the variable of interest, is commonly used to impute missing values to increase estimation precision. If the survey weights are unavailable and unweighted ratio imputation is used to impute missing values, the proposed bias-adjusted estimator with the corresponding variance estimator is suggested for obtaining a better estimation.

Keywords: nonresponse, ratio imputation, imputed estimator, bias-adjusted estimator, unequal probability sampling

在分層不等機率抽樣下母體均數插補估計量之比較分析及應用*

許玉雪**

摘要

調查實務上遺漏值在所難免，如何在複雜抽樣設計下結合遺漏值插補而能得到不偏估計量成為重要的研究課題。本文旨在探討分層不等機率抽樣下結合輔助變數插補遺漏值的插補估計量在不同遺漏機制

* 作者感謝國科會對本研究的經費贊助（計畫編號：NSC 94-2118-M-305-005）及研究助理黃耀億先生在程式模擬的協助，並感謝審查人的寶貴建議使本文更趨完善。

** 國立臺北大學統計系副教授。台北市民生東路三段 67 號，Tel.: +886 2 25024654，E-mail: hsu@mail.ntpu.edu.tw。

(MCAR、MAR) 及輔助變數與興趣變數之不同相關水準下的表現。本文在分層不等機率抽樣下結合比率插補法導出三種母體均數插補估計量（加權、未加權及偏誤調整）及其變異數估計量。利用插補估計量之相對偏誤及變異係數與其變異數估計量之相對偏誤，比較分析插補估計量的表現，並以一實例說明這些插補估計量如何應用於實際調查資料。模擬結果顯示，三個估計量的估計精確度都將隨著輔助變數和興趣變數相關性的增加而增加，插補估計量在 MCAR 遺漏機制表現較為穩定。本文所提偏誤調整插補估計量在輔助變數與興趣變數具有高度相關時，確可減少來自未加權的估計偏誤並降低均方誤的低估。實務上，若無權重資料可用而採未加權比率插補，本文所提的偏差調整插補估計量可用以得到較佳的估計。

關鍵詞：無回應、比率插補、插補估計量、偏誤調整估計量、不等機率抽樣

I. Introduction

The continuing demand for accurate data has caused the designing of sampling procedures to become complex. In order to obtain samples with market representativeness, the method of unequal probability sampling is increasingly employed in sample surveys. Item nonresponse is inevitable, even if a sampling survey has been designed very cautiously and interviewers have put a lot of effort into following up on nonresponse items. There are three types of missing data mechanisms: (1) missing completely at random (MCAR), (2) missing at random (MAR), and (3) non-ignorable nonresponse (Little and Rubin 1987). Cases with missing data can be imputed or simply ignored and deleted. Simply dropping cases with item nonresponse would lose some information, whereas imputation may result in an estimation bias.

In practice, how to have a complete data set with some missing information imputed and without estimation bias is always a topic for research. The imputation methods are usually categorized as single-imputation and multiple-imputation (Chen and Haziza 2019). Random hot-deck, mean imputation, regression imputation, and ratio imputation are commonly used in current surveys. Numerous studies have dealt with nonresponse and imputation items. Most consider imputed non-respondents to be observed data and use standard formulas for estimation. This may lead to bias and inconsistent estimation when there is a large proportion of missing data. Both the imputation method and corresponding estimator must be discussed with respect to the sampling design to preserve the representativeness of the sample.

Because the sampling with probability proportional to size (PPS) has been increasingly used in survey practice, a lot of work has been done in deriving estimators under PPS sampling (Rao 1966; Keeble et al. 2015). Rao (1966) proposed alternative estimators which were poorly correlated with the selection probabilities in PPS sampling schemes for multiple characteristics and showed that the proposed alternative estimators had greater relative efficiency than that of the usual estimators under a super-population model. Keeble et al. (2015) summarized the methods used for reducing selection bias and proposed a tool to choose a method to reduce selection bias. The Horvitz-Thompson estimator (Horvitz and Thompson 1952) is a general technique to estimate a finite population total when a sample is selected with unequal probabilities without replacement. Al-Jararha and Sulaiman (2020) modify the Horvitz-Thompson estimator based on the availability of the auxiliary variable and show that the modified estimator performs significantly better than the original estimator.

Variance estimation that takes nonresponse and imputation into account

has been studied by Särndal (1992), Shao and Steel (1999), Rao and Shao (1992), Skinner and Rao (2002), and Haziza and Rao (2003). Of these, Rao and Shao (1992) as well as Skinner and Rao (2002) used the jackknife technique to derive variance estimators, while Shao and Steel (1999) as well as Haziza and Rao (2003) derived linearization variance estimators. It has been proved that both the jackknife technique and the linearization technique can obtain asymptotically unbiased estimators (Skinner and Rao 2002; Haziza and Rao 2003). In addition, the variance estimators based on the linearization technique are also consistent (Shao and Steel 1999). Haziza and Rao (2003) used the delta method to derive linearization variance estimators, which are asymptotically unbiased and consistent estimators. Särndal (1992) developed an estimation of variance in terms of the sum of sampling variance and an imputation variance. Shao and Steel (1999) proposed a linearization variance estimator for Horvitz-Thompson-type estimated totals, which can be derived under either design-based approach or model-assisted approach, and are asymptotically unbiased and consistent.

Since the imputation procedure may lead to bias in standard estimators, a bias-adjusted estimator has been developed (Rao and Shao 1992; Skinner and Rao 2002; Haziza and Rao 2003; 2005). Skinner and Rao (2002) showed that the imputation procedure may lead to bias in standard estimators. They further derived a bias-adjusted estimator under simple random sampling, which extended the research ideas of Rao and Shao (1992). In addition, the paper pointed out that if the imputed values are treated as actual responses, the standard error of the estimator is usually underestimated. Jackknife variance estimators were used for both the standard imputed and bias-adjusted imputed estimators in the study. Simulation study under hot-deck imputation showed that the empirical standard deviation of the bias-adjusted estimator

was somewhat larger than that of the unadjusted one. Jackknife variance estimators derived by Skinner and Rao (2002) have addressed the underestimation problem of the standard error.

Following Skinner and Rao (2002), Haziza and Rao (2003) proposed bias-adjusted estimators of a population mean under unweighted imputation and derived linearization variance estimators as well. A simulation study is conducted to compare the performance of these methods in terms of bias and mean square error (MSE) under uniform response. The study results show that the bias-adjusted estimator performs better than the unadjusted estimator under unweighted imputation, while the variance estimator of the imputed estimator under unweighted imputation leads to serious underestimation of MSE when there is a large correlation between the variable of interest and the auxiliary variable. Haziza and Rao (2005) studied the estimation of domain totals and means under survey-weighted regression imputation for missing items by using design-based estimation with uniform response within classes and model-assisted estimation with ignorable response and an imputation model. Moreover, previous studies have compared model-based estimation with design-based estimation and emphasized the advantages of using a design-based approach (Särndal 1978; Wheeler et al. 2007; Knaub 2017). In practice, if the survey weights are unavailable and unweighted imputation is used to impute missing values, the unadjusted estimator may lead to estimation bias and underestimation of MSE (Haziza and Rao 2003).

It would be interesting to explore the imputed estimators with a design-based approach for a complex survey. This study thus intends to present imputed estimators for survey data imputed with an auxiliary variable under a stratified unequal probability sampling design, and to compare their performance in different missing data mechanisms and different levels of the

correlation coefficient between the auxiliary variable and the variable of interest. Therefore, using ratio imputation with an auxiliary variable, three imputed estimators (weighted, unweighted, and bias-adjusted imputed estimators) and their corresponding variance estimators are derived under the stratified unequal probability sampling design in section 2. In section 3, a simulation study is conducted to compare the performance of the three proposed imputed estimators in six different cases, in terms of relative bias and coefficient of variation of the estimators. The relative bias of the variance estimators is also studied to compare the performance of the corresponding variance estimators. Furthermore, a practical application is also performed in section 4 to show how to apply the imputed estimators to real data.

2. Imputed Estimators under Stratified Unequal Probability Sampling

In stratified sampling the population of N units is stratified into L strata. The population mean, $\bar{Y}$, is written as (Cochran 1977)

$$\bar{Y}=\frac{\sum_{h=1}^{L}\sum_{i=1}^{N_h} y_{hi}}{N}=\frac{\sum_{h=1}^{L} N_h\bar{Y}_h}{N}=\sum_{h=1}^{L} W_h\bar{Y}_h, \tag{1}$$

where suffix h denotes the stratum and i the unit within the stratum, while N_h, $\bar{Y}_h$, W_h are the total number of units, true mean, and stratum weight from stratum h, respectively. In this study, a probability sample is selected independently from each stratum and the imputation is done independently in each stratum. The imputed estimators are derived and shown in Appendix 1. For each stratum, the imputed estimators of the population mean, $\bar{y}_{imp}$, $\bar{y}_{imp}^{wr}$, $\bar{y}_{imp}^{uwr}$, and $\bar{y}_{imp}^{adj}$, are given by (A1.1)-(A1.3) and (A1.5) in Appendix 1,

respectively. The stratified probability sampling requires some additional notation. The following symbols all refer to stratum h.

$\bar{y}_{hr}$: Sample mean using the subset of units with responses

$\bar{y}_{imp,\,h}$: Imputed estimator of true mean with imputation

$\bar{y}_{imp,\,h}^{wr}$: Imputed estimator of true mean under weighted ratio imputation

$\bar{y}_{imp,\,h}^{uwr}$: Imputed estimator of true mean under unweighted ratio imputation

$\bar{y}_{imp,\,h}^{adj}$: Bias-adjusted estimator of true mean under unweighted ratio imputation

$v(\bar{y}_{imp,\,h}^{wr})$: Variance estimator of $\bar{y}_{imp,\,h}^{wr}$

$v(\bar{y}_{imp,\,h}^{uwr})$: Variance estimator of $\bar{y}_{imp,\,h}^{uwr}$

$v(\bar{y}_{imp,\,h}^{adj})$: Variance estimator of $\bar{y}_{imp,\,h}^{adj}$.

2.1 Estimation of Population Mean

2.1.1 Standard Estimator with Complete-case Method

Suppose that a probability sample is drawn by stratified sampling with unequal probability without replacement. The Horvitz-Thompson estimator proposed by Horvitz and Thompson (1952) is generally used with unequal probability sampling to obtain an unbiased estimator. The designed-based estimator under the sampling design of stratified sampling with unequal probability without replacement is thus expressed with the Horvitz-Thompson estimator. For each stratum we classify the probability sample (S_h) into two subsets (S_{hr} and S_{hm}), where S_{hr} is the set of respondents of size r_h, and S_{hm} is the set of non-respondents of size m_h; $r_h+m_h=n_h$. For a complete data set, an unbiased estimator of the population mean under stratified PPS sampling design, $\bar{y}_{st}$, can be written as follows (see Cochran 1977; Hedayat and Sinha 1991)

$$\bar{y}_{st}=\frac{\sum_{h=1}^{L}N_h\bar{y}_h}{N}=\sum_{h=1}^{L}W_h\bar{y}_h, \tag{2}$$

where $\bar{y}_h=\left(\sum_{i\in s_h}\frac{y_{hi}}{\pi_{hi}}\right)\Big/\left(\sum_{i\in s_h}\frac{1}{\pi_{hi}}\right)$ is the Horvitz-Thompson estimator for stratum h, and π_{hi} is the probability of including unit i in the sample for stratum h. For an incomplete data set (one with missing values), the estimator for the population mean based on the complete-case method[1] in stratified sampling, $\bar{y}_{str}$, is given by

$$\bar{y}_{str}=\sum_{h=1}^{L}W_h\bar{y}_{hr}, \tag{3}$$

where $\bar{y}_{hr}=\left(\sum_{i\in s_{hr}}\frac{y_{hi}}{\pi_{hi}}\right)\Big/\left(\sum_{i\in s_{hr}}\frac{1}{\pi_{hi}}\right)$.

2.1.2 Estimation with Data Imputation

(1) Imputed Estimator with Weighted Imputation

In this study, a ratio imputation, which takes advantage of the relationship between the auxiliary variables and the variable of interest, is used for imputing missing values. Suppose that the imputation is done independently in each stratum. Weighted ratio imputation uses $\hat{R}_r x_i$ to impute missing y_i. $\hat{R}_r=\frac{\bar{y}_r}{\bar{x}_r}$ is the ratio of the weighted means of respondents of variables y and x, where $\bar{y}_r=(\sum_{s_r}w_iy_i)/\sum_{s_r}w_i$, $\bar{x}_r=(\sum_{s_r}w_ix_i)/\sum_{s_r}w_i$, and w_i is the sampling weight of the unit i (see detailed derivation in Appendix 1). The formula of the imputed estimator with weighted ratio imputation for each stratum, $\bar{y}_{imp}^{wr}$, is thus expressed as (A1.2) in Appendix 1. Hence, the imputed estimator of the population mean with stratified probability sampling without replace-

1 The complete-case method is the method that simply drops the cases (units) with nonresponse items from the analysis and uses only complete cases.

ment under weighted ratio imputation, $\bar{y}_{st}^{wr}$ (referred to as weighted imputed estimator), is given by

$$\bar{y}_{st}^{wr} = \sum_{h=1}^{L} W_h \bar{y}_{imp,\,h}^{wr}, \tag{4}$$

where $\bar{y}_{imp,\,h}^{wr} = \hat{R}_{hr}(\sum_{S_h} w_{hi}x_{hi}) \big/ \sum_{S_h} w_{hi})$ is the imputed estimator with weighted ratio imputation in stratum h.

(2) Imputed Estimator with Unweighted Imputation

Similarly, in the case of unweighted ratio imputation, for each stratum, $\hat{R}_r^{uw}x_i$ is used to impute missing y_i, where $\hat{R}_r^{uw} = \bar{y}_r^{uw} / \bar{x}_r^{uw}$ is the ratio of the unweighted means of respondents of variables y and x, $\bar{y}_r^{uw} = (\sum_{S_r} y_i) \big/ r$ and $\bar{x}_r^{uw} = (\sum_{S_r} x_i) \big/ r$ (detailed derivation in Appendix 1). The imputed estimator with unweighted ratio imputation for each stratum, $\bar{y}_{imp}^{uwr}$, is expressed as (A1.3) in Appendix 1. Therefore, the imputed estimator of the population mean in stratified probability sampling without replacement under unweighted ratio imputation, $\bar{y}_{st}^{uwr}$ (referred to as unweighted imputed estimator), is given by

$$\bar{y}_{st}^{uwr} = \sum_{h=1}^{L} W_h \bar{y}_{imp,\,h}^{uwr}, \tag{5}$$

where $\bar{y}_{imp,\,h}^{uwr} = \left(\frac{1}{\sum_{S_h} w_{hi}}\right)\left(\sum_{S_{hr}} w_{hi}y_{hi} + \hat{R}_{hr}^{uw} \sum_{S_{hm}} w_{hi}x_{hi}\right)$ is the imputed estimator with unweighted ratio imputation in stratum h.

2.1.3 Bias-adjusted Estimator with Unweighted Imputation

Under uniform response with assumption of a design-based approach, the imputed estimator under unweighted imputation may have a bias as shown in (A1.4). A bias-adjusted estimator for each stratum under

unweighted ratio imputation, $\bar{y}_{imp}^{adj}$, is thus derived as (A1.5) to adjust the bias in equation (5). Hence, for the imputed estimator of the population mean with stratified sampling under unweighted ratio imputation, the adjusted estimator denoted as $\bar{y}_{st}^{adj}$ (referred to as bias-adjusted imputed estimator) is given by

$$\bar{y}_{st}^{adj} = \sum_{h=1}^{L} W_h \bar{y}_{imp,\,h}^{adj}, \tag{6}$$

where $\bar{y}_{imp,\,h}^{adj} = \bar{y}_{hr} + \hat{R}_{hr}^{uw}(\bar{x}_h - \bar{x}_{hr})$ is the bias-adjusted imputed estimator under unweighted ratio imputation in stratum h.

In summary, estimators of the population mean in equations (2) and (3) use only respondent data. Estimator (2) uses full sample values, while estimator (3) uses only sample values of respondents. Equations (4) and (5) are the imputed estimators based upon the whole sample values consisting of observed values and the imputed values, which are under weighted and unweighted imputation, respectively. Moreover, the bias-adjusted estimator in equation (6) is intended to adjust the bias in the imputed estimator in equation (5). The $\bar{y}_{imp}^{wr}$ is an approximately unbiased estimator of the population mean under uniform response (Haziza and Rao 2003). It is easy to verify that $\bar{y}_{st}^{wr}$ is also an approximately unbiased estimator of the population mean $\bar{Y}$ with stratified random sampling. Haziza and Rao (2003) have verified that the bias-adjusted estimator $\bar{y}_{imp}^{adj}$ is an approximately unbiased estimator of the population mean under uniform response. We would like to further compare the performance of these imputed estimators with different missing mechanisms.

2.2 Variance Estimation

2.2.1 Standard Estimator without Imputation

Under a stratified unequal sampling probability design, taking a probability sample from each stratum independently, with a complete data set, the variance of $\bar{y}_{st}$ for both proportional allocation and Neyman allocation is given by

$$V(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 V(\bar{y}_h). \tag{7}$$

The variance of the Horvitz-Thompson estimator ($\bar{y}_h$) with fixed-size is expressed as

$$V(\bar{y}_h) = \frac{1}{N_h^2} \sum_{i=1}^{N_h} \sum_{j>i}^{N_h} (\pi_{hi}\pi_{hj} - \pi_{hij}) \left(\frac{y_{hi}}{\pi_{hi}} - \frac{y_{hj}}{\pi_{hj}} \right)^2, \tag{8}$$

where π_{hij} is the probability that the i^{th} and j^{th} units are both in the sample of stratum h (see Cochran, 1977). The Sen-Yates-Grundy estimator of $V(\bar{y}_h)$ is written as

$$\upsilon_{SYG}(\bar{y}_h) = \frac{1}{N_h^2} \sum_{i=1}^{n_h} \sum_{j>i}^{n_h} \frac{\pi_{hi}\pi_{hj} - \pi_{hij}}{\pi_{hij}} \left(\frac{y_{hi}}{\pi_{hi}} - \frac{y_{hj}}{\pi_{hj}} \right)^2. \tag{9}$$

Hence, the estimator of $V(\bar{y}_{st})$ is given as

$$\upsilon(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \upsilon_{SYG}(\bar{y}_h). \tag{10}$$

For an incomplete data set, the estimator of the population mean based on the complete-case method is expressed as $\bar{y}_{str}$ in equation (3). The variance of $\bar{y}_{str}$ is usually estimated by the standard formula as equation (10). The variance estimator of $\bar{y}_{str}$ can be written as

$$v(\bar{y}_{str}) = \sum_{h=1}^{L} W_h^2 v_{SYG}(\bar{y}_{hr}), \tag{11}$$

where $\bar{y}_{hr}$ is defined as in equation (3).

2.2.2 Estimation with Imputation for Missing Survey Data

Since imputation does not reproduce the true value of the nonresponse item, the imputation will therefore increase the variance of an estimated mean (Namboodiri 1978). The overall variance may contain two components: a sampling variance and a variance due to imputation (Särndal 1992). Using the variance proposed by Fay (1991), the variance of $\bar{y}_{imp}$ under ratio imputation is expressed as

$$V(\bar{y}_{imp}) = V_1(\bar{y}_{imp} - \bar{Y}) + V_2(\bar{y}_{imp} - \bar{Y}), \tag{12}$$

where $V_1(.)$ is the variance with respect to sampling design, and $V_2(.)$ denotes the variance with respect to the response mechanism. $V(\bar{y}_{imp})$ is estimated by $v_t = v_1 + v_2$, where v_1 and v_2 are the estimators of $V_1(.)$ and $V_2(.)$, respectively. As mentioned above, both the jackknife technique and the linearization technique have been used to derive the variance estimator of the imputed estimator, while the linearization technique can obtain asymptotically unbiased and consistent estimators (Haziza and Rao 2003). We thus use linearization variance estimation with the delta method to derive the variance estimator for each of the imputed estimators ($\bar{y}_{imp}^{wr}$, $\bar{y}_{imp}^{uwr}$, $\bar{y}_{imp}^{adj}$) as shown in Appendix 2.

In the case of survey data with missing items, if the imputation is done independently in each stratum, the variance estimators of $\bar{y}_{st}^{wr}$, $\bar{y}_{st}^{uwr}$, and $\bar{y}_{st}^{adj}$ can be expressed as follows.

$$v(\bar{y}_{st}^{wr}) = \sum_{h=1}^{L} W_h^2 v(\bar{y}_{imp,\,h}^{wr}) = v_1^{wr} + v_2^{wr}, \tag{13}$$

$$v(\bar{y}_{st}^{uwr}) = \sum_{h=1}^{L} W_h^2 v(\bar{y}_{imp,\,h}^{uwr}) = v_1^{uwr} + v_2^{uwr}, \tag{14}$$

$$v(\bar{y}_{st}^{adj}) = \sum_{h=1}^{L} W_h^2 v(\bar{y}_{imp,\,h}^{adj}) = v_1^{adj} + v_2^{adj}. \tag{15}$$

That is, for each stratum, the $v(\bar{y}_{imp,\,h}^{wr})$, $v(\bar{y}_{imp,\,h}^{uwr})$, and $v(\bar{y}_{imp,\,h}^{adj})$ can be respectively estimated by $v_1^{wr}+v_2^{wr}$, $v_1^{uwr}+v_2^{uwr}$, and $v_1^{adj}+v_2^{adj}$. A general formula $v_1 = v(\sum_s w_i q_i)$ is used for estimating v_1^{wr}, v_1^{uwr}, and v_1^{adj}, the first component in equations (13)-(15). For unequal probability sampling, v_1 is estimated with a Sen-Yates-Grundy estimator and given by

$$v_1 = \frac{1}{N^2} \sum_{i=1}^{n} \sum_{j>i}^{n} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{q_i}{\pi_i} - \frac{q_j}{\pi_j} \right)^2, \tag{16}$$

where π_i is the probability of unit i being included in the sample, and π_{ij} is the joint probability of inclusion of units i and j in the sample. The component q_i for each estimator is derived as shown in Appendix 2. It gives $q_i = \frac{1}{\sum_s w_i}(q_{1i} - \bar{y}_{imp}^{wr})$ for v_1^{wr}, $q_i = \frac{1}{\sum_s w_i}(q_{2i} - \bar{y}_{imp}^{uwr})$ for v_1^{uwr}, and $q_i = \frac{1}{\sum_s w_i}(q_{3i} - \bar{y}_{imp}^{adj})$ for v_1^{adj}, where q_{1i}, q_{2i}, and q_{3i} are expressed as in (A2.1), (A2.3), and (A2.5).

Next, the estimators of the second component in equations (13)-(15), v_2^{wr}, v_2^{uwr}, and v_2^{adj}, are derived in Appendix 2 and presented by (A2.2), (A2.4), and (A2.6), respectively, which give

$$v_2^{wr} = \frac{\hat{p}(1-\hat{p}) \left(\frac{\hat{x}}{\hat{x}_a} \right)^2 S_{er}^2}{\hat{N}}, \tag{17}$$

$$v_2^{uwr} = \frac{\hat{p}(1-\hat{p}) \left[S_{er(1)}^2 + \left(\frac{\hat{x} - \hat{x}_a}{\hat{x}_{au}} \right)^2 S_{er(2)}^2 + 2 \left(\frac{\hat{x} - \hat{x}_a}{\hat{x}_{au}} \right)^2 S_{er(3)}^2 \right]}{\hat{N}}, \tag{18}$$

$$v_2^{adj}=\hat{p}(1-\hat{p})\frac{\hat{N}}{\sum_s w_i a_i}[s_{yr}^2+(\hat{R}_r^{uw})^2 s_{xr}^2-2\hat{R}_r^{uw}s_{yxr}+(\hat{h}\sum_s w_i a_i)^2 s_{er(2)}^2$$
$$+2(\hat{h}\sum_s w_i a_i)(s_{eyr}-\hat{R}_r^{uw}s_{exr})], \quad (19)$$

where $\hat{p}=\frac{\sum_{s_r} w_i}{\sum_s w_i}$, $\hat{x}=\sum_s w_i x_i$, $\hat{x}_a=\sum_s w_i a_i x_i$, $\hat{N}=\sum_s w_i$, $s_{er}^2=\frac{\sum_s w_i a_i(y_i-\hat{R}_i x_i)^2}{\sum_s w_i a_i}$, $\hat{x}_{au}=\sum_s a_i x_i$, $s_{er(1)}^2=\frac{\sum_s w_i a_i(y_i-\hat{R}_r^{uw}x_i)^2}{\sum_s w_i a_i}$, $s_{er(2)}^2=\frac{\sum_s w_i^{-1} a_i(y_i-\hat{R}_r^{uw}x_i)^2}{\sum_s w_i a_i}$, $s_{er(3)}^2=\frac{\sum_s a_i(y_i-\hat{R}_r^{uw}x_i)^2}{\sum_s w_i a_i}$, $\hat{h}=\frac{(\bar{x}-\bar{x}_r)}{\sum_s a_i x_i}$, $s_{yr}^2=\frac{\sum_s w_i a_i(y_i-\bar{y}_r)^2}{\sum_s w_i a_i}$, $s_{xr}^2=\frac{\sum_s w_i a_i(x_i-\bar{x}_r)^2}{\sum_s w_i a_i}$, $s_{yxr}^2=\frac{\sum_s w_i a_i(x_i-\bar{x}_r)(y_i-\bar{y}_r)}{\sum_s w_i a_i}$, $s_{er(2)}^2=\frac{\sum_s w_i^{-1} a_i(y_i-\hat{R}_r^{uw}x_i)^2}{\sum_s w_i a_i}$, $s_{eyr}=\frac{\sum_s a_i(y_i-\hat{R}_r^{uw}x_i)(y_i-\bar{y}_r)}{\sum_s w_i a_i}$, $s_{exr}=\frac{\sum_s a_i(y_i-\hat{R}_r^{uw}x_i)(x_i-\bar{x}_r)}{\sum_s w_i a_i}$.

The sum of (16) and (17) gives $v(\bar{y}_{st}^{wr})$ in (13), the estimator of the overall variance of $\bar{y}_{st}^{wr}$. Similarly, the estimator $v(\bar{y}_{st}^{uwr})$ in (14) is estimated by the sum of (16) and (18), while the estimator $v(\bar{y}_{st}^{adj})$ in (15) is estimated by the sum of (16) and (19).

3. Simulation Study

3.1 Simulation Procedure

A Monte Carlo simulation experiment is conducted using R programming language. The simulation procedure includes three steps: (1) several data sets consisting of three strata with some missing values are generated; (2) the missing values in each stratum were imputed independently with the ratio imputation methods proposed by this study; (3) the performance among three proposed imputed estimators of the population mean and cor-

responding variance estimators are compared.

Two types of population were generated in this study. Each population with size $N=1500$ consists of three strata with $N_1=550$, $N_2=450$, and $N_3=500$. The values (x_i, y_i) in each population were generated according to the ratio model $y=\beta x+\varepsilon$ with correlation between x and y equal to 0.9, 0.7, 0.5, 0.3, and 0.1 to examine the performance of estimators in response to the correlation. For population type I, x and ε were independently generated from a normal distribution such that $x\sim N(\mu_i, \sigma=5)$ with $\mu_i=30, 50, 100$ for $i=1, 2, 3$, respectively; $\varepsilon\sim N(0, 1)$. For population type II, x and ε were generated from $x\sim N(\mu_i, \sigma_i)$ with $(\mu_i, \sigma_i)=(30, 5), (40, 10)$, and $(50, 15)$ for $i=1, 2, 3$, respectively; $\varepsilon\sim N(0, 1)$. We assume that x has the same variance in different strata for the type I population (referred to as the homoscedastic population) but different variance in different strata for the type II population (referred to as the heteroscedastic population).

Stratified PPS sampling is used to draw the probability samples without replacement under both proportional allocation and Neyman allocation. With proportional allocation, 10,000 PPS samples, each of size $n=150$ and consisting of three strata (with size $n_1=55$, $n_2=45$, and $n_3=50$), are taken from each population according to Sampford's PPS sampling method (Sampford 1967), using variate x as the measure of size. In order to compare the performance among proposed imputed estimators under proportional and Neyman allocation, Neyman allocation is also used for the simulation study on the type II population. Under Neyman allocation, the sample size $n=150$ was partitioned into three strata of size $n_1=28$, $n_2=46$, and $n_3=76$.

A response rate of 0.7 is commonly used in imputation studies to indicate a certain degree of missing values, so it is used here for simulation studies. In each stratum, nonresponse to item y was generated from each PPS

sample with a response rate of 0.7[2] according to the missing data mechanism and x is taken from all units in the sample. In this study, simulations were conducted for both missing data mechanisms MCAR and MAR. For each stratum, samples satisfy three conditions: (1) partial missing values in y; (2) no missing values in x; (3) y_i and y_j are independent for $i \neq j$. For the cases of MCAR, nonresponse to item y was generated based on a uniform response mechanism. For the cases of MAR, the following algorithms were used to generate data missing at random[3]

$$r_i = \begin{cases} 0, \text{ for } U_i \geq P_{70} \\ 1, \text{ otherwise} \end{cases}, \tag{20}$$

$$U_i = \alpha x_i + \varepsilon_i, \tag{21}$$

where U_i is a random variable depending on the value of standardized x_i and a standard normal error ε_i; α is a constant. In order to meet a response rate of 0.7, y_i is deleted when r_i is zero (as U_i is larger than its 70th percentile, P_{70}).

With two population types, two methods of sample allocation to strata, and two missing data mechanisms, there are six cases of simulation conducted in this study as shown in Table 1.

Each simulation experiment consists of 10,000 stratified PPS samples. For each sample, values of $\bar{y}_{st}^{wr}$, $\bar{y}_{st}^{uwr}$, $\bar{y}_{st}^{adj}$ and their corresponding estimated variances of $v(\bar{y}_{st}^{wr})$, $v(\bar{y}_{st}^{uwr})$, $v(\bar{y}_{st}^{adj})$ are calculated based on equations (4)–(6) and (13)–(15), respectively.

Usually bias and MSE are used to compare the performance of the estimators. Since the parameters may differ in different cases, here we use rel-

---

2 A response rate of 0.7 is also used in Haziza and Rao (2003).

3 There are similar examples in Little (1992) and Nittner (2003).

Table 1 Settings of the simulation experiment

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
Population type	I	II	II	I	II	II
Missing data mechanism	MCAR	MCAR	MCAR	MAR	MAR	MAR
Sample allocation	Proportional	Proportional	Neyman	Proportional	Proportional	Neyman

ative bias and the coefficient of variation (CV) instead of bias and MSE to compare the performance of the estimators. Hence, the imputed estimators are compared in terms of their relative bias and CV. Relative bias of the variance estimators is also used to compare the performance of the variance estimators. Let θ be a finite population parameter and $\hat{\theta}$ be its estimator. The relative bias of $\hat{\theta}$ denoted $RB(\hat{\theta})$ is given by

$$RB(\hat{\theta})=\frac{E(\hat{\theta})-\theta}{\theta}, \tag{22}$$

and the CV of $\hat{\theta}$ denoted $CV(\hat{\theta})$ is given by

$$CV(\hat{\theta})=\frac{\sqrt{MSE(\hat{\theta})}}{\theta}. \tag{23}$$

The relative bias of the variance estimator, $v(\hat{\theta})$, is then defined by

$$RB(v(\hat{\theta}))=\frac{E(v(\hat{\theta}))-MSE(\hat{\theta})}{MSE(\hat{\theta})}. \tag{24}$$

The generated population varies with different cases under different levels of correlation between x and y, thereby obtaining different population parameters (such as population mean and population variance). The relative bias is calculated by dividing bias by the target population param-

eter as shown in equation (22) to avoid influence from the size of the parameter value. The relative bias instead of bias is thus used for comparing the bias among the estimators. Accordingly, the relative bias expressed in equation (22) is used to measure the bias of the imputed estimators for estimating the population mean ($\bar{Y}$) and to compare the performance among the imputed estimators. Similarly, the relative bias of the variance estimator in equation (24) is used to measure the bias of the variance estimator and to compare the performance among the variance estimators. Moreover, CV computed by dividing the standard deviation by the mean is a useful statistic for comparing the variability of variables that have different deviations and different means. CV in equation (23) is computed by dividing the square root of MSE by the population mean. To avoid influence from the size of the population mean, the CV instead of MSE is used in this study for comparing the relative efficiency among the imputed estimators.

The relative bias of the $\bar{y}_{st}^{wr}$, $\bar{y}_{st}^{uwr}$, and $\bar{y}_{st}^{adj}$ are thus denoted $RB(\bar{y}_{st}^{wr})$, $RB(\bar{y}_{st}^{uwr})$, and $RB(\bar{y}_{st}^{adj})$, respectively, while the CV of the $\bar{y}_{st}^{wr}$, $\bar{y}_{st}^{uwr}$, and $\bar{y}_{st}^{uwr}$ are respectively denoted $CV(\bar{y}_{st}^{wr})$, $CV(\bar{y}_{st}^{uwr})$, and $CV(\bar{y}_{st}^{adj})$. Similarly, $RB(v(\bar{y}_{st}^{wr}))$, $RB(v(\bar{y}_{st}^{uwr}))$, and $RB(v(\bar{y}_{st}^{adj}))$ denote the relative bias of the $v(\bar{y}_{st}^{wr})$, $v(\bar{y}_{st}^{uwr})$, and $v(\bar{y}_{st}^{adj})$, respectively.

3.2 Simulation Results

In this section, the relative bias and CV of the imputed estimators are calculated to compare their performance, while the relative bias of the variance estimators is also studied to compare the performance of the corresponding variance estimators. Values of the above measures are calculated from the generated 10,000 stratified PPS samples for each of the six simulation cases at different levels of the correlation coefficient (ρ_{xy}). The

simulation results of the six cases are summarized in Tables 2-7 and also displayed graphically in Appendix 3. The relative bias of imputed estimators, CV of the imputed estimators, and the relative bias of the variance estimators are displayed in Figures A3.1-A3.3 of Appendix 3, respectively. The values given in Tables 2-7 are respectively presented by cases 1-6 in Figures A3.1-A3.3 to show the performance pattern of the estimators (denoted by weighted, unweighted, and bias-adjusted) along with the increase of the correlation coefficient between the auxiliary variable and the variable of interest.

3.2.1 Missing Completely at Random (MCAR)

Simulation results under MCAR for cases 1-3 are reported in Tables 2-4, respectively and graphically displayed with case 1, case 2, and case 3 in Figures A3.1-A3.3. As seen in Tables 2-4 and Figure A3.1, under MCAR, the simulation results clearly show that the absolute relative bias of the unweighted imputed estimator ($\bar{y}_{st}^{uwr}$) is larger than that of the other two imputed estimators ($\bar{y}_{st}^{wr}$ and $\bar{y}_{st}^{adj}$) for all values of ρ_{xy}. The simulation results also show that no significant difference exists on the relative bias between the weighted imputed estimators and the bias-adjusted estimator for all values of ρ_{xy}. That implies that the bias-adjusted estimator may reduce the bias due to unweighted imputation. The simulation results show that all three imputed estimators may slightly overestimate the population mean for large ρ_{xy}, but underestimate the population mean for small ρ_{xy}.

Simulation results on the relative bias of the variance estimator show that all three variance estimators may underestimate MSE. The variance estimator of the bias-adjusted estimator ($v(\bar{y}_{st}^{adj})$) has a relatively small bias for all values of ρ_{xy}, and this is significantly smaller than that of the other two estimators ($v(\bar{y}_{st}^{wr})$ and $v(\bar{y}_{st}^{uwr})$) as shown in Tables 2-4 and Figure A3.3. The

negative bias of the variance estimator will respond to the underestimate of MSE. As seen in Tables 2–4 and Figure A3.2, the simulation results clearly show the CV ratios decrease along with the increase of ρ_{xy} for all three imputed estimators. The unweighted imputed estimator ($\bar{y}_{st}^{uwr}$) has the smallest CV ratio, followed by the weighted imputed estimator ($\bar{y}_{st}^{wr}$) and the bias-adjusted estimator ($\bar{y}_{imp}^{adj}$) for all values of ρ_{xy}. Moreover, the CV ratios shown in Table 2 (case 1) are significantly smaller than those shown in Tables 3 and 4 (cases 2 and 3) because the type II (heteroscedastic) population has a relatively large variance. The small $CV(\bar{y}_{st}^{uwr})$ and large negative $RB(v(\bar{y}_{st}^{uwr}))$ indicate that the variance estimator of the unweighted imputed estimator may lead to underestimation of MSE, which coincides with the problem mentioned by Skinner and Rao (2002) and Haziza and Rao (2003). In contrast, the small absolute $RB(\bar{y}_{st}^{adj})$ and small negative $RB(v(\bar{y}_{st}^{adj}))$ with the slightly high $CV(\bar{y}_{st}^{adj})$ of the bias-adjusted estimator indicate that the bias-

Table 2 Simulation results for case 1
— type I population, proportional allocation, MCAR (%)

	ρ_{yx}=0.9	ρ_{yx}=0.7	ρ_{yx}=0.5	ρ_{yx}=0.3	ρ_{yx}=0.1
$RB(\bar{y}_{st}^{wr})$	0.028	0.036	0.025	0.007	−0.022
$RB(\bar{y}_{st}^{uwr})$	0.028	0.047	0.029	−0.028	−0.073
$RB(\bar{y}_{st}^{adj})$	0.028	0.036	0.025	0.006	−0.022
$CV(\bar{y}_{st}^{wr})$	0.746	1.131	1.566	2.288	6.132
$CV(\bar{y}_{st}^{uwr})$	0.744	1.127	1.556	2.272	6.079
$CV(\bar{y}_{st}^{adj})$	0.777	1.200	1.689	2.516	6.707
$RB(v(\bar{y}_{st}^{wr}))$	−0.145	−0.099	−0.035	−0.001	−0.001
$RB(v(\bar{y}_{st}^{uwr}))$	−0.146	−0.171	−0.026	−0.015	−0.014
$RB(v(\bar{y}_{st}^{adj}))$	−0.134	−0.088	−0.030	−0.001	−0.001

Table 3 Simulation results for case 2
— type II population, proportional allocation, MCAR (%)

	$\rho_{yx}=0.9$	$\rho_{yx}=0.7$	$\rho_{yx}=0.5$	$\rho_{yx}=0.3$	$\rho_{yx}=0.1$
$RB(\bar{y}_{st}^{wr})$	0.057	0.063	0.080	0.054	−0.401
$RB(\bar{y}_{st}^{uwr})$	0.064	0.079	0.130	0.083	−0.469
$RB(\bar{y}_{st}^{adj})$	0.057	0.063	0.080	0.053	−0.400
$CV(\bar{y}_{st}^{wr})$	1.602	2.164	3.268	5.519	15.051
$CV(\bar{y}_{st}^{uwr})$	1.596	2.144	3.220	5.436	14.774
$CV(\bar{y}_{st}^{adj})$	1.664	2.292	3.490	5.978	16.360
$RB(v(\bar{y}_{st}^{wr}))$	−0.128	−0.084	−0.060	−0.010	−0.071
$RB(v(\bar{y}_{st}^{uwr}))$	−0.161	−0.136	−0.164	−0.023	−0.101
$RB(v(\bar{y}_{st}^{adj}))$	−0.119	−0.076	−0.053	−0.008	−0.060

Table 4 Simulation results for case 3
— type II population, Neyman allocation, MCAR (%)

	$\rho_{yx}=0.9$	$\rho_{yx}=0.7$	$\rho_{yx}=0.5$	$\rho_{yx}=0.3$	$\rho_{yx}=0.1$
$RB(\bar{y}_{st}^{wr})$	0.024	0.092	0.027	0.092	−0.043
$RB(\bar{y}_{st}^{uwr})$	0.031	0.107	0.079	0.115	−0.128
$RB(\bar{y}_{st}^{adj})$	0.024	0.093	0.028	0.092	−0.044
$CV(\bar{y}_{st}^{wr})$	1.794	2.385	3.645	6.044	16.878
$CV(\bar{y}_{st}^{uwr})$	1.787	2.362	3.592	5.941	16.565
$CV(\bar{y}_{st}^{adj})$	1.816	2.416	3.641	6.162	16.959
$RB(v(\bar{y}_{st}^{wr}))$	−0.018	−0.150	−0.006	−0.023	−0.001
$RB(v(\bar{y}_{st}^{uwr}))$	−0.031	−0.204	−0.048	−0.038	−0.006
$RB(v(\bar{y}_{st}^{adj}))$	−0.017	−0.147	−0.006	−0.022	−0.001

adjusted estimator may reduce the bias made by unweighted imputation, and its corresponding variance estimator may also reduce the underestimation of MSE more or less. However, the simulation results show that the CV of the bias-adjusted estimator is slightly larger than that of the weighted imputed estimator for all values of ρ_{xy}. This implies that from an efficiency perspective, the performance of the bias-adjusted estimator may not be better than the imputed estimator with weighted imputation.

As expected, the simulation results show that the performance of the imputed estimators depends on the correlation coefficients ρ_{xy} and varies with the population type and sample allocation. Under the MCAR missing mechanism, the simulation results show that the MSE of the imputed estimator may decrease as the ρ_{xy} increases for all of the three imputed estimators. This result shows that an auxiliary variable with high ρ_{xy} can be used to increase precision. However, the unweighted imputed estimator may overestimate the population mean, while its corresponding variance estimators may also underestimate MSE of the estimator if the correlation is high. The study results imply that for the cases of the MCAR, the proposed bias-adjusted estimator and its corresponding variance estimator may decrease the estimation bias and also reduce the underestimation of MSE due to unweighted imputation.

3.2.2 Missing at Random (MAR)

Simulation results under MAR for cases 4–6 are reported in Tables 5–7, respectively and graphically displayed with case 4, case 5, and case 6 in Figures A3.1–A3.3. Simulation results under MAR show that the relative bias does not reveal an apparent pattern accompanying the change of ρ_{xy}. However, the absolute relative bias of the imputed estimators is small if ρ_{xy}

>0.5 and is large if $\rho_{xy}<0.5$ as shown in Figure A3.1, cases 4-6. The results in Tables 5-7 show that the bias-adjusted estimator ($\bar{y}_{st}^{adj}$) can slightly reduce the estimation bias from unweighted imputation when ρ_{xy} is high. Like in the MCAR cases, Tables 5-7 and Figure A3.3 show that the variance estimator of the bias-adjusted estimator ($v(\bar{y}_{st}^{adj})$) has a relatively small bias compared to that of the other two estimators ($v(\bar{y}_{st}^{wr})$ and $v(\bar{y}_{st}^{uwr})$). That is, the variance estimator of the bias-adjusted estimator has the smallest relative bias for estimating MSE. Tables 5-7 and Figure A3.2 clearly show that the CV ratios of the imputed estimators decrease as ρ_{xy} increases for all three imputed estimators. Especially for the cases under MAR, the CV of the bias-adjusted estimator is smaller than the two others at $\rho_{xy}=0.9$. In general, the bias-adjusted estimator proposed by this study under MAR has better performance at a high level of correlation. Under MAR, the bias-adjusted estimator does not perform better than the imputed estimator with weighted imputation except at $\rho_{xy}=0.9$.

Table 5 Simulation results for case 4
— type I population, proportional allocation, MAR (%)

	$\rho_{yx}=0.9$	$\rho_{yx}=0.7$	$\rho_{yx}=0.5$	$\rho_{yx}=0.3$	$\rho_{yx}=0.1$
$RB(\bar{y}_{st}^{wr})$	0.012	−0.147	0.004	0.451	0.721
$RB(\bar{y}_{st}^{uwr})$	0.013	−0.142	−0.006	0.429	0.683
$RB(\bar{y}_{st}^{adj})$	0.012	−0.147	0.002	0.447	0.716
$CV(\bar{y}_{st}^{wr})$	0.755	1.153	1.620	2.403	6.400
$CV(\bar{y}_{st}^{uwr})$	0.753	1.149	1.610	2.384	6.348
$CV(\bar{y}_{st}^{adj})$	0.729	1.207	1.752	2.661	7.085
$RB(v(\bar{y}_{st}^{wr}))$	−0.025	−1.635	−0.001	−3.528	−1.268
$RB(v(\bar{y}_{st}^{uwr}))$	−0.031	−1.528	−0.002	−3.233	−1.156
$RB(v(\bar{y}_{st}^{adj}))$	−0.028	−1.481	0.000	−2.829	−1.021

Table 6 Simulation results for case 5
— type II population, proportional allocation, MAR (%)

	$\rho_{yx}=0.9$	$\rho_{yx}=0.7$	$\rho_{yx}=0.5$	$\rho_{yx}=0.3$	$\rho_{yx}=0.1$
$RB(\bar{y}_{st}^{wr})$	0.050	−0.003	−0.410	−0.190	0.278
$RB(\bar{y}_{st}^{uwr})$	0.057	0.016	−0.366	−0.207	0.257
$RB(\bar{y}_{st}^{adj})$	0.052	0.000	−0.402	−0.194	0.271
$CV(\bar{y}_{st}^{wr})$	1.622	2.232	3.443	5.773	15.940
$CV(\bar{y}_{st}^{uwr})$	1.615	2.211	3.392	5.694	15.680
$CV(\bar{y}_{st}^{adj})$	1.612	2.357	3.759	6.497	18.037
$RB(v(\bar{y}_{st}^{wr}))$	−0.095	0.000	−1.420	−0.108	−0.030
$RB(v(\bar{y}_{st}^{uwr}))$	−0.126	−0.005	−1.164	−0.132	−0.027
$RB(v(\bar{y}_{st}^{adj}))$	−0.013	0.000	−1.145	−0.089	−0.023

Table 7 Simulation results for case 6
— type II population, Neyman allocation, MAR (%)

	$\rho_{yx}=0.9$	$\rho_{yx}=0.7$	$\rho_{yx}=0.5$	$\rho_{yx}=0.3$	$\rho_{yx}=0.1$
$RB(\bar{y}_{st}^{wr})$	0.041	−0.005	−0.344	−0.223	0.448
$RB(\bar{y}_{st}^{uwr})$	0.048	0.015	−0.301	−0.242	0.394
$RB(\bar{y}_{st}^{adj})$	0.042	−0.002	−0.336	−0.227	0.435
$CV(\bar{y}_{st}^{wr})$	1.815	2.466	3.802	6.357	17.879
$CV(\bar{y}_{st}^{uwr})$	1.808	2.442	3.748	6.278	17.591
$CV(\bar{y}_{st}^{adj})$	1.789	2.570	4.083	7.070	19.920
$RB(v(\bar{y}_{st}^{wr}))$	−0.050	0.000	−0.820	−0.123	−0.063
$RB(v(\bar{y}_{st}^{uwr}))$	−0.071	−0.004	−0.643	−0.148	−0.050
$RB(v(\bar{y}_{st}^{adj}))$	−0.056	0.000	−0.678	−0.014	−0.048

For MAR cases, comparing the values in Table 5 with the values in Tables 6 and 7 we can see that the performance of the estimators has a significant difference between homoscedastic and heteroscedastic populations. Besides, Figures A3.1–3.3 clearly show a similar pattern between case 5 and case 6. That implies that the performance of imputed estimators under MAR is more affected by population distribution than by sample allocation. Although the bias-adjusted estimator under MAR may not reduce bias as stably as that under MCAR, the simulation results indicate that the bias-adjusted estimator has better performance at a high level of correlation.

In summary, performance of the proposed bias-adjusted estimator varies in different cases. For the cases of MCAR with proportional allocation, the proposed bias-adjusted estimator works very well. Comparing the simulation results of MCAR cases and MAR cases, the performance of the bias-adjusted estimator proposed by this study under MAR is not as stable as that under MCAR. The simulation results under MAR show that the relative bias of all the estimators does not reveal an apparent pattern but is unstable at different levels of correlation coefficients, ρ_{xy}. Especially for the cases of the MAR missing data mechanism and a heteroscedastic population, the performance of imputed estimators is significantly affected by the level of correlation coefficients. Furthermore, comparing the simulation results under strata with a homoscedastic population and a heteroscedastic population, the bias-adjusted estimator performs better under strata with a homoscedastic population. Thus, the performance of the bias-adjusted estimator proposed in this study depends on the missing data mechanisms, the types of the population distribution and the sample allocation used with the stratified unequal probability sampling.

In practice, we usually use auxiliary variables which have a high corre-

lation with the variable of interest, to increase the precision of estimation. The simulation results in this study clearly show that the MSE of the imputed estimator may decrease as the ρ_{xy} increases for all three imputed estimators. This study result demonstrates that auxiliary variables with high ρ_{xy} can be used to increase the estimators' precision. Moreover, although from an efficiency perspective, the simulation results did not show that the bias-adjusted estimator performs better than the imputed estimator with weighted imputation except at $\rho_{xy}=0.9$ under MAR, the variance estimator of the bias-adjusted estimator has the smallest relative bias for estimating MSE. However, in cases of high correlation between the auxiliary variable and the variable of interest, the proposed bias-adjusted estimator works well with the stratified unequal probability under an MCAR or MAR missing mechanism to reduce the estimation bias and the underestimation of MSE due to unweighted imputation.

4. Practical Application

In practice, the stratified PPS sampling design is commonly used to obtain more accurate data. How to deal with missing values in a data set is always an issue in practical surveys. This paper presents three imputed estimators (weighted, unweighted, and bias-adjusted) and their corresponding variance estimators derived with a stratified PPS sampling design under ratio imputation in section 2. A simulation study has been conducted in section 3 to compare the performance of those estimators. This section intends to show how to apply the imputed estimators derived in section 2 to a real data set collected with a complex survey under a stratified PPS sampling design.

Considering the practical use of imputing methods and corresponding

estimations, this practical study uses data taken from the sampling survey in Hsu et al. (2001), which was designed with stratified PPS sampling. The sampling survey was conducted in 2001 to estimate collection and recycling costs of recycling home appliances by interviewing collectors and recycling plants in Taiwan. Under stratified sampling with PPS, collectors were first stratified by region into four strata. Some collectors are drawn from each stratum (region) with probabilities proportional to the collector's size as measured by the collection quantity of the collector. The collection quantity provided by official statistics was used as an auxiliary variable for PPS sampling in the survey. That is, the sampling probability is proportional to the collection quantity of the collectors for each region to select collectors with higher market share.

Table A4.1 in Appendix 4 gives part of the survey data. In this data set, the cost variable has two missing values occurring in the central and eastern regions. The missing data are imputed by both weighted and unweighted ratio imputation expressed in Appendix 1. The collection quantity (x), which has high correlation ($\rho_{yx} > 0.75$) with collection cost (y), is used as an auxiliary variable for both weighted and unweighted ratio imputations to take advantage of the relationship between collection cost and collection quantity. The imputed results are shown in Table 8. Based on three imputed estimators ($\bar{y}_{st}^{wr}$, $\bar{y}_{st}^{uwr}$, and $\bar{y}_{st}^{adj}$) presented in equations (4)–(6) and their corresponding variance estimators ($v(\bar{y}_{st}^{wr})$, $v(\bar{y}_{st}^{uwr})$, and $v(\bar{y}_{st}^{adj})$) presented in equations (13)–(15), the estimated means and their corresponding estimated variances are calculated and shown in Table 9.

With respect to the estimated mean, estimation results indicate that the estimate based on the bias-adjusted imputed estimator is close to that based on the weighted imputed estimator, while the estimate obtained by

the unweighted imputed estimator is relatively low. Moreover, regarding the estimated standard error, the estimate based on the unweighted imputed estimator is less than the others. The estimates as shown in Table 9 coincide with the simulation results in section 3; that is, the standard error based on the unadjusted estimator under unweighted ratio imputation may be underestimated. The average collection cost is estimated to be around NT$31,231 with standard error of NT$9,404 based on the bias-adjusted estimator, and is estimated at around NT$31,316 with standard error of NT$8,930 based on the weighted imputed estimator.

In practice, if the survey weights are not available, we cannot estimate population mean by the weighted imputation estimator. The unadjusted imputed estimator with unweighted imputation (i.e. unweighted imputed estimator) is usually used to estimate the population mean, which may lead to biased estimation and underestimation of MSE as shown in the simulation

Table 8 Imputed results

Region	Collection quantity (unit)	Imputed collection cost (NT$)	
		Weighted	Unweighted
Central	160	22,478.30	20,208.19
Eastern	700	100,083.69	41,081.45

Table 9 The estimated mean and variance

Imputed estimators	Mean (NT$)	Standard error (NT$)
Weighted	$\bar{y}_{st}^{wr}=31{,}316$	$se(\bar{y}_{st}^{wr})=8{,}930$
Unweighted	$\bar{y}_{st}^{uwr}=30{,}786$	$se(\bar{y}_{st}^{uwr})=8{,}925$
Bias-adjusted	$\bar{y}_{st}^{adj}=31{,}231$	$se(\bar{y}_{st}^{adj})=9{,}404$

study. The bias-adjusted estimator proposed in this paper can be applied in this case to adjust the bias from unweighted imputation.

5. Conclusion

In practice, most complex surveys treat imputed missing values as observed data and use standard formulas to estimate population parameters. This may not only lead to serious bias and inconsistent estimation but also cause variance underestimation, especially when the proportion of missing values is not small. How to obtain unbiased estimation with data imputation under a complex survey is thus an important issue for research. This study takes nonresponse and imputation into account for estimating the population mean and derives three imputed estimators with corresponding variance estimators for a data set with missing values under a stratified unequal probability sampling design. Six cases are selected for study to compare the performance of the proposed estimators under different conditions (missing data mechanisms, population distribution, and sample allocation). A simulation study is conducted to see the performance of three imputed estimators in estimating the population mean with data imputation under stratified unequal probability sampling. A practical application is also presented to show how the imputed estimators work with real data.

As expected, the simulation results demonstrated that the performance of the estimators varies depending on the missing data mechanisms, population distributions, and methods of sample allocation. Obviously, the imputed estimators perform better in the case of stratified unequal probability sampling with proportional allocation under the strata with homogeneous variance. Simulation results also indicate that the imputed estimators perform

more stably in MCAR cases than in MAR cases. In practice, we usually use an auxiliary variable, which has high correlation with the variable of interest, to increase estimation precision. For all three imputed estimators, simulation results indicate that the estimation precision of the imputed estimator increases as the correlation between the auxiliary variable and the variable of interest increases.

Comparing the performance among three imputed estimators, the results of this study show that in cases of high correlation between the auxiliary variable and the variable of interest, the proposed bias-adjusted estimator works well with the stratified unequal probability sampling under MCAR or MAR missing mechanism to reduce the estimation bias and the underestimation of MSE due to unweighted imputation. Moreover, the variance estimator of the bias-adjusted estimator has the smallest relative bias for estimating MSE among the three. The unadjusted imputed estimator with unweighted imputation may cause estimation bias, while its corresponding variance estimators may also underestimate the MSE of the estimator. However, simulation results do not reveal that the bias-adjusted estimator performs better than the imputed estimator with weighted imputation except at a high level of correlation between the auxiliary variable and the variable of interest. In practice, if the survey weights are unavailable and unweighted ratio imputation is used to impute missing values, the proposed bias-adjusted estimator with the corresponding variance estimator is suggested for better estimation.

In the cases of MAR and the cases of populations with heterogeneous variance, the imputed estimators proposed in this study are unstable. Improving the imputed estimators for stratified PPS sampling with heteroscedastic population under an MAR missing data mechanism would be an interesting

topic for future studies. Moreover, this study only explores imputed estimators with single value imputation and focuses on the univariate estimation. Extension of the imputed estimator in conjunction with multiple imputations or incorporation into a multivariate estimation under a complex sample survey will be more complicated and therefore is a subject for future study.

This study focuses on the comparative analysis of imputed estimators with ratio imputation. The study results of the comparative analysis may not be changed at different response rates. Therefore, this paper takes a response rate (0.7) to represent a certain degree of missing values for the simulation study and does not conduct a comparative analysis of different response rates. However, the performance of imputed estimators may change in response to different response rates, especially when using different imputation methods. The comparative analysis of the imputed estimators with different imputation methods (such as hot-deck imputation, or regression imputation) at different response rates is also important and is left for further study.

In addition, the availability of the auxiliary variable (or control variate) is usually used for increasing estimation precision. This study applies an auxiliary variable with ratio imputation for imputing missing data. The Horvitz-Thompson estimation method is a general technique used with unequal probability sampling to obtain an unbiased estimator. Therefore, this study adopts the Horvitz-Thompson estimator to drive the imputed estimators with ratio imputation under stratified unequal probability sampling. However, the auxiliary variable here is only used in the imputation stage and is not combined with the Horvitz-Thompson estimator. A modified Horvitz-Thompson estimator based on auxiliary variables (Al-Jararha and Sulaiman 2020) may further improve the efficiency of the imputed estimator. The imputed estimator combining the Horvitz-Thompson estimator with

auxiliary variables (or control variates) is therefore of interest for further study.

Appendix 1 Imputed Estimators of Population Mean

Suppose a probability sample, s, is classified into two subsets, s_r and s_m, where s_r is the subset of respondents of size r, s_m is the subset of non-respondents of size m, and sample size $n=r+m$. The imputed estimator of the population mean after imputation, $\bar{y}_{imp}$, is given by the following (see Haziza and Rao 2003)

$$\bar{y}_{imp}=\left(\frac{1}{\sum_s w_i}\right)\left(\sum_{s_r} w_i y_i+\sum_{s_m} w_i y_i^*\right), \tag{A1.1}$$

where w_i is the sampling weight of the unit i, and y_i^* is the imputed value for missing y_i. In this study, the Horvitz-Thompson weight $w_i=\frac{1}{\pi_i}$ is used, where π_i is the probability of unit i being included in the sample.

In the case of weighted ratio imputation, a variate x_i correlated with y_i is used as an auxiliary variable. $\hat{R}_r x_i$ is used for y_i^*. The imputed estimator with weighted ratio imputation, $\bar{y}_{imp}^{wr}$ (weighted imputed estimator) is given by

$$\bar{y}_{imp}^{wr}=\hat{R}_r\left(\frac{\sum_s w_i x_i}{\sum_s w_i}\right), \tag{A1.2}$$

where $\hat{R}_r=\frac{\bar{y}_r}{\bar{x}_r}$ with $\bar{y}_r=\frac{\sum_{s_r} w_i y_i}{\sum_{s_r} w_i}$, and $\bar{x}_r=\frac{\sum_{s_r} w_i x_i}{\sum_{s_r} w_i}$.

In the case of unweighted ratio imputation, $\hat{R}_r^{uw} x_i$ is used for y_i^*. The imputed estimator with unweighted ratio imputation, $\bar{y}_{imp}^{uwr}$ (unweighted imputed estimator) is given by

$$\bar{y}_{imp}^{uwr}=\left(\frac{1}{\sum_s w_i}\right)\left(\sum_{s_r} w_i y_i+\hat{R}_r^{uw}\sum_{s_m} w_i x_i\right), \tag{A1.3}$$

where $\hat{R}_r^{uw}=\frac{\bar{y}_r^{uw}}{\bar{x}_r^{uw}}$ with $\bar{y}_r^{uw}=\frac{\sum_{s_r} y_i}{r}$, and $\bar{x}_r^{uw}=\frac{\sum_{s_r} x_i}{r}$.

The bias of $\bar{y}_{imp}^{uwr}$ under uniform response in assumption of design-based approach is estimated by

$$b(\bar{y}_{imp}^{uwr})=(1-\hat{p})\,(\hat{R}_r^{uw}\bar{x}-\bar{y}_r), \tag{A1.4}$$

where $\hat{p}$ is an estimator of the probability of response. Subtracting the bias estimator $b(\bar{y}_{imp}^{uwr})$ from $\bar{y}_{imp}^{uwr}$, the bias-adjusted estimator of $\bar{Y}$ under unweighted ratio imputation is derived and expressed as[4]

$$\bar{y}_{imp}^{adj}=\bar{y}_r+\hat{R}_r^{uw}(\bar{x}-\bar{x}_r), \tag{A1.5}$$

where $\bar{x}$ uses full sample x-values[5].

4 See also in Haziza and Rao (2003).

5 No imputations for x-values; all data of x-values are actual observations.

Appendix 2 Variance Estimation of Imputed Estimators

1. With weighted ratio imputation

Under weighted ratio imputation, $V_1(.)$ and $V_2(.)$ are derived and estimated by v_1^{wr} and v_2^{wr}, respectively as follows (Haziza and Rao 2003)

$$v_1^{wr}=v(q). \tag{A2.1}$$

In (A2.1), $q=\sum_s w_i q_i$, the value of q_i for $i\in s$ under weighted ratio imputation is given by $q_i=\frac{1}{\sum_s w_i}(q_{1i}-\bar{y}_{imp}^{wr})$, with $q_{1i}=a_i y_i+(1-a_i)\hat{R}_r x_i+\hat{c}a_i(y_i-\hat{R}_r x_i)$, where $\hat{c}=\frac{\sum_s w_i(1-a_i)x_i}{\sum_s w_i a_i x_i}$ and a_i is an indicator, $a_i=1$ if $i\in s_r$, otherwise $a_i=0$.

In addition, under weighted ratio imputation, v_2^{wr} is derived and expressed as

$$v_2^{wr}=\frac{\hat{p}(1-\hat{p})\left(\frac{\hat{x}}{\hat{x}_a}\right)^2 s_{er}^2}{\hat{N}}, \tag{A2.2}$$

where $\hat{p}=\frac{\sum_{s_r} w_i}{\sum_s w_i}$, $\hat{x}=\sum_s w_i x_i$, $\hat{x}_a=\sum_s w_i a_i x_i$, $\hat{N}=\sum_s w_i$, $s_{er}^2=\frac{\sum_s w_i a_i(y_i-\hat{R}_i x_i)^2}{\sum_s w_i a_i}$.

2. With unweighted ratio imputation

Similarly, for the case of unweighted ratio imputation, using linearization variance estimation with the delta method, $V_1(.)$ and $V_2(.)$ are estimated by v_1^{uwr} and v_2^{uwr}, respectively as follows.

$$v_1^{uwr}=v(q). \tag{A2.3}$$

In (A2.3), $q=\sum_s w_i q_i$, the value of q_i for $i\in s$ under unweighted ratio imputation is given by $q_i=\frac{1}{\sum_s w_i}(q_{2i}-\bar{y}_{imp}^{uwr})$, with $q_{2i}=a_i y_i+(1-a_i)\hat{R}_r^{uw}x_i+\hat{d}\left(\frac{a_i}{w_i}\right)(y_i-\hat{R}_r^{uw}x_i)$,

where $\hat{d}=\frac{\sum_s w_i(1-a_i)x_i}{\sum_s a_i x_i}$.

Under unweighted ratio imputation, v_2^{uwr} is derived and given as

$$v_2^{uwr}=\frac{\hat{p}(1-\hat{p})\left[s_{er(1)}^2+\left(\frac{\hat{x}-\hat{x}_a}{\hat{x}_{au}}\right)^2 s_{er(2)}^2+2\left(\frac{\hat{x}-\hat{x}_a}{\hat{x}_{au}}\right)^2 s_{er(3)}^2\right]}{\hat{N}}, \tag{A2.4}$$

where $\hat{x}_{au}=\sum_s a_i x_i$, $s_{er(1)}^2=\frac{\sum_s w_i a_i(y_i-\hat{R}_r^{uw}x_i)^2}{\sum_s w_i a_i}$, $s_{er(2)}^2=\frac{\sum_s w_i^{-1} a_i(y_i-\hat{R}_r^{uw}x_i)^2}{\sum_s w_i a_i}$,

$s_{er(3)}^2=\frac{\sum_s a_i(y_i-\hat{R}_r^{uw}x_i)^2}{\sum_s w_i a_i}$.

3. For bias-adjusted estimator under unweighted ratio imputation

For the bias-adjusted estimator under unweighted ratio imputation, using linearization variance estimation with the delta method, $V_1(.)$ and $V_2(.)$ are derived and estimated by v_1^{adj} and v_2^{adj} (Haziza and Rao 2003), respectively as follows.

$$v_1^{adj}=v(q). \tag{A2.5}$$

In (A2.5), $q=\sum_s w_i q_i$, the value of q_i for $i\in s$ under unweighted ratio imputation with bias-adjusted is given by $q_i=\frac{1}{\sum_s w_i}(q_{3i}-\bar{y}_{imp}^{adj})$, with

$$q_{3i}=\frac{a_i}{\sum_s w_i a_i}[(y_i-\bar{y}_r)+\hat{R}_r^{uw}(x_i-\bar{x}_r)]+\frac{\hat{R}_r^{uw}}{\hat{N}}(x_i-\bar{x})+\frac{1}{\sum_s a_i x_i}\left(\frac{a_i}{w_i}\right)(x_i-\bar{x}_r)(y_i-\hat{R}_r^{uw}x_i).$$

Under unweighted ratio imputation with bias-adjusted, v_2^{adj} is derived and presented as

$$v_2^{adj} = \hat{p}(1-\hat{p})\frac{\hat{N}}{\sum_s w_i a_i}[s_{yr}^2 + (\hat{R}_r^{uw})^2 s_{xr}^2 - 2\hat{R}_r^{uw} s_{yxr} + (\hat{h}\sum_s w_i a_i)^2 s_{er(2)}^2 + 2(\hat{h}\sum_s w_i a_i)(s_{eyr} - \hat{R}_r^{uw} s_{exr})], \quad \text{(A2.6)}$$

where $\hat{h} = \frac{(\bar{x} - \bar{x}_r)}{\sum_s a_i x_i}$, $s_{yr}^2 = \frac{\sum_s w_i a_i (y_i - \bar{y}_r)^2}{\sum_s w_i a_i}$, $s_{xr}^2 = \frac{\sum_s w_i a_i (x_i - \bar{x}_r)^2}{\sum_s w_i a_i}$,

$s_{yxr} = \frac{\sum_s w_i a_i (x_i - \bar{x}_r)(y_i - \bar{y}_r)}{\sum_s w_i a_i}$, $s_{er(2)}^2 = \frac{\sum_s w_i^{-1} a_i (y_i - \hat{R}_r^{uw} x_i)^2}{\sum_s w_i a_i}$,

$s_{eyr} = \frac{\sum_s a_i (y_i - \hat{R}_r^{uw} x_i)(y_i - \bar{y}_r)}{\sum_s w_i a_i}$, $s_{exr} = \frac{\sum_s a_i (y_i - \hat{R}_r^{uw} x_i)(x_i - \bar{x}_r)}{\sum_s w_i a_i}$.

Appendix 3 Simulation results

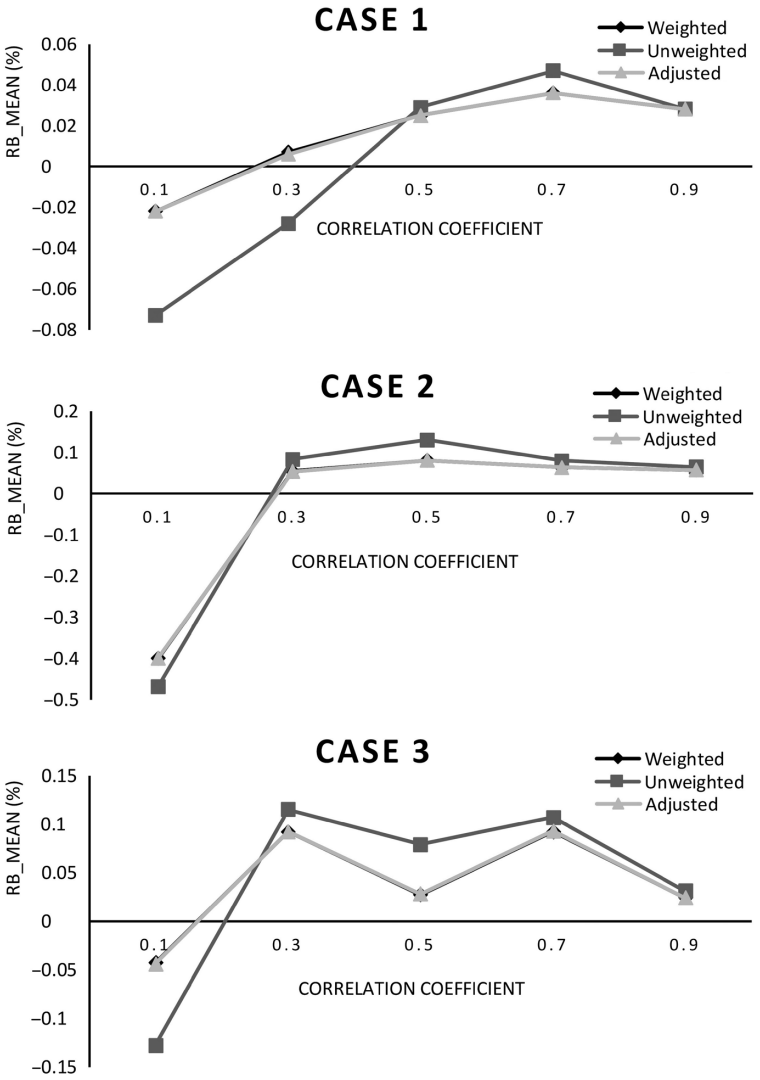

Figure A3.1 Relative bias of the imputed estimators for six cases

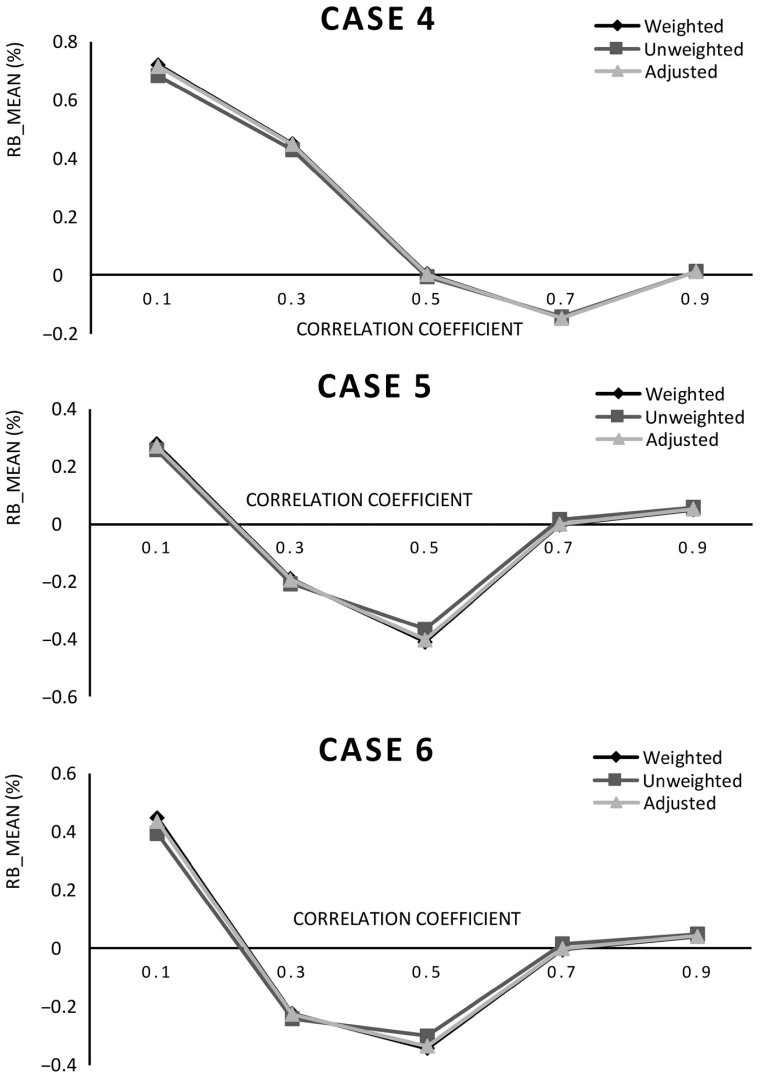

Figure A3.1 Relative bias of the imputed estimators for six cases (Continued)

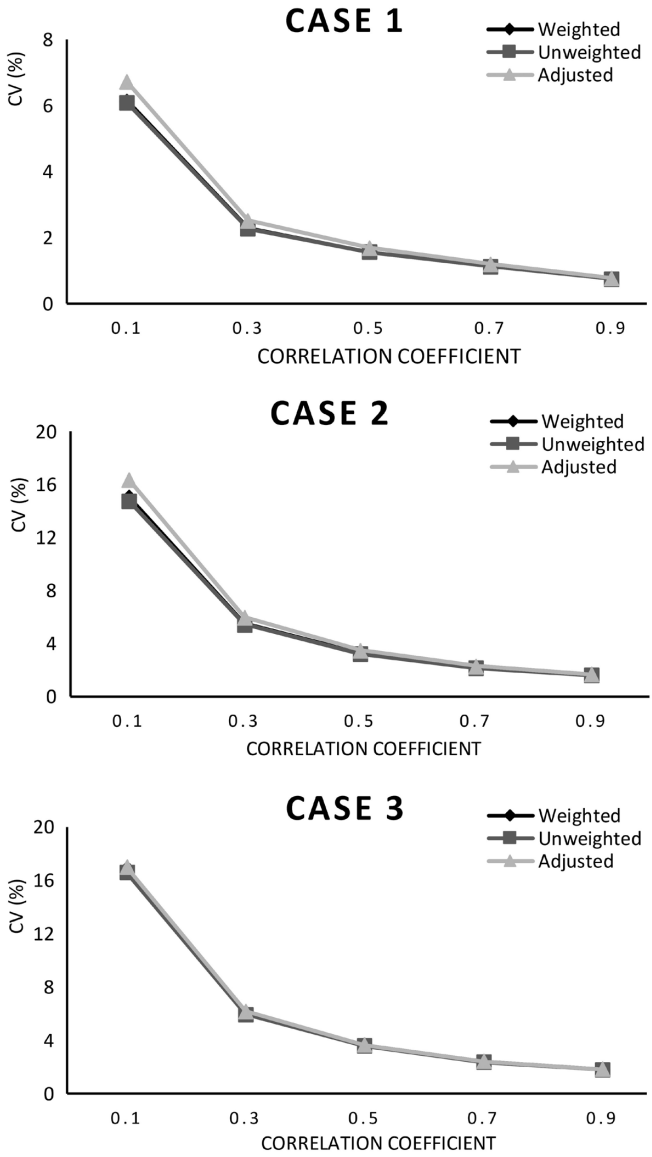

Figure A3.2 CV of the imputed estimators for six cases

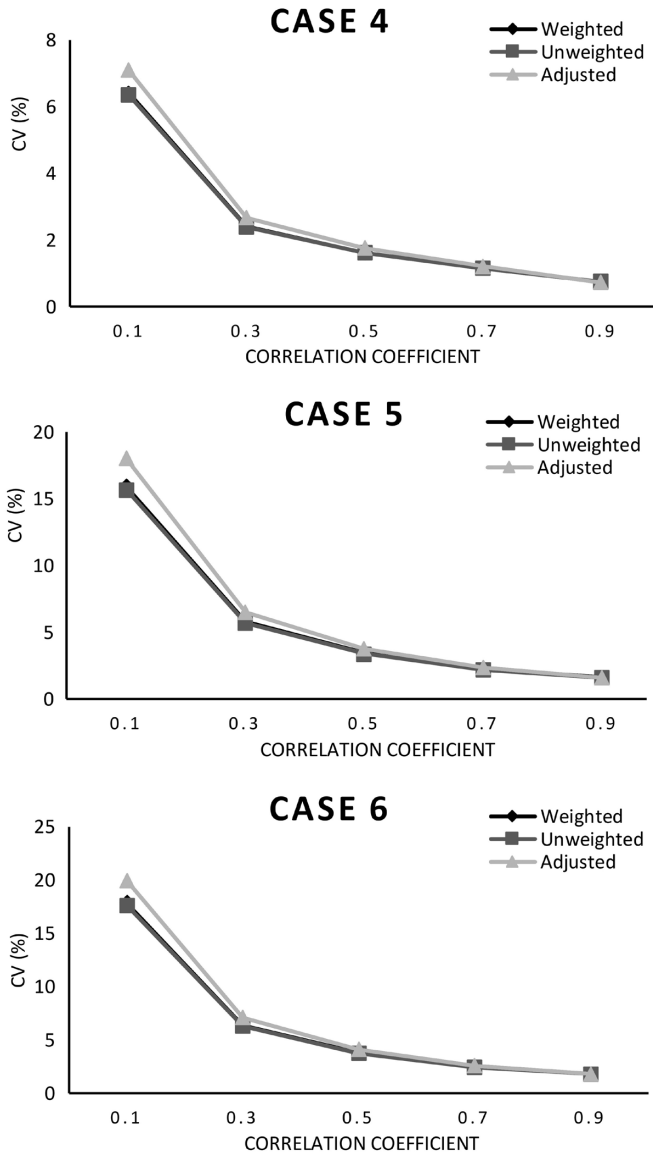

Figure A3.2 CV of the imputed estimators for six cases (Continued)

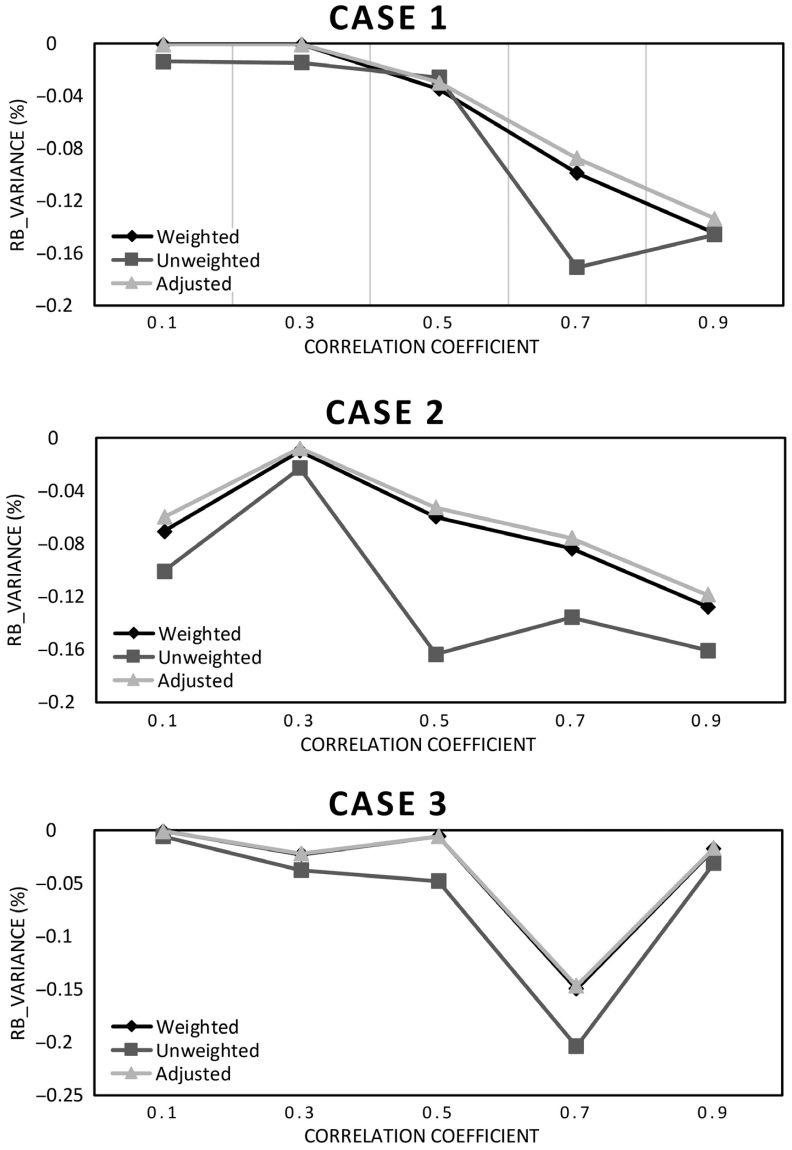

Figure A3.3 Relative bias of the variance estimators for six cases

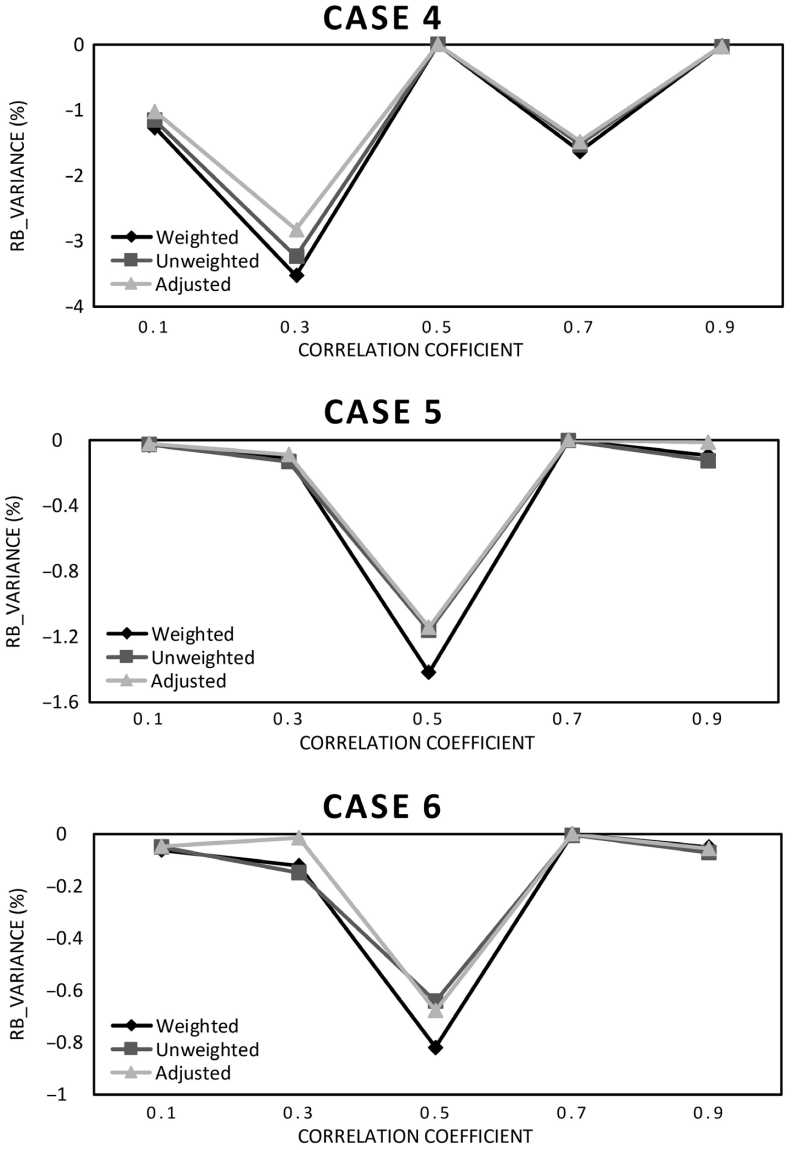

Figure A3.3 Relative bias of the variance estimators for six cases (Continued)

Appendix 4 Survey data for practical application

Table A4.1 Survey data obtained by the sampling survey in 2001

No.	Northern Region		Central Region		Southern Region		Eastern Region	
	Cost (NT$)	Quantity (Unit)	Cost (NT$)	Quantity (Unit)	Cost (NT$)	Quantity (Unit)	Cost (NT$)	Quantity (Unit)
1	142700	2700	280000	1200	70000	600	28000	1095
2	36000	400	450000	3500	50000	600	280000	10050
3	90000	700	110000	1000	35000	450	1032000	16200
4	10000	120	1100000	10000	15000	380	32500	200
5	7500	90	180000	1200	12000	250	22500	200
6	12500	160	70000	400	11000	200	28000	200
7	36000	350	18000	160	8000	230	130000	400
8	18000	170	48000	400	30000	600	55000	290
9	38000	1500	25000	200	100000	700	96000	400
10	190000	1100	—	160	9000	100	—	700
11	37500	350			1800	120		
12	16000	200			38000	165		
13	26000	200			18000	160		
14	16800	200			19500	160		
15	36000	350			11200	140		
16	30000	150			18800	160		
17	36000	400			240000	1800		
18	18900	200			14000	200		
19	17500	200			37500	210		
20	45000	400			90000	850		
21	40000	200			12000	520		
22	20000	120			8500	240		
23	16250	160			36000	650		
24	29000	150			9500	120		
25	45000	400			32000	400		
26	33000	350			4500	1000		
27	112000	1100			22500	475		
28	10400	120			9000	175		
29	30000	160			52000	620		
30	24000	160						

Source: Sampling survey data from Hsu et al. (2001).

REFERENCES

Al-Jararha, Jehad M., and Mazen Sulaiman, 2020, "Horvitz-Thompson Estimator Based on the Auxiliary Variable." *Statistics in Transition New Series* 21(1): 37–53.

Chen, Sixia, and David Haziza, 2019, "Recent Developments in Dealing with Item Non-response in Surveys: A Critical Review." *International Statistical Review* 87(S1): S192–S218.

Cochran, William G., 1977, *Sampling Techniques* (3rd ed.). New York: John Wiley & Sons.

Fay, Robert E., 1991, "A Design-Based Perspective on Missing Data Variance." Pp. 429–440 in *Proceedings of the 1991 Annual Research Conference*, edited by Bureau of the Census. Washington, DC: U.S. Bureau of the Census.

Haziza, David, and Jon N. K. Rao, 2003, "Inference for Population Means under Unweighted Imputation for Missing Survey Data." *Survey Methodology* 29: 81–90.

——, 2005, "Inference for Domains under Imputation for Missing Survey Data." *The Canadian Journal of Statistics* 33(2): 149–161.

Hedayat, Samad, and Bikas K. Sinha, 1991, *Design and Inference in Finite Population Sampling*. New York: John Wiley & Sons.

Horvitz, Daniel G., and Donovan J. Thompson, 1952, "A Generalization of Sampling without Replacement from A Finite Universe." *Journal of the American Statistical Association* 47(260): 663–685.

Hsu, Esher, Chien-Fu J. Lin, Chen-Meng Kuo, Nan-Min Wu, Hsiao-Kan Ma, Yunchang J. Bor, and Yu-Lan Chien, 2001, "Formulating Recycling Charges and Subsidies of Waste Home Appliances." *Report*, No. EPA–90–HA31–03–060. Taipei: Environmental Protection Administration.

Keeble, Claire, Graham R. Law, Stuart Barber, and Paul D. Baxter, 2015, "Choosing a Method to Reduce Selection Bias: A Tool for Researchers." *Open Journal of Epidemiology* 5: 155–162.

Knaub, James R., 2017, "Comparison of Model-Based to Design-Based Ratio Estimators." Paper presented at The 2017 JSM, Baltimore, Maryland, USA, August 1.

Little, Roderick J. A., 1992, "Regression with Missing X's: A Review." *Journal of the American Statistical Association* 87(420):1227–1237.

Little, Roderick J. A., and Donald B. Rubin, 1987, *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.

Namboodiri, N. Krishnan, 1978, *Survey Sampling and Measurement*. New York: Academic Press, Inc.

Nittner, Thomas, 2003, "Missing at Random (MAR) in Nonparametric Regression—A Simulation Experiment." *Statistical Methods & Applications* 12: 195–210.

Rao, Jon N. K., 1966, "Alternative Estimators in PPS Sampling for Multiple Characteristics." *Sankhyā: The Indian Journal of Statistics, Series A (1961–2002)* 28(1): 47–60.

Rao, Jon N. K., and Jun Shao, 1992, "Jackknife Variance Estimation with Survey Data under Hot Deck Imputation." *Biometrika* 79: 811–822.

Sampford, Michael R., 1967, "On Sampling without Replacement with Unequal Probabilities of Selection." *Biometrika* 54: 499–513.

Särndal, Carl-Erik, 1978, "Design-Based and Model-Based Inference in Survey Sampling." *Scandinavian Journal of Statistics* 5(1): 27–52.

——, 1992, "Methods for Estimating the Precision of Survey Estimates When Imputation Has Been Used." *Statistics Canada* 18: 241–252.

Shao, Jun, and Philip Steel, 1999, "Variance Estimation for Survey Data with Composite Imputation and Nonnegligible Sampling Fractions." *Journal of the American Statistical Association* 94 (445): 254–265.

Skinner, Chris J., and Jon N. K. Rao, 2002, "Jackknife Variance Estimation for Multivariate Statistics Under Hot-deck Imputation from Common Donors." *Journal of Statistical Planning and Inference* 102: 149–167.

Wheeler, David C., Jason E. VanHorn, and Electra D. Paskett, 2007, "A Comparison of Design-based and Model-based Analysis of Sample Surveys in Geography." *Technical Report* No. 07–11 December 2007. Georgia: Department of Biostatistics Rollins School of Public Health Emory University.