Journal of Statistical Sciences & Data Analytics (JOSSDA)
Volume 1, Issue 2, December 2026 | Print ISSN XXXX-XXXX | Online ISSN XXXX-XXXX

DOI: 10.XXXXX/jossda.2026.1.1.011

JOURNAL OF STATISTICAL SCIENCES & DATA ANALYTICS
Department of Statistics, Ahmadu Bello University, Zaria, Nigeria | www.jossda.abu.edu.ng www.jossda.com.ng

Generalized Regression-Based Imputation Using Multiple Auxiliary Variables for Missing Data in Survey Sampling

U. A. Huguma[1*], O. O. Ishaq[1], S. A. Sabo[1], H. A. Hamisu[1], I. A. Isah[1], U. A. Abubakar[1], N. Yusuf[1], S. S. Suleiman[1], J. S. Ogheneovo[1], F. U. Muhammad[1], A. F. Kabara[1]
[1]*Department of Statistics, Aliko Dangote University of Science and Technology, Wudil, Nigeria*

Corresponding Author:
U. A. Huguma, Department of Statistics, Aliko Dangote University of Science and Technology, Wudil, Nigeria. umarhuguma2023@gmail.com

Abstract

Non-response remains a major source of bias and efficiency loss in sample surveys, particularly when conventional imputation procedures use only one auxiliary variable or assume that missingness is completely random. This paper develops a generalised regression (GREG)-based imputation estimator for the finite-population mean when the study variable is subject to missingness and several auxiliary variables are completely observed. Under a missing-at-random mechanism, the respondent data are used to estimate a multiple linear regression model, predictions are generated for nonrespondents, and the completed-data estimator is expressed in the familiar model-assisted GREG form. The estimator is shown to be approximately unbiased to the first order, while its mean squared error is governed by the residual variation after projecting the study variable on the auxiliary vector. Four finite-population data sets are used for empirical evaluation against established imputation estimators. The proposed estimator recorded percentage relative efficiencies of 443.26% for Population I, 184.61% for Populations II and III, and approximately 184.81% across the complete response-rate scenarios for Population IV. These values correspond to substantial reductions in mean squared error relative to the respondent-mean baseline. The results further show that absolute mean squared error declines as the number of respondents increases, whereas the relative advantage of the proposed estimator remains stable. The findings support multiple-auxiliary GREG imputation as an efficient and implementable approach for finite-population estimation under non-response, provided that the working regression model and missing-at-random assumption are reasonable.

Keywords: auxiliary information; finite population mean; generalised regression estimator; imputation; missing at random; non-response; percentage relative efficiency

1. Introduction

Sample surveys provide essential information for public policy, economic planning, health research, social investigation and environmental monitoring. Their validity, however, depends on the quality and completeness of the observations collected from sampled units. Non-response occurs when selected units fail to provide all or part of the required information. Unit non-response removes the complete record for a sampled unit, whereas item non-response leaves one or more study variables unobserved. In both cases, the effective sample size is reduced, and the resulting estimators may be biased when respondents and nonrespondents differ systematically (Groves *et al.*, 2009; Särndal & Lundström, 2005). Imputation replaces a missing observation with a plausible value so that a complete analytic data set can be retained. Classical procedures include respondent-mean, ratio and regression imputation. These methods are attractive because of their simplicity, but many of them use only one auxiliary variable and therefore fail to exploit the wider information now available in modern surveys. Demographic, socioeconomic, administrative and geographic variables may jointly predict the study variable and the propensity to respond. Discarding this multivariate information can increase residual variation and weaken finite-population inference. Another important limitation concerns the assumed missing-data mechanism. Much of the classical theory is presented under missing completely at random (MCAR), for which response is independent of observed and unobserved measurements. In practice, non-response frequently depends on observed characteristics. Missing at random (MAR), under which the probability of response may depend on observed auxiliary variables but not on the missing study value after conditioning on those variables, is therefore a more credible working assumption (Rubin, 1976; Heitjan & Basu, 1996; Little & Rubin, 2019). The generalised regression estimator is a well-established model-assisted procedure that combines probability-sample observations with auxiliary totals or means to improve precision (Särndal *et al.*, 1992). Although multiple regression is widely used for prediction, a fully articulated imputation scheme that embeds multiple-auxiliary prediction in the GREG framework and establishes its properties under non-response is less common in applied survey work. This paper addresses that gap by developing a multiple-auxiliary GREG imputation estimator for the finite-population mean, deriving its approximate bias and mean squared error (MSE), and evaluating its efficiency with four empirical populations. The specific objectives are to: (i) formulate a GREG-based imputation scheme and resultant point estimator; (ii) derive its first-order bias and MSE; (iii) compare its empirical efficiency with existing imputation estimators; and (iv) examine its behaviour across different sample and respondent sizes. The contribution is both methodological and practical: it provides a transparent estimator that uses several auxiliary variables simultaneously and can be implemented with standard regression and survey-estimation software.

Respondent-mean imputation replaces every missing study value with the mean calculated from responding units. It is easy to apply, but it suppresses variation and ignores auxiliary information. Ratio imputation improves on this procedure when a study variable is approximately proportional to a positive auxiliary variable. Its performance deteriorates when the relationship does not pass through the origin or when the chosen auxiliary variable is weak. Conventional regression imputation accommodates an intercept and is more flexible, although its common single-predictor formulation still ignores the joint predictive contribution of other variables (Lee *et al*., 1994). Several refinements have been proposed for one-auxiliary-variable settings. Singh and Horn (2000) introduced compromised imputation, combining information from responding and nonresponding units. Singh and Deo (2003) used a power transformation, while Kadilar and Cingi (2008) developed modified ratio-type procedures in the presence of missing observations. Exponential and ratio-exponential variants were later developed to reduce MSE under particular combinations of coefficients of variation, correlations and tuning parameters. Singh *et al*. (2022) proposed another improved alternative for estimating a population mean with missing survey values. These methods can be efficient under their defining conditions, but their dependence on a single auxiliary variable and optimally selected constants may limit their generality. Multiple auxiliary variables can be used sequentially, but separate one-variable imputations do not adequately represent correlation among predictors. Multiple regression imputation treats the auxiliary variables jointly, yet a purely model-based implementation may not preserve the design-based or model-assisted rationale needed for finite-population inference. GREG estimation closes this gap by proposing an estimator to known multi-auxiliary information while retaining a regression representation. The present method combines these two roles: regression predictions provide values for missing observations, and calibration corrects the respondent estimate using the discrepancy between respondent and known auxiliary means. The missingness mechanism determines whether the procedure is defensible. Under MAR, conditioning on the observed auxiliary vector makes response ignorable for likelihood- or model-based estimation, provided that the response and outcome models are adequately specified (Little & Rubin, 2019). The proposed GREG framework is therefore especially relevant when the variables related to response are available for both respondents and nonrespondents.

2. Materials and Methods

2.1. Population, sample and notation

Let $U = \{1,2,\dots,N\}$ denote a finite population of size N. A simple random sample without replacement, s, of size n is selected. The study variable is y_i, and $\mathbf{x}_i = (x_{i1},\dots,x_{ip})'$ is a p-dimensional auxiliary vector observed for every sampled unit. Let $R_i = 1$ when unit i responds to the study variable and $R_i = 0$ otherwise. The respondent set is $s_r = \{i \in s: R_i = 1\}$ with size r, and the nonrespondent set is $s_m = s \setminus s_r$ with size $m = n - r$.

The parameter of interest is the finite-population mean

$$\bar{Y} = \frac{1}{N}\sum_{i\in U} y_i.. \tag{1}$$

Journal of Statistical Sciences & Data Analytics (JOSSDA)
Volume 1, Issue 2, December 2026 | Print ISSN XXXX-XXXX | Online ISSN XXXX-XXXX

DOI: 10.XXXXX/jossda.2026.1.1.011

The auxiliary population mean $\bar{X} = N^{-1}\sum_{i \in U} x_i$ is assumed known from a frame, census or administrative source. The respondent and complete-sample auxiliary means are respectively

$$\bar{X}_r = r^{-1}\sum_{i \in U} x_i \tag{2}$$

and

$$\bar{X}_s = n^{-1}\sum_{i \in S} x_i \tag{3}$$

2.2. Working model and response assumption

For respondents, consider the linear working model.

$$y_i = \beta_0 + \mathbf{x}_i{}'\boldsymbol{\beta} + e_i, \qquad E(e_i \mid \mathbf{x}_i) = 0, \qquad \mathrm{Var}(e_i \mid \mathbf{x}_i) = \sigma_i^2 \tag{4}$$

The missingness mechanism is assumed to be MAR:

$$P(R_i = 1 \mid y_i, \mathbf{x}_i) = P(R_i = 1 \mid \mathbf{x}_i). \tag{5}$$

Thus, after conditioning on the complete auxiliary vector, the response does not depend on the missing study value. The respondent sample must be sufficiently large to estimate the regression coefficients, and the auxiliary cross-product matrix must be nonsingular.

2.3. Proposed imputation scheme

Define $\mathbf{z}_i = (1, \mathbf{x}_i{}')'$ and $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\beta}')'$. The ordinary least-squares estimator calculated from respondents is

$$\hat{\boldsymbol{\theta}} = (\mathbf{Z}_r{}'\mathbf{Z}_r)^{-1}\mathbf{Z}_r{}'\mathbf{y}_r.$$

After centering the auxiliary variables, the slope estimator becomes

$$\hat{\boldsymbol{\beta}} = \left[\sum_{i \in s_r} (\mathbf{x}_i - \bar{\mathbf{x}}_r)(\mathbf{x}_i - \bar{\mathbf{x}}_r)'\right]^{-1} \sum_{i \in s_r} (\mathbf{x}_i - \bar{\mathbf{x}}_r)(y_i - \bar{y}_r) \tag{6}$$

Because $\hat{\beta}_0 = \bar{y}_r - \bar{\mathbf{x}}_r{}'\hat{\boldsymbol{\beta}}$, a missing value for $i \in s_m$ is imputed by

$$\hat{y}_i = \hat{\beta}_0 + \mathbf{x}_i{}'\hat{\boldsymbol{\beta}} = \bar{y}_r + (\mathbf{x}_i - \bar{\mathbf{x}}_r)'\hat{\boldsymbol{\beta}}. \tag{7}$$

The completed value is

$$y_i^* = R_i y_i + (1 - R_i)\hat{y}_i.$$

The completed-data mean is consequently.

$$\hat{\bar{Y}}_I = \frac{1}{n}\left(\sum_{i \in s_r} y_i + \sum_{i \in s_m} \hat{y}_i\right) \tag{8}$$

Substitution of (6) into (8) gives

$$\begin{aligned}\hat{\bar{Y}}_I &= \frac{1}{n}\left[r\bar{y}_r + (n-r)\bar{y}_r + \textstyle\sum_{i\in s_m}(\mathbf{x}_i - \bar{\mathbf{x}}_r)'\hat{\boldsymbol{\beta}}\right] \\ &= \bar{y}_r + (\bar{\mathbf{x}}_s - \bar{\mathbf{x}}_r)'\hat{\boldsymbol{\beta}}.\end{aligned} \tag{9}$$

When $\bar{\mathbf{X}}$ is known, calibration replaces $\bar{\mathbf{x}}_s$ with $\bar{\mathbf{X}}$, producing the proposed GREG imputation estimator

$$\boxed{\hat{\bar{Y}}_{\text{GREG}} = \bar{y}_r + (\bar{\mathbf{X}} - \bar{\mathbf{x}}_r)'\hat{\boldsymbol{\beta}}.} \tag{10}$$

Equation (10) shows that the estimator adds a regression adjustment to the respondent mean. The size and direction of this adjustment depend on the difference between the known population auxiliary means and their respondent counterparts.

2.4. Approximate bias

Write the finite-population regression decomposition as

$$y_i = \bar{Y} + (\mathbf{x}_i - \bar{\mathbf{X}})'\boldsymbol{\beta} + e_i, \qquad \textstyle\sum_{i\in U} e_i = 0. \tag{11}$$

Substitution into (10) gives

$$\begin{aligned}\hat{\bar{Y}}_{\text{GREG}} - \bar{Y} &= (\bar{y}_r - \bar{Y}) + (\bar{\mathbf{X}} - \bar{\mathbf{x}}_r)'\hat{\boldsymbol{\beta}} \\ &= \bar{e}_r + (\bar{\mathbf{x}}_r - \bar{\mathbf{X}})'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}).\end{aligned} \tag{12}$$

Under standard regularity conditions, $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = O_p(r^{-1/2})$, and the second term in the final expression is of second order. Therefore,

$$\hat{\bar{Y}}_{\text{GREG}} - \bar{Y} \doteq \bar{e}_r, \qquad \text{Bias}\left(\hat{\bar{Y}}_{\text{GREG}}\right) \doteq 0. \tag{13}$$

The proposed estimator is thus approximately unbiased to the first order when the response process is ignorable after conditioning on the auxiliary variables and the working model is adequate.

2.5. First-order mean squared error

From (13),

$$\text{MSE}\left(\hat{\bar{Y}}_{\text{GREG}}\right) \doteq \text{Var}(\bar{e}_r) = \left(\frac{1}{r} - \frac{1}{N}\right)S_e^2,$$

where

$$S_e^2 = S_y^2 + \boldsymbol{\beta}'\mathbf{S}_x\boldsymbol{\beta} - 2\boldsymbol{\beta}'\mathbf{S}_{xy}.$$

Journal of Statistical Sciences & Data Analytics (JOSSDA)
Volume 1, Issue 2, December 2026 | Print ISSN XXXX-XXXX | Online ISSN XXXX-XXXX

DOI: 10.XXXXX/jossda.2026.1.1.011

At the least-squares optimum $\boldsymbol{\beta} = \mathbf{S}_x^{-1}\mathbf{S}_{xy}$,

$$S_e^2 = S_y^2 - \mathbf{S}_{yx}\mathbf{S}_x^{-1}\mathbf{S}_{xy} = S_y^2(1 - R^2),$$

where R^2 is the finite-population coefficient of multiple determination. Accordingly,

$$\mathrm{MSE}\left(\hat{\bar{Y}}_{\mathrm{GREG}}\right) \doteq \left(\frac{1}{r} - \frac{1}{N}\right) S_y^2(1 - R^2).$$

For respondent-mean imputation,

$$\mathrm{MSE}(\bar{y}_r) \doteq \left(\frac{1}{r} - \frac{1}{N}\right) S_y^2.$$

Therefore,

$$\mathrm{MSE}(\bar{y}_r) - \mathrm{MSE}\left(\hat{\bar{Y}}_{\mathrm{GREG}}\right) \doteq \left(\frac{1}{r} - \frac{1}{N}\right) S_y^2 R^2 \geq 0.$$

Strict improvement occurs whenever the auxiliary vector explains a nonzero proportion of variation in y. Percentage relative efficiency is defined as

$$\mathrm{PRE}(T_j) = \frac{\mathrm{MSE}(T_0)}{\mathrm{MSE}(T_j)} \times 100,$$

where T_0 is respondent-mean imputation. A value above 100% indicates improvement over the baseline.

2.6. Empirical evaluation

Four finite populations were used to compare the proposed estimator with eleven existing or modified estimators presented in the source study. Because the original estimator symbols were embedded equation objects, the competitors are indexed as $T_0, T_1, \ldots, T_{10}$ in their original table order; T_0 denotes the respondent-mean baseline and T_{GREG} denotes the proposed estimator. MSE and PRE were computed for Populations I–III. For Population IV, performance was evaluated at several combinations of sample size n and respondent count r.

3. Results

3.1. Populations I–III

Table 1 reports the complete MSE and PRE results for Population I. The proposed GREG estimator attained an MSE of 12,028.93 and PRE of 443.26%. Its MSE was the smallest reported value, although the same value was shared by T_2, T_3 and T_7. Relative to the baseline, the result implies an MSE reduction of approximately 77.44%.

Journal of Statistical Sciences & Data Analytics (JOSSDA)
Volume 1, Issue 2, December 2026 | Print ISSN XXXX-XXXX | Online ISSN XXXX-XXXX

DOI: 10.XXXXX/jossda.2026.1.1.011

Table 1. MSE and PRE of estimators for Population I

Estimator	MSE	PRE (%)	Estimator	MSE	PRE (%)
T_0	53,319.74	100.0000	T_6	33,907.86	157.2489
T_1	15,085.49	353.4505	T_7	12,028.93	443.2624
T_2	12,028.93	443.2624	T_8	14,029.59	380.0520
T_3	12,028.93	443.2624	T_9	39,055.64	136.5225
T_4	33,907.86	157.2489	T_{10}	14,029.59	380.0520
T_5	46,613.82	114.3861	T_{GREG}	12,028.93	443.2624

For Population II, the proposed estimator recorded an MSE of 156,356,484 and PRE of 184.61% (Table 2), corresponding to an MSE reduction of approximately 45.83% relative to T_0. The proposed estimator shared the smallest internally consistent MSE with T_2, T_3 and T_7. The original T_{10} entry (MSE = 14,029.59; PRE = 2,057,479%) is inconsistent with the scale and with the defining PRE formula and is therefore flagged for verification rather than interpreted as a valid efficiency result.

Table 2. MSE and PRE of estimators for Population II

Estimator	MSE	PRE (%)	Estimator	MSE	PRE (%)
T_0	288,655,816	100.0000	T_6	405,176,555	71.2420
T_1	174,605,502	165.3189	T_7	156,356,484	184.6139
T_2	156,356,484	184.6139	T_8	171,330,037	168.4794
T_3	156,356,484	184.6139	T_9	264,909,782	108.9638
T_4	405,176,555	71.2420	T_{10}	14,029.59	2,057,479
T_5	309,569,795	93.2442	T_{GREG}	156,356,484	184.6139

Population III produced a similar ranking (Table 3). The proposed estimator had an MSE of 709,280.8 and PRE of 184.61%, again implying an MSE reduction of approximately 45.83%. The T_{10} source entry is inconsistent with the scale of the population and is not used in the substantive comparison.

Table 3. MSE and PRE of estimators for Population III

Estimator	MSE	PRE (%)	Estimator	MSE	PRE (%)
T_0	1,309,431	100.0000	T_6	1,159,630	112.9180
T_1	720,016.5	181.8612	T_7	709,280.8	184.6139
T_2	709,280.8	184.6139	T_8	718,892.4	182.1456
T_3	709,280.8	184.6139	T_9	1,246,592	105.0409
T_4	1,159,630	112.9180	T_{10}	14,029.59	9,333.351
T_5	1,293,746	101.2124	T_{GREG}	709,280.8	184.6139

3.2. Effect of response level in Population IV

Population IV examined the estimators under several sample and respondent sizes. Table 4 presents the complete, internally aligned rows for $n = 61$ and $n = 611$. The absolute MSE of every estimator decreased as r increased. At $n = 61$, the MSE of the proposed estimator fell from 16,281,084,146 when $r = 6$ to 3,230,619,317 when $r = 30$. At $n = 611$,

Journal of Statistical Sciences & Data Analytics (JOSSDA)
Volume 1, Issue 2, December 2026 | Print ISSN XXXX-XXXX | Online ISSN XXXX-XXXX
 DOI: 10.XXXXX/jossda.2026.1.1.011

it declined from 1,572,568,458 when $r = 61$ to 288,916,180 when $r = 305$. Despite these reductions in absolute error, the proposed estimator's PRE remained approximately 184.81% in every complete row.

Table 4. Proposed GREG performance across response scenarios in Population IV

Sample size (n)	Respondents (r)	Response rate (%)	MSE of T_0	MSE of T_{GREG}	PRE (%)
61	6	9.84	30,089,362,284	16,281,084,146	184.8118
61	12	19.67	15,015,114,130	8,124,543,628	184.8118
61	18	29.51	9,990,364,746	5,405,696,789	184.8118
61	24	39.34	7,477,990,053	4,046,273,369	184.8118
61	30	49.18	5,970,565,238	3,230,619,317	184.8118
611	61	9.98	2,906,291,843	1,572,568,458	184.8118
611	122	19.97	1,423,578,910	770,285,784	184.8118
611	183	29.95	929,341,265	502,858,226	184.8118
611	244	39.93	682,222,443	369,144,447	184.8118
611	305	49.92	533,951,150	288,916,180	184.8118

The stable PRE indicates that the baseline and proposed MSEs change proportionally with the response level in the reported formulation. The result is consistent with Equation (20): increasing r reduces the finite-population variance factor, while the multiple coefficient of determination controls the relative reduction. Rows with merged or misaligned cells in the original Population IV table were not reconstructed beyond values that could be identified unambiguously.

3.3 Overall empirical performance

Table 5 summarises the performance of the proposed estimator. It shared the smallest verified MSE in all four populations. Its greatest relative gain occurred in Population I, where PRE exceeded 443%. Populations II–IV produced comparable PRE values of about 185%.

Table 5. Cross-population performance of the proposed estimator

Population	MSE of T_{GREG}	PRE (%)	Reduction relative to T_0 (%)	Verified standing
I	12,028.93	443.2624	77.44	Joint minimum
II	156,356,484	184.6139	45.83	Joint minimum
III	709,280.8	184.6139	45.83	Joint minimum
IV	Varies with n, r	184.8118	45.89	Joint minimum in complete rows

3.4. Discussion

The empirical findings support the theoretical argument for using multiple auxiliary variables in imputation. From Equation (20), the MSE of the GREG estimator is proportional to $S_y^2(1 - R^2)$. When the auxiliary vector explains a substantial proportion of variation in the study variable, the residual variance is smaller than the original study-variable variance, and a sizeable efficiency gain follows. Population I appears to contain particularly strong predictive information, producing a PRE of 443.26% and reducing MSE by approximately 77.44% relative to respondent-mean imputation. The PRE of about 185% in Populations II–IV still represents an important improvement: the proposed estimator used only about 54% of the baseline MSE. Population IV demonstrates the expected effect of response level. The absolute MSE fell consistently as more sampled units responded because the variance factor involving r decreased. The near-constant PRE suggests that the relative amount of variation explained by the auxiliary variables remained unchanged over the reported response scenarios. In practice, this means that GREG adjustment can improve precision at both low and moderate response rates, but it cannot fully substitute for effective response follow-up: both the baseline and adjusted estimators become more stable when r increases.

Several competitors tied the proposed estimator in the reported data. These exact ties may reflect algebraically equivalent minimum-MSE forms, identical use of the available population moments, or repeated formulas in the computational implementation. Consequently, the evidence supports the proposed method as jointly best rather than uniquely dominant. The results are consistent with the broader literature. Ratio and power-transformation estimators can perform well when their structural assumptions are satisfied (Singh & Horn, 2000; Singh & Deo, 2003), whereas regression and calibration procedures are more flexible when several correlated predictors are available (Särndal et al., 1992). Under MAR, including variables related to both outcome and response improves the plausibility of ignorability and reduces residual bias. Nevertheless, MAR cannot be verified from observed data alone. Sensitivity analysis remains necessary when nonresponse may depend on unobserved study values.

4. Conclusion

This study developed a multiple-auxiliary GREG-based imputation estimator for estimating a finite-population mean when the study variable is missing for some sampled units. The estimator combines respondent regression prediction with known auxiliary means. Under standard regularity conditions and an ignorable MAR response process, it is approximately unbiased to the first order and has MSE proportional to the residual variance $S_y^2(1 - R^2)$.

Across the four empirical populations, the proposed estimator recorded PRE values ranging from 184.61% to 443.26% and shared the smallest verified MSE. Increasing the number of respondents reduced absolute MSE, while the relative efficiency of the proposed estimator remained stable in Population IV. The results demonstrate that complete and predictive auxiliary information can materially improve finite-population estimation under non-response.

The following recommendations arise from the study. First, survey organisations should collect and maintain auxiliary variables that are predictive of both the outcome and response. Second, analysts should assess model adequacy and multicollinearity before applying GREG imputation. Third, future work should evaluate empirical bias, confidence-interval coverage and Monte Carlo error under both MCAR and MAR using a fully reported simulation design. Finally, the method should be extended to stratified, cluster and unequal-probability sampling and compared with nonlinear and machine-learning-assisted imputation when the linear working model is inadequate.

5. Practical Implementation

The proposed procedure can be implemented through the following steps:

i. identify auxiliary variables observed for respondents and nonrespondents and, preferably, with known population totals or means;
ii. examine missingness patterns, response rates and relationships between response indicators and auxiliary variables;
iii. fit the multiple linear regression model using respondents, checking linearity, influential observations, heteroscedasticity and multicollinearity;
iv. calculate predictions for nonrespondents and form the completed data values;
v. apply the GREG correction in Equation (10) or equivalent calibration weights in Equation (13);
vi. estimate uncertainty using a design-consistent linearization, replication or bootstrap procedure that repeats coefficient estimation and imputation within each replicate; and
vii. compare the result with simpler benchmark procedures and perform sensitivity analyses for model specification and departures from MAR.

This workflow is readily programmable in R, Python, SAS or other statistical software. When the sampling design is complex, the design weights, strata and clusters should be incorporated in both estimation and variance calculation.

Data Availability

The empirical populations and computational outputs used in this study are contained in the underlying dissertation materials. Additional reproducibility files can be deposited on request.

Conflict of Interest

The author declares no conflict of interest. Ethical approval: The study is methodological; no personally identifiable human-subject data are reported in this manuscript.

Acknowledgements

The author acknowledges Dr O. O. Ishaq and S. A. Sabo of the Department of Statistics, Aliko Dangote University of Science and Technology, Wudil, for their academic support.

References

Carpenter, J. R., & Kenward, M. G. (2013). *Multiple imputation and its application*. Wiley.
Chen, Q., & Kott, P. S. (2020). Model-assisted imputation for complex surveys. *Journal of Official Statistics*, *36*(4), 757–778.
Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (2nd ed.). Wiley.
Haziza, D., & Beaumont, J.-F. (2017). *Imputation in surveys: From theory to practice*. Statistics Canada.
Heitjan, D. F., & Basu, S. (1996). Distinguishing “missing at random” and “missing completely at random.” *The American Statistician*, *50*(3), 207–213.
Kalton, G., & Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, *12*(1), 1–16.

Journal of Statistical Sciences & Data Analytics (JOSSDA)
Volume 1, Issue 2, December 2026 | Print ISSN XXXX-XXXX | Online ISSN XXXX-XXXX
 DOI: 10.XXXXX/jossda.2026.1.1.011

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

Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). Wiley.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Särndal, C.-E., & Lundström, S. (2005). *Estimation in surveys with nonresponse*. Wiley.

Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. Springer.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177.

Singh, G. N., Jaiswal, A. K., Singh, C., & Usman, M. (2022). An improved alternative method of imputation for missing data in survey sampling. *Journal of Statistics Applications & Probability*, *11*(2), 535–543.

Singh, S., & Deo, B. (2003). Imputation by power transformation. *Statistical Papers*, *44*(4), 555–579.

Singh, S., & Horn, S. (2000). Compromised imputation in survey sampling. *Metrika*, *51*(3), 267–276.