Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Multiple imputation for nonresponse in surveys using design weights and auxiliary margins

by Kewei Xu and Jerome P. Reiter

Release date: June 29, 2026

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, the Agency has developed standards of service which its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Survey Methodology, June 2026
Vol. 52, No. 1, pp. 259-270
Statistics Canada, Catalogue No. 12-001-X

Multiple imputation for nonresponse in surveys using design weights and auxiliary margins

Kewei Xu and Jerome P. Reiter[1]

Abstract

Survey data typically have missing values due to unit and item nonresponse. Sometimes, survey organizations know the marginal distributions of certain categorical variables in the target population. As shown in previous work, survey organizations can leverage these distributions in multiple imputation for nonignorable unit nonresponse, generating imputations that result in plausible completed-data estimates for the variables with known margins. However, this prior work does not use the design weights for unit nonrespondents. We extend this previous work to utilize the design weights for all sampled units. We illustrate the approach using simulation studies.

Key Words: Item; Missing; Nonignorable; Unit.

1. Introduction

Survey data usually suffer from both unit and item nonresponse. As a result, survey organizations have to make strong assumptions about the reasons for missingness, for example, the data are missing at random. One way to lessen reliance on assumptions is to utilize information in auxiliary data sources. For example, and pertinent to the setting of our work, survey organizations may know the population percentages or totals of some categorical variables in the survey. These could be available from recent censuses, administrative databases, or other high quality surveys (Sadinle and Reiter, 2017). Indeed, such information is frequently used in methods for handling survey nonresponse, such as calibration and raking.

We consider settings where analysts seek to integrate auxiliary information on marginal distributions in multiple imputation for nonresponse (Rubin, 1987) in surveys. We build on the missing data with auxiliary margins, or MD-AM, framework introduced by Akande, Madson, Hillygus and Reiter (2021) and extended to surveys by Akande and Reiter (2022), Tang, Hillygus and Reiter (2024), and Yang and Reiter (2025). The latter three works impute missing values to ensure completed-data, survey-weighted inferences are plausible given the known margins. However, these works impose a simplifying condition on the unit nonrespondents, namely that the analyst does not use their design weights and instead replaces these weights with a convenient constant.

In this short note, we propose MD-AM models that allow use of the design-based weights for all sampled units. These models are intended especially for unequal probability samples where one wants to incorporate relationships between the design variables and survey variables in imputation. They presume design weights are available for all sampled units, as should be true for the organization that collected the data. Section 2 provides a brief review of the hybrid missing MD-AM model (Tang et al., 2024; Yang and Reiter, 2025),

1. Kewei Xu and Jerome P. Reiter, Department of Statistical Science, Box 90251, Duke University, Durham, NC 27708-0251. E-mail: jreiter@duke.edu.

which we extend to incorporate design weights for unit nonrespondents. Section 3 introduces the methodology. Section 4 illustrates the methods using simulation studies. Codes and additional results are available at https://github.com/kevinxu47/MDAM.

2. Review of hybrid missingness MD-AM modeling

In reviewing the hybrid missingness MD-AM model, we closely follow the notation in Yang and Reiter (2025). Let $\mathcal{P}$ be a finite population comprising N units. For $i = 1,\ldots,N$, let $\mathbf{x}_i = (x_{i1},\ldots,x_{ip})$ denote the p survey variables for unit i; let $I_i = 1$ if unit i is selected for inclusion in the survey and $I_i = 0$ otherwise; let $\pi_i = \Pr(I_i = 1)$ represent the probability that unit i is selected into the survey; let $w_i^d = 1/\pi_i$ be its design weight; and, let $\mathbf{z}_i$ be its design variables, e.g., size measures or stratum indicators. Let $n = \sum_{i=1}^{N} I_i$ be the intended sample size. We refer to the sampled units as $\mathcal{S}$.

Let U be the unit nonresponse indicator so that, for each unit $i \in \mathcal{S}$, $U_i = 1$ if the unit provides no responses to any survey questions and $U_i = 0$ otherwise. Let $\mathbf{R} = (R_1,\ldots,R_p)$ be item nonresponse indicators corresponding to $\mathbf{X} = (X_1,\ldots,X_p)$ so that, for any unit i with $U_i = 0$ and any X_j, $R_{ij} = 1$ if unit i does not respond to the question for X_j and $R_{ij} = 0$ otherwise. When $U_i = 1$, $(R_{i1},\ldots,R_{ij})$ are undefined.

The survey organization has auxiliary information $\mathcal{A}$ on the marginal distributions of a subset of $\mathbf{X}$. We write $X_j \in \mathcal{A}$ whenever X_j has a known margin in $\mathcal{A}$, which we write as $\mathcal{A}_j$. For convenience, we presume each $\mathcal{A}_j$ comprises population totals; incorporating percentages is a simple modification. For any categorical X_j taking on levels $c = 1,\ldots,m_j$, let $T_{jc} = \sum_{i=1}^{N} I(x_{ij} = c)$ be the total number of units in $\mathcal{P}$ at level c. Here, $I(\cdot) = 1$ when the condition inside the parenthesis is true and $I(\cdot) = 0$ otherwise.

As described in Tang et al. (2024) and Yang and Reiter (2025), the hybrid missingness MD-AM model specifies a joint distribution of $(\mathbf{X},\mathbf{R},U)$. Specifically, the model presumes $U \sim \text{Bernoulli}(\pi_u)$, where $\pi_u = \Pr(U = 1)$ is the marginal probability of unit nonresponse. For $j = 1,\ldots,p$, let $g_j(-)$ represent the model for X_j given $X_1,\ldots,X_{j-1}$, and let Ω_j and θ_j be the corresponding model parameters. For example, $g_j(-)$ could be a logistic regression of X_j on some function of $X_1,\ldots,X_{j-1}$, possibly including interactions, as well as a main effect for U when $X_j \in \mathcal{A}$. Thus, for $\mathbf{X} \mid U$, the hybrid missingness MD-AM model uses

$$X_1 \mid U \sim g_1(\Omega_1, \theta_1 U I(X_1 \in \mathcal{A}) \tag{2.1}$$

$$X_j \mid X_1,\ldots,X_{j-1},U \sim g_j(X_1,\ldots,X_{j-1},\Omega_j,\theta_j U I(X_j \in \mathcal{A}), \text{ for } j = 2,\ldots,p. \tag{2.2}$$

Because the distribution for any $X_j \in \mathcal{A}$ differs for unit nonrespondents and respondents when $\theta_j \neq 0$, the model encodes potentially missing not at random mechanisms for unit nonresponse. The key identifying assumption is that the predictor function for any $X_j \in \mathcal{A}$ not include interactions between U and elements of $(X_1,\ldots,X_{j-1})$. This assumption is a version of the additive nonignorable missingness mechanism (Hirano, Imbens, Ridder and Rubin, 2001; Sadinle and Reiter, 2019).

For the model for each R_j, let $h_j(-)$ be some predictor function that excludes the corresponding X_j but may include main and interaction effects involving other variables in $\mathbf{X}$. For $j = 1,\ldots,p$, we have

$$R_j \mid X_1,\ldots,X_p,U=0 \sim \text{Bernoulli}(\pi_{R_j}),\ \text{logit}(\pi_{R_j}) = h_j(X_1,\ldots,X_{j-1},X_{j+1},\ldots X_p,\Phi_j). \tag{2.3}$$

The models in (2.3) encode itemwise conditionally independent nonresponse (ICIN) mechanisms (Sadinle and Reiter, 2017), which are known to have identifiable parameters.

The imputation and estimation algorithms in Tang et al. (2024) and Yang and Reiter (2025) do not use the design weights for unit nonrespondents. Instead, they replace the design weights for unit nonrespondents with a constant that makes $\sum_{i=1}^{N} w_i = N$; that is, they base imputations and completed-data inferences on, for all units $i \in \mathcal{S}$,

$$w_i = \begin{cases} w_i^d & \text{if } U_i = 0 \\ \dfrac{N - \sum_{k\in\mathcal{S}} w_k^d}{\sum_{k\in\mathcal{S}} U_k} & \text{if } U_i = 1. \end{cases} \tag{2.4}$$

For any unit $i \in \mathcal{S}$ and $j = 1,\ldots,p$, let $x_{ij}^* = x_{ij}$ when $(R_{ij} = 0, U_i = 0)$, and let x_{ij}^* be an imputed value when $R_{ij} = 1$ or $U_i = 1$. For any $X_j \in \mathcal{A}$, Akande and Reiter (2022), Tang et al. (2024), and Yang and Reiter (2025) suppose the imputations adhere to (2.1)-(2.3), with the additional constraint that

$$\hat{T}_{jc}^* = \sum_{i\in\mathcal{S}} w_i\, I(x_{ij}^* = c) \sim N(T_{jc}, V_{jc}), \tag{2.5}$$

where the weights are from (2.4). Here, V_{jc} is set by the analyst and determines how closely $\hat{T}_{jc}^*$ matches T_{jc} in any completed dataset.

Using (2.5) for imputations has commonalities with balanced random imputation (Chauvet, Deville and Haziza, 2011; Chauvet and Do Paco, 2018), although the goals differ. In balanced random imputation, the imputed values are required always to satisfy some enforced constraint, e.g., the survey-weighted mean of the imputed values equals the survey-weighted mean of the observed values. Work on balanced random imputation tends to consider single imputations for item nonresponse, presume values are missing (completely) at random, and set constraints based on observed data quantities. In contrast, the hybrid missingness MD-AM model considers multiple imputation for unit nonresponse that is missing not at random, and it uses auxiliary marginal information to establish distributional constraints on imputations.

Although not done in Tang et al. (2024) or Yang and Reiter (2025), the model for each X_j or R_j can include the design variables $\mathbf{Z}$ as predictors. Including $\mathbf{Z}$ can improve explanatory power, and hence ultimately imputation quality and estimation accuracy, when they are associated with the survey variables or nonresponse indicators (Reiter, Raghunathan and Kinney, 2006). Alternatively, the models can include some function of W as a predictor. This is common advice for imputation modeling (e.g., Kim, Brick, Fuller and Kalton, 2006; Quartagno, Carpenter and Goldstein, 2020) that is particularly salient when the analyst doing the imputations has the survey weights but not the design variables, e.g., because they are confidential. In the MD-AM models of Tang et al. (2024) and Yang and Reiter (2025), however, every unit nonrespondent has the same value of w_i given in (2.4). Thus, including W in the models would not

influence the imputation probabilities, and hence not contribute to incorporating the design in the imputation process, for the unit nonrespondents.

3. Methodology

We now modify the hybrid missingness MD-AM model to use the design weights for all sampled individuals, including unit nonrespondents. For brevity of notation, we now let W refer to the design weights w_i^d rather than the weights in (2.4). We follow the general strategy outlined in Yang and Reiter (2025). First, we impute values for item nonresponse. Second, we use $\mathcal{A}$ to impute values of variables with margins for unit nonrespondents. Third, we impute values of the remaining variables for unit nonrespondents. Our innovation is to modify the second step so that imputations and the completed datasets use the design weights rather than (2.4) for unit nonrespondents. This change requires new imputation algorithms, which we present in Section 3.2. For the first and third steps, we use techniques presented in Yang and Reiter (2025), which we briefly summarize in Section 3.1 and Section 3.3. For convenience, we presume that $X_j \in \mathcal{A}$ for $j = 1,\ldots,k < p,$ and the remaining $p-k$ variables do not have auxiliary margins.

3.1 Variables with item nonresponse

We first create multiple imputations for all values of x_{ij} missing due to item nonresponse for units with $U_i = 0.$ We implement multiple imputation by chained equations (MICE) using only units with $U_i = 0$ and potentially including $\mathbf{Z}$ as predictors. The specification of the MICE algorithm does not require any extraordinary considerations. In our simulations, we use the "mice" package in R (Van Buuren, 2018), following its default settings including the ordering of variables. We run the MICE procedure to create L completed datasets for the units with $U_i = 0.$

3.2 Variables with margins for unit nonrespondents

After imputing values for item nonresponse, in each completed dataset we impute the unit nonrespondents' values for all $X_j \in \mathcal{A}.$ The analyst orders these variables based on ease of modeling. For convenience, we suppose that X_1 is imputed first, X_2 is imputed second, and so on sequentially until $X_k.$

To begin, for each unit i with $U_i = 1$ and in each of the L completed datasets, the analyst sets an initial probability distribution for imputing $x_{i1}.$ For example, in each completed dataset, the analyst estimates a (multinomial) logistic regression of X_1 on some function of $\mathbf{Z}$ (or W), and uses the predicted probabilities from the estimated model. Alternatively, the analyst could regress X_1 on an intercept only or compute marginal probabilities of X_1 using survey-weighted ratio estimators. For these two alternatives, the initial probabilities are identical for all unit nonrespondents. This approach may be especially appropriate when the analyst does not expect associations between the design weights and the missing values of X_1 for unit nonrespondents.

Regardless of how it is derived, we refer to the initial distribution for imputing x_{i1} as its working distribution, notated as $\{p_{i1c} : c = 1,\ldots,m_1; U_i = 1; i \in \mathcal{S}\}$. Because of imputation for item nonresponse, the working distribution for any unit can differ across completed datasets; however, for convenience, we forego notation designating completed datasets. The analyst performs the computations of this section for each of the L completed datasets.

Imputing each missing x_{i1} using its working distribution does not utilize the information in $\mathcal{A}_1$. The resulting imputations could generate completed-data Horvitz and Thompson (1952) estimates that are far from the known totals for some levels of X_1, particularly when values are missing not at random. We therefore modify $\{p_{i1c}\}$ to ensure the resulting imputations result in reasonable design-based, completed-data estimates of these totals. This is the primary objective and motivation of our extension of the MD-AM framework.

To do so, we first simulate plausible values of $\hat{T}_{1c}$ for each level c by appealing to large-sample central limit theorems. For $c = 1,\ldots,(m_1 - 1)$, we sample $\hat{T}_{1c}$ from $N(T_{1c}, V_{1c})$ and set $\hat{T}_{1m_1} = N - \sum_{c=1}^{m_1-1} \hat{T}_{1c}$. We then find adjusted probabilities $\{\tilde{p}_{i1c} : c = 1,\ldots,m_1; U_i = 1; i \in \mathcal{S}\}$ so that, for $c = 1,\ldots,m_1$, the expectation of each completed-data estimator $\hat{T}^*_{1c}$ from (2.5) over imputations for unit nonresponse approximately equals $\hat{T}_{1c}$. Put another way, recognizing that

$$\begin{aligned}\mathbb{E}\left[\sum_{i\in\mathcal{S}} w_i\, I(x^*_{ij} = c)\right] &= \sum_{i\in\mathcal{S}} w_i\, I(x^*_{ij} = c)\, I(U_i = 0) + \mathbb{E}\left[\sum_{i\in\mathcal{S}} w_i\, I(x^*_{ij} = c)\, I(U_i = 1)\right] \\ &= \sum_{i\in\mathcal{S}} w_i\, I(x^*_{ij} = c)\, I(U_i = 0) + \sum_{i\in\mathcal{S}} w_i \tilde{p}_{i1c}\, I(U_i = 1),\end{aligned}$$

we want

$$\sum_{i\in\mathcal{S}} w_i \tilde{p}_{i1c}\, I(U_i = 1) \;=\; \hat{T}_{1c} - \sum_{i\in\mathcal{S}} w_i\, I(x^*_{i1} = c)\, I(U_i = 0).$$

To keep the imputation probabilities tied to the working distribution, for $c = 1,\ldots,m_1 - 1$, we set $\tilde{p}_{i1c} = f_{1c} p_{i1c}$, where each constant $f_{1c} > 0$ is given by

$$f_{1c} \;=\; \frac{\hat{T}_{1c} - \sum_{i\in\mathcal{S}} w_i\, I(x^*_{i1} = c, U_i = 0)}{\sum_{i\in\mathcal{S}} p_{i1c} w_i\, I(U_i = 1)}. \tag{3.1}$$

We impute x^*_{i1} for each unit with $U_i = 1$ by drawing randomly from the adjusted probability distribution, $\{\tilde{p}_{i1c} : c = 1,\ldots,m_1\}$, where $\tilde{p}_{i1m_1} = 1 - \sum_{c=1}^{m_1-1} f_{1c} p_{i1c}$. If $\sum_{c=1}^{m_1-1} f_{1c} p_{i1c} > 1$, we compute (3.1) for $c = 1,\ldots,m_1$ and set $\tilde{p}_{i1c} = f_{1c} p_{i1c} \Big/ \sum_{c=1}^{m_1} f_{1c} p_{i1c}$. In the event that some $f_{1c} p_{i1c} < 0$, we set those $\tilde{p}_{i1c} = 0$. Analysts may consider reducing V_{1c} in this case.

We now turn to imputing $X_2 \in \mathcal{A}$ for unit nonrespondents. We follow a similar strategy: in each completed dataset, start by setting each unit nonrespondent's working distribution for imputation of x^*_{i2} given x^*_{i1} and $(\mathbf{z}_i, w_i)$, then adjust the probabilities to make $(\hat{T}^*_{21},\ldots,\hat{T}^*_{2m_2})$ approximately match a sampled value of $(\hat{T}_{21},\ldots,\hat{T}_{2m_2})$. We present two options for imputing X_2, one based on applying a multiplicative

adjustment to the working probabilities (Section 3.2.1) and the other based on solving a system of equations making use of $\mathcal{A}$ (Section 3.2.2).

3.2.1 Multiplicative adjustment method

Let $p_{i2c} = \Pr(X_{i2} = c \mid x_{i1}^* = d, \mathbf{z}_i, w_i)$ be the working probability that $X_2 = c$ for unit i, given its imputed $x_{i1}^* = d$ and $(\mathbf{z}_i, w_i)$. We determine these from a (multinomial) logistic regression of X_2 on X_1 and possibly some function of $\mathbf{Z}$ or W, with coefficients estimated from the completed data for units with $U_i = 0$. Note that if we disregard $\mathbf{Z}$ and W, $p_{i2c} = \Pr(X_2 = c \mid X_1 = d)$ for $c = 1,\ldots,m_2$ and $d = 1,\ldots,m_1$; that is, the conditional probability is the same for all units with $x_{i1}^* = d$. For this case (which we use in Section 3.2.2), to simplify notation we drop the index i for individual units and add a subscript d for the value of X_1, writing $p_{i2c} = p_{2cd}$.

For $c = 1,\ldots,m_2 - 1$, we sample a plausible value of $\hat{T}_{2c}$ from a $N(T_{2c}, V_{2c})$, setting the estimated total for the final level as $\hat{T}_{2m_2} = N - \sum_{c=1}^{m_2-1} \hat{T}_{2c}$. As in (3.1), we find suitable constants $f_{2c} > 0$ to adjust each p_{i2c}. We have

$$f_{2c} = \frac{\hat{T}_{2c} - \sum_{i \in \mathcal{S}} w_i I(x_{i2}^* = c, U_i = 0)}{\sum_{i \in \mathcal{S}} p_{i2c} w_i I(U_i = 1)}. \tag{3.2}$$

We impute each x_{i2}^* given x_{i1}^* and $\mathbf{z}_i$ (or w_i) for units with $U_i = 1$ using a random draw from the adjusted probability mass function $\{\tilde{p}_{i2c} : c = 1,\ldots,m_2\}$, where $\tilde{p}_{i2c} = f_{2c} p_{i2c}$ for $c = 1,\ldots,m_2 - 1$ and $\tilde{p}_{i2m_2} = 1 - \sum_{c=1}^{m_2-1} f_{2c} p_{i2c}$. If $\sum_{c=1}^{m_2-1} f_{2c} p_{i2c} > 1$, we compute (3.2) for $c = 1,\ldots,m_2$ and set $\tilde{p}_{i2c} = f_{2c} p_{i2c} \Big/ \sum_{c=1}^{m_2} f_{2c} p_{i2c}$. As for X_1, if some $f_{2c} p_{i2c} < 0$, we set those $\tilde{p}_{i2c} = 0$. This process ensures the completed-data Horvitz and Thompson (1952) estimates approximately match the sampled $(\hat{T}_{21},\ldots,\hat{T}_{2m_2})$ in expectation, while also incorporating relationships between X_2, X_1 and $\mathbf{Z}$ (or W) implied in the working probabilities. We independently repeat the estimation and imputation process in each completed dataset. We refer to this method as MDAM–adj.

When $\{p_{i2c}\}$ derives from a logistic regression of X_2 on X_1 and $\mathbf{Z}$ (or W), using $\{\tilde{p}_{i2c}\}$ is equivalent to adjusting the intercept in that regression, leaving other coefficients alone. This implies a model in which the log-odds for X_1 are the same for unit respondents and nonrespondents, in accordance with no interactions between U and X_1 in (2.2).

3.2.2 Systems of equations method

MDAM–adj captures the relationship between X_2 and X_1 through the logistic regression used in the working probabilities. Another approach is to impose constraints on the imputations through a system of equations implied by the assumptions underpinning (2.2) and (2.5), as we now describe.

As in Section 3.2.1, we first sample plausible values of $(\hat{T}_{21},\ldots,\hat{T}_{2m_2})$. For now, we presume the working probabilities follow $p_{i2c} = p_{2cd}$ for all units. Within any completed dataset, we encode the constant log-odds assumption in (2.2) as a set of $(m_1-1)(m_2-1)$ equations. For $c = 2,\ldots,m_2$ and $d = 2,\ldots,m_1$, we have

$$\log\left[\frac{p_{2cd}}{p_{21d}}\right] - \log\left[\frac{p_{2c1}}{p_{211}}\right] = \log\left[\frac{\Pr(X_2 = c \mid X_1 = d, U = 0)}{\Pr(X_2 = 1 \mid X_1 = d, U = 0)}\right] - \log\left[\frac{\Pr(X_2 = c \mid X_1 = 1, U = 0)}{\Pr(X_2 = 1 \mid X_1 = 1, U = 0)}\right]. \quad (3.3)$$

We estimate the conditional probabilities for unit respondents in (3.3) via survey-weighted ratio estimators using the completed data; for example,

$$\Pr(X_2 = c \mid X_1 = d, U = 0) = \frac{\sum_{i\in\mathcal{S}} w_i\, I(x^*_{i1} = d, x^*_{i2} = c, U_i = 0)}{\sum_{i\in\mathcal{S}} w_i\, I(x^*_{i1} = d, U_i = 0)}. \quad (3.4)$$

We encode the conditions on $(\hat{T}^*_{21},\ldots,\hat{T}^*_{2m_2})$ in (2.5) as a set of m_2 equations. We have

$$\sum_{d=1}^{m_1}\sum_{i\in\mathcal{S}} w_i p_{2cd}\, I(x^*_{i1} = d, x^*_{i2} = c, U_i = 1) = \hat{T}_{2c} - \sum_{i\in\mathcal{S}} w_i\, I(x^*_{i2} = c, U_i = 0). \quad (3.5)$$

We solve these equations for the $m_1(m_2-1)$ conditional probabilities, $\{p_{2cd} : c = 2,\ldots,m_2; d = 1,\ldots m_1\}$. We impute x^*_{i2} given x^*_{i1} for the unit nonrespondents using Bernoulli draws with probabilities $\{p_{2cd}\}$. We refer to this method as MDAM–sys.

It is possible to modify MDAM–sys to use working probabilities $\{p_{i2c}\}$ that vary for units with the same value of x^*_{i1}. To do so, we impute using Bernoulli draws with probabilities $\{\tilde{p}_{i2c}\}$, where $\tilde{p}_{i2c} = f_{2cd}\, p_{i2c}$ and

$$f_{2cd} = \frac{\sum_{i\in\mathcal{S}} w_i p_{2cd}}{\sum_{i\in\mathcal{S}} w_i p_{i2c}} \quad (3.6)$$

for each $(X_2 = c, X_1 = d)$. The adjustment in (3.6) is motivated by matching $\sum_{i\in\mathcal{S}} w_i p_{2c} = \sum_{i\in\mathcal{S}} w_i f_{2cd}\, p_{i2c}$.

3.2.3 Accounting for additional variables with margins

When $k > 2$, for MDAM–adj we can apply the process used in (3.2) to each $X_j \in \mathcal{A}$. For example, if $X_3 \in \mathcal{A}$, the analyst specifies a set of working probabilities, $p_{i3c} = \Pr(X_{i3} = c \mid x^*_{i1}, x^*_{i2}, \mathbf{z}_i, w_i)$ for $c = 1,\ldots,m_3$ via a logistic regression of X_3 on X_1, X_2, and $\mathbf{Z}$ (or W). The analyst samples values of $(\hat{T}_{31},\ldots,\hat{T}_{3m_3})$ and computes values of f_{3c} using expressions analogous to (3.2), replacing quantities based on X_2 with those based on X_3. For MDAM–sys, the analyst can solve a system of equations akin to those in (3.3) and (3.5). However, MDAM–sys becomes increasingly complicated as more variables are added to $\mathcal{A}$.

3.3 Variables without margins for unit nonrespondents

After creating L partially completed datasets using Section 3.1 and 3.2, we impute values x^*_{ij} for unit nonrespondents for all $X_j \notin \mathcal{A}$. We use the random hot deck imputation procedure developed by Yang and Reiter (2025). Paraphrasing from their presentation, for any unit $i \in \mathcal{S}$ with $U_i = 0$, in any completed

dataset, let $\mathbf{x}_i^{\mathcal{A}} = \{x_{ij} : X_j \in \mathcal{A}; j = 1,\ldots,p; U_i = 0\}$ be the values of the (possibly imputed) survey variables in the completed data for those $X_j \in \mathcal{A}$. Similarly, for any unit $i' \in \mathcal{S}$ with $U_{i'} = 1$, let $\mathbf{x}_{i'}^{\mathcal{A}*} = \{x_{i'j}^* : X_j \in \mathcal{A}; j = 1,\ldots,p; U_i = 1\}$. For each unit i' with $U_{i'} = 1$, in each completed dataset we construct its donor set, $\mathcal{D}_{i'} = \{(x_{i1},\ldots,x_{ip}) : \mathbf{x}_i^{\mathcal{A}} = \mathbf{x}_{i'}^{\mathcal{A}*}, U_i = 0, i \in \mathcal{S}\}$. We randomly sample one record i from $\mathcal{D}_{i'}$ and append its $\{(x_{i1},\ldots,x_{ik}) : X_j \notin \mathcal{A}\}$ to $\mathbf{x}_{i'}^{\mathcal{A}*}$ to make the full imputation $\mathbf{x}_{i'}^*$ for unit i'. We apply this procedure for all units in $\mathcal{S}$ with $U_{i'} = 1$, resulting in a completed dataset for all n units. We repeat this process in each of the L datasets.

4. Simulation studies

We construct a population $\mathcal{P}$ comprising $N =$ 3,405,809 individuals from the 2023 American Community Survey public use files available from IPUMS. Each individual has the six variables described in Table 4.1. We fill in any missing values using a single bespoke run of the "mice" package. We treat the survey weight on the file as a known size measure, $Z = (z_1,\ldots,z_N)$. To generate any $\mathcal{S}$, we sample records from $\mathcal{P}$ independently with probabilities $\pi_i = 1/10z_i$, where $i = 1,\ldots,N$. This Poisson sampling results in approximately $n =$ 7,000 individuals in any $\mathcal{S}$. We take 500 independent Poisson samples.

Table 4.1
Description of variables in the simulation study

Variable	Notation	Range
Sex	X_1	1 = Male, 0 = Female
Marital status	X_2	1 = Married, 0 = Other
Any health insurance coverage	X_3	1 = Yes, 0 = No
School attendance	X_4	1 = Yes, 0 = No
Travel time to work	X_5	0 to 888
Occupational income score	X_6	0 to 80

For each $\mathcal{S}$, we generate unit nonresponse by sampling U_i for all $i \in \mathcal{S}$ from

$$U \sim \text{Bernoulli}(\pi_U) \quad \text{logit}(\pi_U) = \omega_0 + \omega_1 X_1 + \omega_2 X_2. \tag{4.1}$$

We set $(\omega_0, \omega_1, \omega_2) =$ (-1.6, 0.5, 0.5) for a unit nonresponse rate around 25%. We make all data other than Z and W completely missing for every unit where $U_i = 1$. To generate item nonresponse for any unit i with $U_i = 0$, we sample R_{ij} for $j = 1,\ldots,6$ using

$$\begin{aligned} &R_j \mid X_1, X_2, X_3, X_4, X_5, X_6, Z, U = 0 \sim \text{Bernoulli}(\pi_{R_j}), \\ &\text{logit}(\pi_{R_j}) = \phi_{j0} + \sum_{k \neq j} \phi_{jk} X_k + \phi_Z Z. \end{aligned} \tag{4.2}$$

We set parameters in (4.2) so that approximately 20% of each survey variable is missing for units with $U_i = 0$. Specifically, ϕ_Z = 0.00001; ϕ_{j0} = -1.5 when $j \leq 4$ and ϕ_{j0} = -1.6 when $j = 5, 6$; and, ϕ_{jk} = 0.1 when $k \leq 4$ and ϕ_{jk} = 0.0001 when $k = 5, 6,$ for all j. We blank every x_{ij} for which the sampled $R_{ij} = 1$.

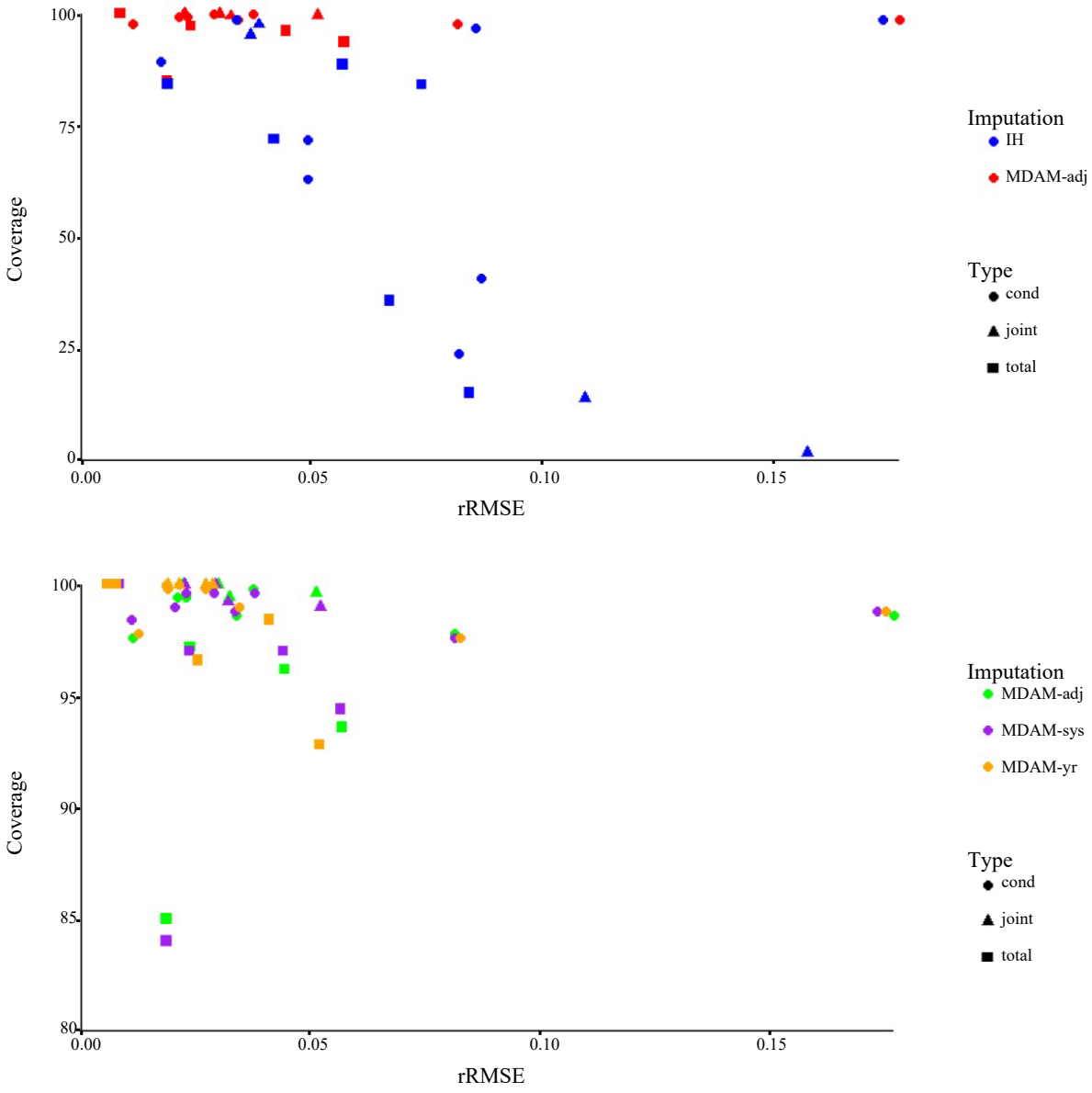
We let $\mathcal{A}$ include $T_{X_1} = \sum_{i=1}^{N} x_{i1}$ and $T_{X_2} = \sum_{i=1}^{N} x_{i2}$. Using Section 3.1, we generate L = 10 completed datasets for item nonresponse using a bespoke implementation of the "mice" package. Using Section 3.2, we implement MDAM–adj and MDAM–sys with working probabilities derived from logistic regressions of X_1 on Z and of X_2 on (X_1, Z). For V_{X_1} and V_{X_2} used to sample $\hat{T}_{X_1} \sim N(T_{X_1}, V_{X_1})$ and $\hat{T}_{X_2} \sim N(T_{X_2}, V_{X_2})$, we use "mice" to make one completed dataset for all of $\mathcal{S}$ and compute the expressions for the unbiased variance estimators for $\hat{T}_{X_1}$ and $\hat{T}_{X_2}$ under Poisson sampling. Using Section 3.3, we create donor pools for the hot deck by matching on (X_1, X_2).

After generating $L = 10$ completed datasets, we use the inferential methods in Rubin (1987) for the marginal totals for $(X_1, \ldots, X_6)$ and several conditional and joint probabilities; see https://github.com/kevinxu47/MDAM for the full list. In each completed dataset, we compute Horvitz and Thompson (1952) point and variance estimators under Poisson sampling based on the design weights for all sampled records. For comparison, we also use the method in Yang and Reiter (2025) with the weights in (2.4). We refer to this method as MDAM–yr. Finally, we use a bespoke implementation of "mice" to impute missing items and randomly sample whole records with replacement as imputations for unit nonresponse. We refer to this method as IH. It does not utilize $\mathcal{A}$.

Figure 4.1 summarizes the relative root mean squared errors (rRMSEs) of the point estimates and empirical coverage rates of 95% multiple imputation confidence intervals for the methods across the 500 samples. Results in tabular form are available at https://github.com/kevinxu47/MDAM. MDAM–adj tends to produce lower rRMSEs and higher coverage rates than IH. The coverage rates for the MD-AM models tend to exceed 95% due to over-estimation of variances. As discussed by Yang and Reiter (2025), this is because the combining rules of Rubin (1987) do not account for using known totals in imputation. MDAM–adj and MDAM–sys offer qualitatively similar results. For both, only the interval for T_{X_3} has a lower than nominal coverage rate (around 85%). This rate approximately matches the coverage rate of design-based confidence intervals estimated before introducing any missing data (see the tabular results), suggesting that the lower than nominal coverage is a feature of the complete data estimator, not a consequence of MDAM modeling. MDAM–yr performs similarly to MDAM–adj and MDAM–sys. As evident in the tabular results, MDAM–yr tends to have lower variances, which results from flattening all weights for unit nonrespondents to a constant.

Overall, the results suggest that the new MD-AM models enable analysts to incorporate known marginal totals in multiple imputation when using all sampled units' design weights.

Figure 4.1 Simulated rRMSEs and coverage rates for population totals, two-way conditional probabilities, and joint probabilities



Notes: - Top panel compares MDAM–adj and IH. Bottom panel displays MDAM–adj, MDAM–sys, and MDAM–yr.
- MDAM-adj and MDAM-sys are missing data with auxiliary margins models proposed here. MDAM-yr is the missing data with auxiliary margins model of Yang and Reiter. Imputation Hot-deck (IH) is the model that uses itemwise conditional independence nonreponse modeling for item nonresponse and a hot deck for unit nonresponse. rRMSE is relative root mean squared error.

References

Akande, O., Madson, G., Hillygus, D.S. and Reiter, J.P. (2021). Leveraging auxiliary information on marginal distributions in nonignorable models for item and unit nonresponse. *Journal of the Royal Statistical Society Series A*, 184, 643-662.

Akande, O. and Reiter, J.P. (2022). Multiple imputations for nonignorable item nonresponse in complex surveys using auxiliary margins. In *Statistics in the Public Interest: In Memory of Stephen E. Fienberg*, 289-306. Cham: Springer International Publishing.

Chauvet, G., Deville, J.-C. and Haziza, D. (2011). On balanced random imputation in surveys. *Biometrika*, 98, 459-471.

Chauvet, G. and Do Paco, W. (2018). Exact balanced random imputation for sample survey data. *Computational Statistics & Data Analysis*, 128, 1-16.

Hirano, K., Imbens, G.W., Ridder, G. and Rubin, D.B. (2001). Combining panel data sets with attrition and refreshment samples. *Econometrica*, 69(6), 1645-1659.

Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

Kim, J.K., Brick, J.M., Fuller, W.A. and Kalton, G. (2006). On the bias of the multiple-imputation variance estimator in survey sampling. *Journal of the Royal Statistical Society: Series B*, 68, 509-521.

Quartagno, M., Carpenter, J.R. and Goldstein, H. (2020). Multiple imputation with survey weights: A multilevel approach. *Journal of Survey Statistics and Methodology*, 8, 965-989.

Reiter, J.P., Raghunathan, T.E. and Kinney, S.K. (2006). The importance of modeling the sampling design in multiple imputation for missing data. *Survey Methodology*, 32(2), 143-149. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2006002/article/9548-eng.pdf.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.

Sadinle, M. and Reiter, J.P. (2017). Itemwise conditionally independent nonresponse modelling for incomplete multivariate data. *Biometrika*, 104, 207-220.

Sadinle, M. and Reiter, J.P. (2019). Sequentially additive nonignorable missing data modelling using auxiliary marginal information. *Biometrika*, 106, 889-911.

Tang, J., Hillygus, D.S. and Reiter, J.P. (2024). Using auxiliary marginal distributions in imputations for nonresponse while accounting for survey weights, with application to estimating voter turnout. *Journal of Survey Statistics and Methodology*, 12, 155-182.

van Buuren, S. (2018). *Flexible Imputation of Missing Data*. CRC Press.

Yang, Y. and Reiter, J.P. (2025). Imputation of nonignorable missing data in surveys using auxiliary margins via hot deck and sequential imputation. *Survey Methodology*, 51(1), 251-274. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2025001/article/00004-eng.pdf.