**RUNNING HEAD:** Improved Mass Imputation in Probability Samples

**Improved Mass Imputation in Probability Samples via Adjustment of Imputation Models Based on Non-Probability Samples for Selection Bias**

**Brady T. West**

Survey Research Center, Institute for Social Research

University of Michigan-Ann Arbor

426 Thompson Street, Room 4118

Ann Arbor, MI 48109

[bwest@umich.edu](mailto:bwest@umich.edu)

ORCID: 0000-0003-0160-1388

**Rebecca R. Andridge**

Division of Biostatistics, College of Public Health

The Ohio State University

[andridge.1@osu.edu](mailto:andridge.1@osu.edu)

ORCID: 0000-0001-9991-9647

**Abstract**

Methods for integrating probability and non-probability samples have grown in popularity in recent years, in part because it may be infeasible to measure a variable of scientific interest in a large probability sample, but feasible to do so in a non-probability sample. One such method is mass imputation, where a variable that is not measured in the probability sample is fully imputed in that sample using a model fitted to the non-probability sample. A fundamental assumption of this method is that the imputation model fitted to the non-probability sample is transportable to the probability sample, meaning that the same regression function for the variable of interest holds for both samples. If the selection mechanism for the non-probability sample is non-ignorable, given the variable of interest, then there may be significant bias in the imputation model coefficients, resulting in poor imputations and subsequent bias in the mass imputation estimator. This paper leverages recent work on novel measures of selection bias for regression coefficients to propose an adjustment of the imputation model coefficients for selection bias prior to imputation. Via two simulation studies and a case study, we demonstrate the ability of this method to improve estimates produced using mass imputation procedures.

**Key Words:** imputation models, mass imputation, non-ignorable selection, non-probability samples, probability samples, selection bias

**1. Introduction**

Given the current drawbacks associated with collecting large amounts of data from large probability samples to make inference about finite populations, including high costs and low response rates that raise questions about whether the final measured sample is truly still a probability sample, many survey researchers worldwide are turning to the integration of probability-based and non-probability samples for the purpose of making finite population inferences (Wisniowski et al. 2020). One popular and effective approach for this type of data integration, known as *mass imputation*, involves collecting measures on unique variables of interest from a non-probability sample, fitting models to that sample enabling predictions of values on the variables of interest as a function of a specified set of covariates, and then imputing values on those variables for all cases in a large probability-based sample, ultimately making inference based on the predictions in the probability sample (Kim et al. 2021). This approach is fairly straightforward and clearly relies on well-specified models that can effectively predict the unique variables of interest.

The mass imputation approach relies on the crucial assumption that the imputation models fitted to the non-probability sample are *transportable* from the non-probability sample to the probability sample (Kim et al. 2021). Put differently, under the assumption that the non-probability sample is arising from the same population from which the probability-based sample arises, this assumption means there is no selection bias associated with the estimated model parameters, and that the same imputation model holds for both the selected non-probability sample and the probability sample. This crucial assumption is violated if the (often unknown) selection mechanism that gives rise to the non-probability sample is *non-ignorable* when conditioning on the outcome variables of interest, which could introduce substantial bias in estimates of the parameters in the imputation model (West et al. 2021). Applying this biased imputation model to the probability sample for the purpose of making inference is likely to result in biased imputations and subsequent bias in imputation-based estimators of the finite population parameters of interest.

Prior work in this area has addressed this selection bias problem to a limited extent, generally focusing on models for outcome variables as a function of known, and possibly unequal, selection probabilities in probability-based samples (Choi et al. 2021; Xu et al. 2020; Ma et al. 2018; Pfeffermann 2017; Nandram et al. 2013) or the design approach known as *balanced sampling* (Royall and Pfeffermann 1982). This paper proposes a novel method for adjusting estimates of the regression coefficients in the imputation model fitted to a *non-probability sample* for non-ignorable selection bias when using this mass imputation framework for making finite population inferences. In the presence of non-ignorable selection mechanisms, this type of adjustment is important for avoiding the bias in imputation-based estimators described above. Such an adjustment would eliminate the need to rely on the assumption of transportability of the imputation model, enabling more flexible approaches for finite population inference. We also evaluate the performance of this novel method against other popular adjustment or imputation-based approaches for making finite-population inference when attempting to integrate non-probability and probability samples.

Section 2 outlines the approach underlying our novel adjustment methodology. Section 3 reviews alternative estimators in this general framework of integrating non-probability and probability samples, including discussion of estimators that allow for non-ignorable selection in the non-probability sample. Section 4 presents two simulation studies evaluating the alternative approaches empirically, while Section 5 presents a case study based on real survey data. Section 6 concludes with a discussion of the findings and suggested directions for future work in this area.

**2. Adjustment Approach**

**2.1. Notation and Data Sources**

We consider the following context for this methodology. A researcher wishes to make inference related to the mean of a variable *Y*, denoted by , in a finite population of size *N* elements. The researcher obtains a selected non-probability sample of size *n* from the finite population, with individual elements in the selected sample indexed by *i* (*i* = 1, …, *n*). Without loss of generality, we assume that the same element from the finite population cannot appear in the selected non-probability sample multiple times.

The researcher is able to measure three types of variables for elements in the selected non-probability sample. The variable *Y* is the dependent variable of interest, and the measurement of *Y* is unique to the selected sample (no information on the variable *Y* is available in any other data sources). In addition to measures of *Y*, the researcher also collects measures on a vector of covariates *Z*, representing scientifically relevant predictors of *Y* that the researcher has primary interest in analyzing. In addition to the collected measures on *Z*, the researcher also has measures available on a vector of additional auxiliary variables denoted by *A*. These are variables that are not of primary scientific interest to the researcher, but may still be correlated with the variable of interest *Y* when conditioning on the predictors of primary interest *Z*. The microdata on *Y*, *Z*, and *A* are only available for the selected non-probability sample.

In addition to the selected non-probability sample, the researcher also has access to a large probability sample (selected from the same target population) that includes microdata for *Z* and *A*, but not *Y*. A dummy variable *S* indicates sample membership, being equal to 1 if a case was selected into the non-probability sample and 0 if a case was selected into the probability sample. We assume that 1) a given case cannot appear in both samples, and 2) the probability sampling is based on an *ignorable* sample selection mechanism, where after proper weighting based on the known probabilities of selection, finite population estimates of parameters from the probability sample are unbiased with respect to the sample design. Given the researcher’s goal of making inference about the mean of *Y* in the target population, the researcher needs to impute values of *Y* for all units in the probability sample, using an imputation model fitted to *Y* that includes the *Z* variables as the most important predictors of *Y*. The researcher will then base their finite population inference on the predicted values of *Y* in the probability sample (Kim et al. 2021). The researcher can also use the (weighted) probability sample to compute estimates of means, variances, and covariances of the predictors and auxiliary variables *Z* and *A* for the finite population of interest, with minimal sampling error.

Clearly the researcher could also include the *A* variables in the imputation model fitted to the non-probability sample. We will demonstrate that these auxiliary variables are more important for reducing selection bias in the estimated relationships of the (scientifically relevant) *Z* variables with *Y* than as additional predictors in an imputation model for *Y* that is limited by biased estimates of the relationships of the *Z* variables with *Y*.

**2.2. Adjustment of Imputation Models for Selection Bias**

Our proposed approach, which we denote as PPMM-Reg (or, regression based on a proxy pattern-mixture model), draws on work by West et al. (2021). Given the data sources described above, one can use the following four-step approach to *adjust* draws of the coefficients in the imputation model for potential selection bias introduced by a non-ignorable selection mechanism:

1. **Define an auxiliary proxy variable for the dependent variable of interest *Y*.**

We use the notation  for the intercept, the regression coefficients (written as a row vector) and residual variance of the regression of *Y* on *Z* and *A* given *S* = *s*. Suppose that

, (1)

where  denotes the best predictor of *Y* in the non-probability sample (*S*  = 1) after conditioning on *Z*, and  represents all of the parameters defining this distribution. We refer to *X* as the *auxiliary proxy variable* for *Y*, given the predictors of interest *Z*. Without loss of generality, we allow for selected variables in *A* (denoted by *V* in West et al. 2021) to be orthogonal to *X*, but do not introduce that notation here; see West et al. (2021) for details.

1. **Given the auxiliary proxy variable *X*, specify a bivariate normal pattern-mixture model (PMM) for *X* and *Y*, given the covariates of interest *Z* and the sample indicator *S*.**

We now assume the following bivariate normal PMM for the distribution of *X* and *Y* given *S* and *Z*:

(2)

An important component of the PMM in (2) is the probability model for the selection indicator *S*. Following Little (1994) and Andridge and Little (2011), this distribution is left unspecified, and the unidentified parameters in the PMM are identified by assuming that given *X*, *Y*, and *Z*, selection into the non-probability sample depends on *X* and *Y* only through a known linear combination of these variables. Specifically, we assume that

, (3)

where *g*() is an arbitrary unknown function, and  is *X* rescaled to have the same residual variance as *Y* after conditioning on *Z*. We call *X*\* the *scaled auxiliary proxy* for *Y*; the transformation from *X* to *X*\* simplifies the interpretation of  by putting *Y* and *X*\* on the same scale. West et al. (2021) refer to  as a *sensitivity parameter* that governs the dependence of selection into the non-probability sample on the scaled auxiliary proxy and the dependent variable of interest *Y*.

This assumed probability model in (3) identifies the PMM in (2) when  is known. In practice,  is unknown; because *X* is an auxiliary proxy for *Y*, we assume that  is positive, that is, .When , selection depends only on the observed variables *Z* and *X*, and hence selection is ignorable, per Rubin (1976). At the other extreme, when , selection depends only on *Y* and *Z*. The sensitivity parameter  is therefore a measure of the “degree of non-random selection,” after conditioning on *Z* and *X*, and no information is available on  in the data.

1. **Adjust the coefficients of interest and the residual variance based on the PMM.**

Given the PMM specified in (2) and (3), the transformation  introduced above (West et al. 2021; Andridge and Little 2011) yields adjusted versions of the intercept, regression coefficients, and residual variance for the population from which the probability sample was selected:

|  |  |
| --- | --- |
|  | (4) |
|  |

where the correlation of *X* and *Y* conditional on *Z* for the selected cases (*S* = 1) is  Estimates (or draws) of the parameters  from the regression of *X* on *Z*, which are critical for the results in (4), can also be obtained from regression models fitted to each respective pattern (*S* = 0,1), where survey weights for the probability sample would be used to obtain these estimates for *S* = 0. We can use the microdata from the non-probability sample to obtain estimates of all parameters for the selected cases (*S* = 1).

As described in West et al. (2021), we can also define measures of unadjusted bias for the regression coefficients. Taking the differences between the estimates (or Bayesian posterior draws) of the parameters of the regression of *Y* on *Z* for the selected (non-probability) sample and the population from which the probability sample was selected (from (4)) yields a *Measure of Unadjusted Bias* for the regression coefficients in the selected sample as compared to the *Non-Selected* cases (MUBNS) as a function of . Looking at the posterior distribution of MUBNS for each parameter can provide a sense of the magnitude of potential bias in the coefficients and serves as an indicator of how much adjustment will be made to the coefficients estimated from the selected sample when adjusting them for use in mass imputation.

One can obtain posterior distributions for the parameters above by performing a fully Bayesian analysis with a Uniform(0,1) prior distribution on , as described by Little et al. (2020) and Andridge et al. (2019). This Bayesian approach allows us to fully account for the uncertainty in all input estimates required for the proposed indices (e.g., the estimated regression coefficients used to form the auxiliary proxy *X*). We note that as there is no information about in the data (without additional assumptions), the posterior distributions will be strongly dependent on the choice of prior. A Uniform(0,1) prior implies no knowledge about the magnitude of nonignorable selection; other priors are possible (see for example the discrete prior in West et al. (2021)).

1. **Given the adjusted draws of the coefficients and the residual variance (for *S* = 0), draw imputed values for each case in the large probability sample from their respective posterior predictive distributions, and compute finite population estimates of the mean of interest in the probability sample.**

For a given draw of the adjusted intercept and regression coefficients from Step 3 (for *S* = 0), we compute the expected value of the variable of interest *Y* for each case in the large probability sample using the microdata for *Z*. This expected value defines the mean of a Normal posterior predictive distribution, while the draw of the adjusted residual variance defines the variance of that distribution. One value of *Y* is drawn from each posterior predictive distribution for each case in the large probability sample, resulting in a new version of the probability sample with an imputed value of *Y* for each case based on the *adjusted* regression coefficients and residual variance. These predicted values are used to estimate the mean of *Y*, using the appropriate design weights associated with the large probability sample. Following Chen et al. (2022), this process is then repeated a large number of times (using the large number of draws from Step 4) to generate a 95% credible interval for the mean of *Y*, enabling inference about the finite population mean.

**3. Alternative Estimators of Population Means Integrating Probability and Non-Probability Samples**

The proposed mass imputation approach based on adjusted versions of the imputation model coefficients from the non-probability sample (PPMM-Reg) is certainly not the only approach available for making finite population inference based on combined probability and non-probability samples. In this section, we describe alternative approaches in the literature that one could use for a similar purpose. We will evaluate each of these competing approaches in later sections.

**3.1. Competing Approach 1: Mass Imputation under the Transportability Assumption**

As a first alternative approach to making inference about the finite population mean, we consider the mass imputation approach outlined in Kim et al. (2021), assuming that the imputation model fitted to the non-probability sample is transportable to the probability sample. This approach involves fitting a parametric model to the variable of interest for all cases in the non-probability sample, with no adjustment for the possibility of non-ignorable selection, and using that model (and the common covariates) to compute predicted values (once) for each case in the probability sample. Variance estimation can be conducted using linearization or a bootstrapping approach (Kim et al. 2021), modified to account for complex sampling features (if applicable). This competing approach represents the “standard” mass imputation approach for making inference based on the probability sample and relies on the presence of microdata on all variables for all cases in the probability sample. This “standard” approach, which we label as MI, is implemented in the *nonprobsvy* contributed package for the R software (Chrostowski et al. 2025).

**3.2. Competing Approach 2: Adjustment Based on the SMUB Measure**

As a second alternative approach, we consider adjusted estimates of the finite population mean based on the *Standardized Measure of Unadjusted Bias* (SMUB) originally proposed and evaluated by Little et al. (2020). This approach relies on *aggregate* population information on *Z* and *A* to compute an auxiliary proxy for the variable of interest *Y* and does not require microdata for subsequent inference about the finite population mean. We note that this approach does not result in imputed values for the variable of interest in the probability sample, so if that is a goal of the analysis (enabling various subsequent analyses of the imputed data), the SMUB approach would not be appropriate. One can compute adjusted estimates of means based on the SMUB approach using R software available in the public domain (*site to be inserted*).

**3.3. Competing Approach 3: Doubly Robust Inference for Non-Probability Samples**

As a third alternative approach, we consider recently developed doubly robust (DR) approaches for estimating finite population means based on non-probability samples (Chen et al. 2020). These approaches generate inference based on two models: one for the probability of selection into the (non-probability) sample, and one for making predictions on the variable of interest for cases in the probability sample. An attractive feature of these DR methods is the property that either model can be misspecified (but not both) and the resulting estimator of the finite population mean will still be consistent. The DR approaches assume that the prediction model fitted to the selected sample holds for the entire finite population, i.e., assume transportability. The DR approaches are implemented in the *nonprobsvy* contributed package for the R software (Chrostowski et al. 2025).

We also consider estimation approaches based on inverse probability/propensity weighting (IPW), also known as quasi-randomization, where the probabilities of inclusion in the non-probability sample are estimated based on a data set combining both the non-probability sample and a (weighted) reference probability sample, and one then computes weighted estimates based on “pseudo” weights, which are computed as the inverse of these estimated probabilities (in addition to some form of replicated variance estimation). These IPW approaches, which are also implemented in the *nonprobsvy* package, are not doubly robust, in that they do not incorporate a prediction model for the probability sample. See Elliott and Valliant (2017) for details.

**3.4. Competing Approach 4: Use of “Shadow Variables”**

So-called “shadow variable” approaches (Miao et al. 2024; see also Zhao and Ma 2022, Tang and Ju 2018, Shao and Wang 2016, and Wang et al. 2014) rely on an auxiliary variable that is assumed to be correlated with the outcome of interest *Y* but independent of the selection mechanism when one conditions on *Y* and a fully-observed set of covariates. When a shadow variable is available, one can identify efficient estimators for the parameter(s) of interest in a full data set (where *Y* is not observed for selected cases, e.g., the probability sample) using nonparametric methods (Miao et al. 2024).

We note that the shadow variable approach is effectively a special case of our proposed PPMM-Reg approach. If we have only a single *A* variable (the auxiliary variable needed for the PPMM-Reg approach), then using *A* as a shadow variable will provide results similar to those of the PPMM-Reg approach with = 1. Both methods will produce unbiased estimates if *A* is truly a shadow variable (i.e., independent of selection, conditional on *Y* and *Z*). However, if *A* *is* associated with selection even after conditioning on *Y* and *Z*, then the shadow variable approach will produce biased estimates. In this case, using the PPMM-Reg approach and putting a Uniform prior on will reduce bias relative to the shadow variable approach (or fixing = 1). We will demonstrate these properties via empirical simulation studies in Section 4.

One limitation of the shadow variable approach is that it is limited to a *single* shadow variable, whereas the PPMM-Reg approach can accommodate multiple *A* variables. In addition, the shadow variable must have larger support than the outcome variable. This may be particularly limiting when using probability samples as the reference sample for mass imputation, where the majority of variables are likely to be categorical. Thus, finding a shadow variable for a continuous *Y* variable may be challenging or even impossible in some situations.

**4. Simulation Studies**

**4.1. Simulation 1**

***Setup.*** For our first simulation study, we generate three fixed populations of size *N* = 100,000 under three different scenarios. Following our earlier notation, let *Y* be the outcome variable of interest, let *Z*1 and *Z*2 be two predictor variables of interest in a target linear regression model, and let *A* be an auxiliary variable. Each of the three finite populations of size 100,000 are generated from the following superpopulation model:

(6)

Based on the results in West et al. (2021), we fix *,* , and to be 0.4 (given that these parameters did not significantly affect the performance of the proposed methodology), and we allow the correlation of *Y* and *A* given *Z*1 and *Z*2 be either 0.2, 0.5, or 0.8 (which determines the value of the covariance ). These three values reflect differing levels of strength of the auxiliary proxy and define the three different scenarios used to generate the three finite populations.

Given the simulation of a single finite population of size 100,000 under one of these three scenarios, we then select a probability sample of size *n*A = 5,000 using simple random sampling (Sample A) and an independent non-probability sample of size *n*B = 500 (Sample B). These relative sample sizes were chosen to mimic the common situation where measurement of *Y* is not not possible in the large probability sample but can be done in a small non-probability sample. Following our earlier notation, we let *S* indicate selection into the non-probability sample (vs. the probability sample). The probability of being selected into the non-probability sample (Sample B) is determined by the following logit model:

(7)

In this model, we fix and to be equal to ln(1.1), we allow the values of to be either 0, ln(1.1), or ln(2) (defining the extent of the non-ignorable selection), and we allow to be either 0, ln(1.1), or ln(2). We set in (7) to the value that results in the desired NP sample size of nB = 500.

The chosen values for and define specific selection mechanisms of interest. If is 0, then selection into the non-probability sample is ignorable, whereas if is greater than 0, then selection into the non-probability sample is non-ignorable. If = 0, then *A* is a shadow variable, since , and when > 0, *A* is not a shadow variable (i.e., is associated with selection even after conditioning on *Y* and *Z*).

Overall, this results in nine possible selection mechanisms that will be applied to each of the three fixed finite populations. Collectively, there are therefore 27 simulation scenarios defined by the combinations of finite population data generation mechanism (3 scenarios) and selection mechanism for the non-probability sample (9 scenarios). We focus on four sets of scenarios:

1. *ignorable selection with a shadow variable* (), in which case the PPMM-Reg approach and the SMUB approach with , the methods of Chen et al. (2020), and the shadow variable approach should produce unbiased estimates (3 scenarios);
2. *ignorable selection without a shadow variable* (), in which case the PPMM-Reg approach and the SMUB approach with and the methods of Chen et al. (2020) should produce unbiased estimates (6 scenarios);
3. *non-ignorable selection with a shadow variable* (), in which case the PPMM-Reg approach with and the shadow variable approach should produce unbiased estimates (6 scenarios); and
4. *non-ignorable selection without a shadow variable* (), in which case no method will produce completely unbiased estimates, but the PPMM-Reg and SMUB approaches should provide decent coverage under a Uniform prior for (12 scenarios).

For each of the three fixed finite populations, we repeat the sampling process 1,000 times under each of the nine selection mechanisms. For each of these 1,000 “paired” non-probability and probability samples, we then compute eight possible estimates of the finite population mean of *Y* (described below).

***Estimation Methods.*** We evaluate the performance of eight possible estimation methods based on the 1,000 pairs of samples selected under each scenario:

1) the sample mean of *Y* from the probability sample (treated as a benchmark), assuming simple random sampling for variance estimation;

2) the sample mean of *Y* from the non-probability sample (considered the naïve estimator), assuming simple random sampling for variance estimation;

3) estimation of the mean of *Y* based on the mass imputation approach (implemented in *nonprobsvy*), using *Z*1, *Z*2, and *A* as predictors, assuming transportability, and using linearization for variance estimation;

4) estimation of the mean of *Y* based on the doubly robust method of Chen et al. (2020), which also assumes transportability, using *Z*1, *Z*2, and *A* in all models and linearization for variance estimation (implemented in *nonprobsvy*),

5) weighted estimation of the mean of *Y* based on a simple IPW method for the non-probability sample, using the estimated probability of selection into the non-probability sample as a function of *Z*1, *Z*2, and *A* based on a combination of the two samples (implemented in *nonprobsvy*), with linearization for variance estimation;

6) use of the SMUB approach from Little et al. (2020) to compute an adjusted mean, following a Bayesian approach;

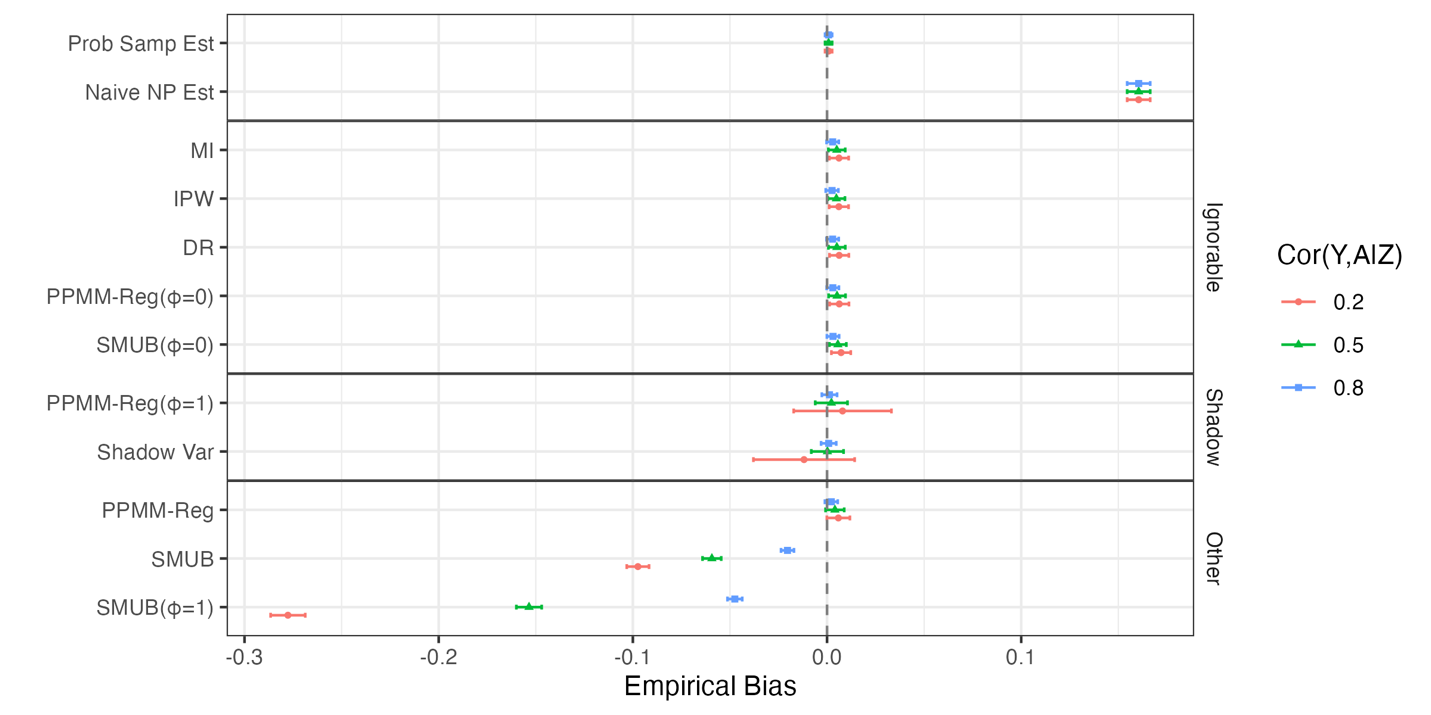
7) use of the proposed PPMM-Reg methodology (with *A* used to define the auxiliary proxy), where the imputation model focuses on prediction of *Y* using *Z*1 and *Z*2, following a Bayesian approach; and

8) estimation of the mean of *Y* based on the doubly robust shadow variable method (using an analytic estimate of the variance).

For each of these eight estimation methods that we apply to the 1,000 paired samples selected under a given scenario, we evaluate the empirical bias of the estimator (using the mean of a given fixed finite population as the true value), the coverage of 95% confidence or credible intervals based on each estimator, and the mean width of the intervals.

***Results.*** We summarize the results of our first simulation study by the four sets of scenarios outlined above. First, for set 1 (*ignorable selection with a shadow variable*), most of the methods performed well, as expected. Notable deviations in terms of empirical bias (Figure 1 below) are observed for the SMUB approach based on the PPMM, which allows for draws of . Notably, the PPMM-Reg approach does not suffer from this same limitation when taking Uniform(0,1) draws of the parameter. The empirical interval coverage rates were found to be largely nominal (close to 0.95), with the exception of the SMUB approach with , which also produced a larger median CI width; the SMUB approach coverage also decreased when the strength of the auxiliary proxy decreased (see Figures 1A and 1B in the supplemental materials). The IPW approach from Chen et al. (2020) was found to have higher than nominal coverage (1.0) and the highest median CI width. As expected, CI width increased in general for all of the estimators as the strength of the auxiliary proxy decreased.

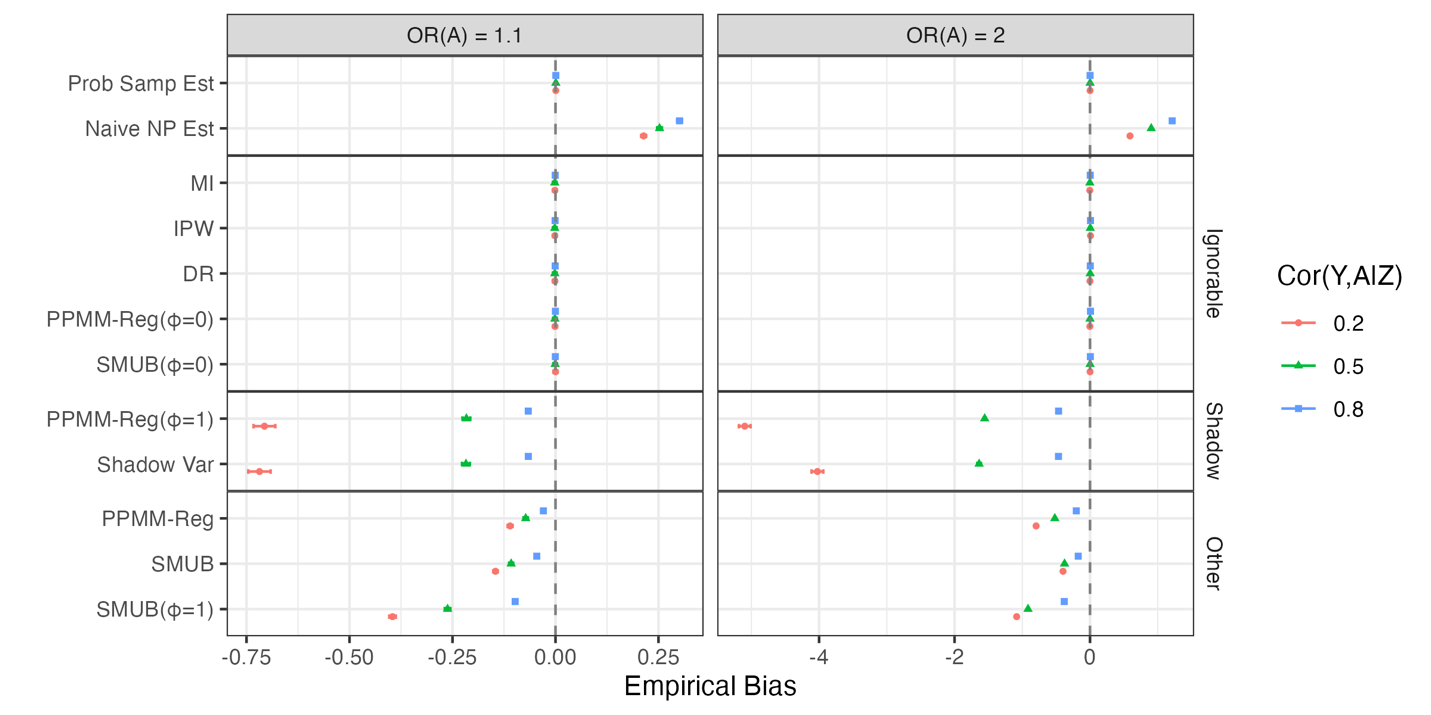
**Figure 1.** Empirical bias of the estimators under ignorable selection with a shadow variable. Estimators are grouped by their underlying assumptions (ignorability, existence of a shadow variable, and other estimators). The top two rows are the gold standard estimate that would be obtained if *Y* were observed in the probability sample and the naïve estimate based on the nonprobability sample.



Second, for set 2 (*ignorable selection without a shadow variable*), all methods that assume ignorability/transportability produce unbiased estimates as expected, including the PPMM-Reg and SMUB approaches with as well as the MI, IPW, and DR methods (see Figure 2). Both the PPMM-Reg approach with and the shadow variable approach are biased and follow a similar pattern, with bias increasing as the dependence of selection into the non-probability sample on the auxiliary proxy *A* becomes stronger, as expected. Both the PPMM-Reg approach and the SMUB approach with drawn from a Uniform(0,1) distribution essentially average over a set of selection mechanisms, with a completely nonignorable mechanism as one bound, and thus they also produce biased estimates. However, the magnitude of this bias is much smaller than when or for the shadow variable approach.

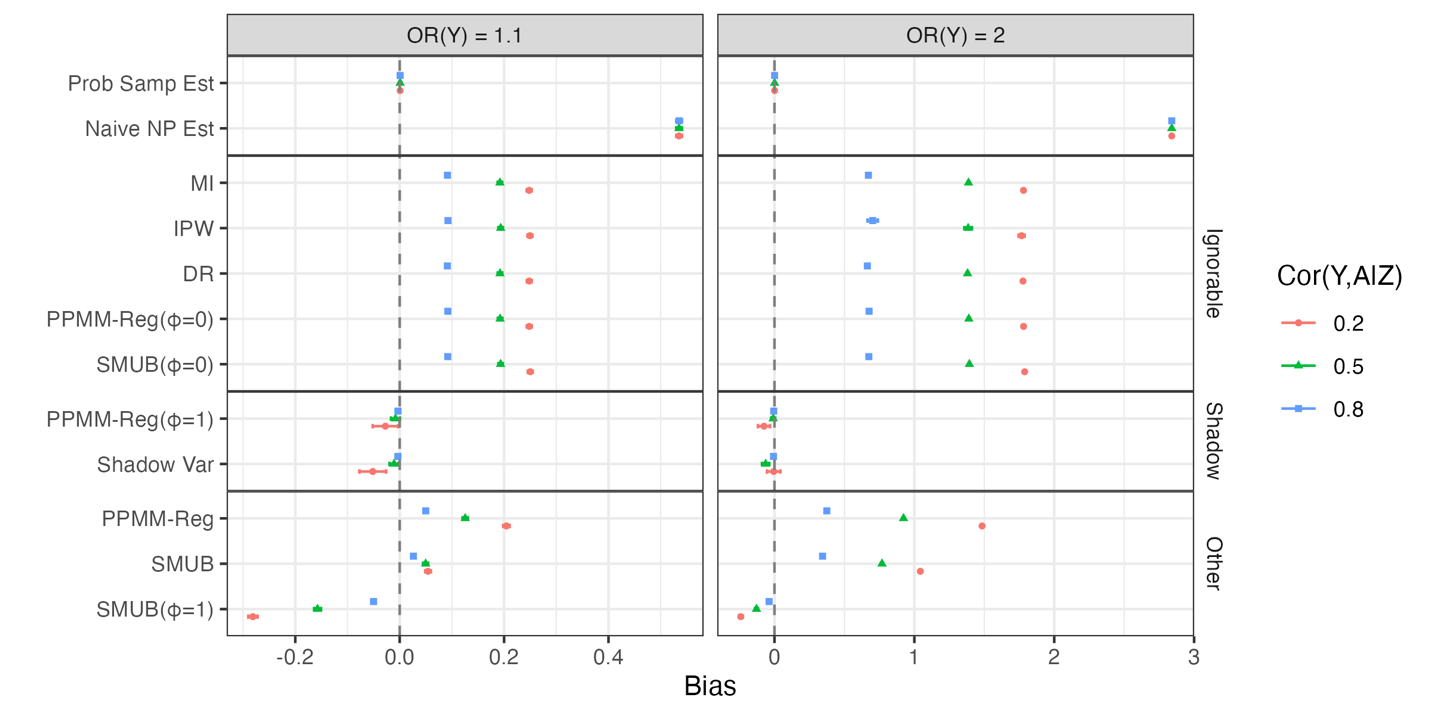
In terms of empirical coverage and median interval width (see Figures 2A and 2B in the supplemental materials), despite producing biased estimates, the PPMM-Reg approach produces nominal coverage when the relationship of *A* with selection is weak, driven by very large credible intervals. All other methods that produce biased estimates have lower than nominal coverage, as expected. Among the methods that assume ignorability (and thus produce unbiased estimates), most have approximately nominal coverage. The exceptions are the IPW approach, which yields consistently large median widths and over-coverage, and the DR approach, which shows some over-coverage when the relationship of selection with *A* is strong.

**Figure 2.** Empirical bias of the estimators under ignorable selection without a shadow variable (the two columns represent the two non-zero associations of the auxiliary proxy *A* with selection). Estimators are grouped by their underlying assumptions (ignorability, existence of a shadow variable, and other estimators). The top two rows are the gold standard estimate that would be obtained if *Y* were observed in the probability sample and the naïve estimate based on the nonprobability sample.



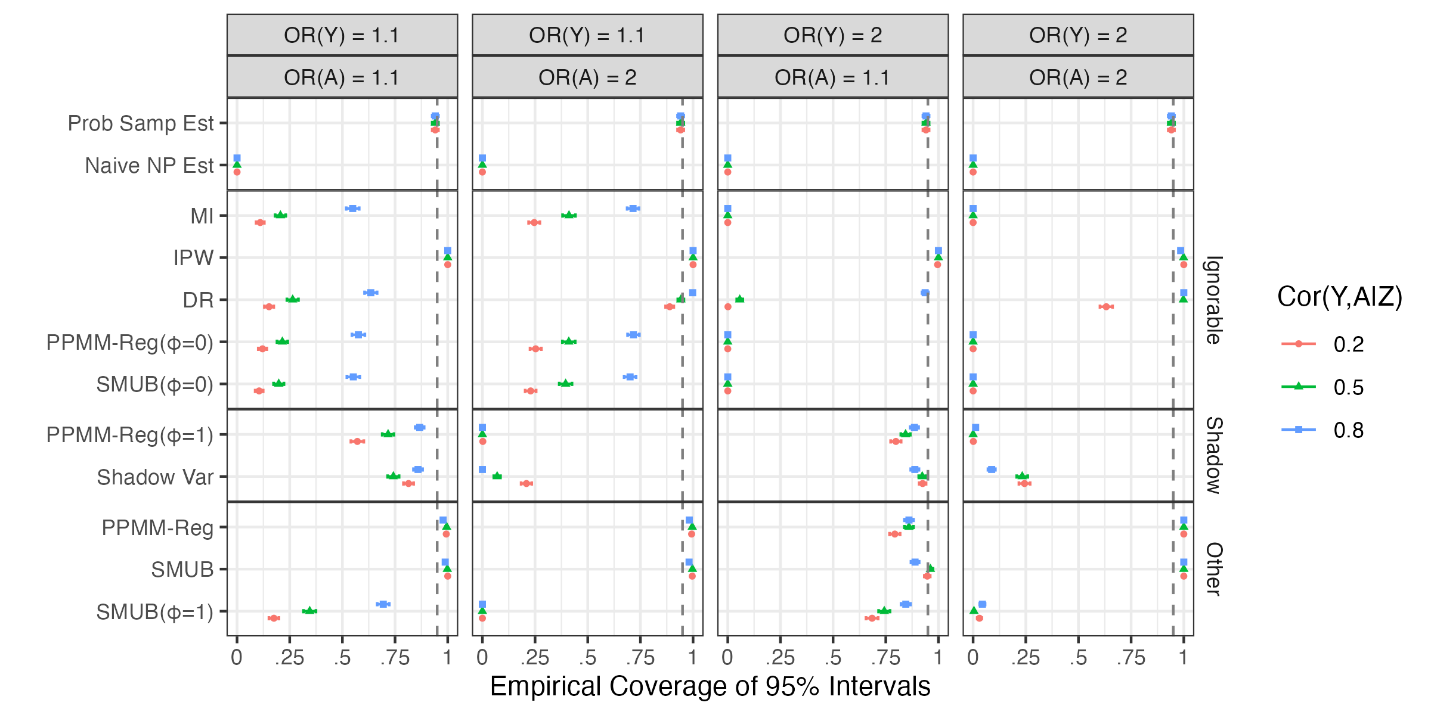
Third, for set 3 (*non-ignorable selection with a shadow variable*), only the PPMM-Reg method with set to 1 and the shadow variable approach produce unbiased estimates regardless of the scenario (see Figure 3). Methods that assume ignorability/transportability perform poorly and similarly to each other across scenarios, as expected. The SMUB-based method that uses a Uniform(0,1) prior produced biased estimates, but less so than the methods assuming ignorability. In terms of empirical coverage and width (see Figures 3A and 3B in the supplemental materials), we find a similar pattern: only the PPMM-Reg method with set to 1 and the shadow variable approach produce coverage close to nominal, and all other methods result in poor coverage (with the IPW method continuing to result in over-coverage despite substantial bias in point estimates). Patterns seen with the empirical interval width results are similar to the first two sets of scenarios.

**Figure 3.** Empirical bias of the estimators under non-ignorable selection with a shadow variable (the two columns are the non-zero association of the outcome variable *Y* with selection). Estimators are grouped by their underlying assumptions (ignorability, existence of a shadow variable, and other estimators). The top two rows are the gold standard estimate that would be obtained if *Y* were observed in the probability sample and the naïve estimate based on the nonprobability sample.



Finally, for set 4 (*non-ignorable selection without a shadow variable*), no method is expected to produce unbiased estimates, as the selection mechanism is non-ignorable but there is no shadow variable available. However, we expect the PPMM-Reg and SMUB methods (with a Uniform(0,1) prior) to perform relatively well, since the selection mechanism corresponds to a value of somewhere between 0 and 1, as selection depends in part on observed covariates and in part on the outcome *Y*. Thus, these methods that average across values should cover the true mean. Figure 4 shows the empirical coverage of the methods for combinations of the strength of associations of *A* and *Y* with selection (empirical bias is shown in Figure 4A in the supplemental materials). Both the PPMM-Reg and SMUB methods consistently achieve closest to nominal coverage. Methods that assume ignorability have low coverage; the exception is the IPW approach (and for some scenarios the DR approach), which show over-coverage once again due to extremely large interval widths (see Figure 4B in the supplemental materials). Methods that assume that *A* is a shadow variable exhibit severe undercoverage when the association of *A* with selection is strong (i.e., when the assumption that *A* is a shadow variable is severely violated); coverage of these methods is closer to nominal but still below when the association of *A* with selection is weak.

**Figure 4.** Empirical coverage of the estimators under non-ignorable selection without a shadow variable (the four columns are the four different selection mechanisms). Estimators are grouped by their underlying assumptions (ignorability, existence of a shadow variable, and other estimators). The top two rows are the gold standard estimate that would be obtained if *Y* were observed in the probability sample and the naïve estimate based on the nonprobability sample.



Overall, we draw the following conclusions from this first simulation study:

* If one is willing to assume an ignorable selection mechanism given *Y*, *Z*, and *A*, and this reflects the truth, then the methods discussed in Chen et al. (2020), including mass imputation assuming transportability (Kim et al. 2021), tend to have the best performance. Specifically, the doubly robust and mass imputation approaches appear to be optimal choices (given the over-coverage and large interval width associated with the IPW approach). The PPMM-based approaches (including PPMM-Reg and SMUB) with set to 0 perform similarly well (as expected).
* Under non-ignorable selection mechanisms, whether or not *A* is a shadow variable is of critical importance. The PPMM-based methods with set to 1 or the shadow variable approach (which are essentially applying the same adjustments) perform well if *A* is in fact a shadow variable (see the Discussion for additional remarks on this issue). If no shadow variable is available, the PPMM-based approaches using random Uniform(0,1) draws of appear to have the best performance across the scenarios. Regardless of the presence of a shadow variable, mass imputation assuming transportability consistently has poor performance in these scenarios.
* In terms of interval width and coverage, having stronger auxiliary proxies is generally always important for the PPMM-based methods. The SMUB approach tends to have slightly better performance than the PPMM-Reg approach, possibly due to the absence of uncertainty associated with the imputed values. As the SMUB approach does not produce imputed values in the probability sample, an application that seeks to perform additional analyses using the imputed values (beyond just estimating the finite population mean) should consider the PPMM-Reg approach.

**4.2. Simulation 2**

Next, we compare the performance of these competing approaches using a simulation study based on real data. This study draws on the example from the National Survey of Family Growth (NSFG) discussed in West et al. (2021). We extracted data from four years of the NSFG, between September 2012 and August 2016. During this period, NSFG respondents were asked whether they owned a smartphone. We treated the full probability sample of NSFG respondents from this time period as the hypothetical larger finite “population” (*N* = 19,800), and smartphone owners with less than high school education as a hypothetical selected non-probability sample from this larger population (*n* = 2,977).

The finite population parameter of interest in this simulation study was the mean number of months worked in the past year for this hypothetical population (true mean = 7.61). We assume that microdata exist for both the selected cases (assumed to be a non-probability sample) and the non-selected cases (the remainder of the population, rather than a probability sample) on two covariates of interest (the *Z* variables) that are theoretically strong predictors of the number of months worked in the past year: sex (male/female) and age (15-18, 19-29, 30-49). We also assume that there are microdata available for the selected sample and summary statistics available for the non-selected cases on additional auxiliary variables *A* that may be strong predictors of the number of months worked: race/ethnicity (non-Hispanic White, non-Hispanic Black, Hispanic, Other), marital status (married, divorced/widowed/separated), household income (<$19,999, $20,000-$59,999, $60,000+), region of the United States where housing unit is located (Midwest, Northeast, South, West), current employment status (working / not working), and presence of children under the age of 16 in the household (yes / no).

***Imputation Approach.*** The imputation model of interest to be used for predicting the number of months worked for all non-selected cases indexed by *i* in the finite population is thus

 (9)

We seek the best possible estimates of the regression coefficients and the residual variance in (9) to use when making predictions for the non-selected cases. This simulation study is therefore more of an empirical evaluation of how well these methods do when following a “superpopulation modeling” approach to finite population inference, per Elliott and Valliant (2017). These approaches involve making predictions for all non-selected cases in a finite population based on a model fitted to a selected sample.

We considered a “standard” multiple imputation approach for this particular simulation study, given that we aren’t working with a selected non-probability sample and a separate probability sample, but rather imputing missing values for all other non-selected cases in the hypothetical population based on a selected sample of that population. Within this context, we evaluated the performance of 1) a standard multiple imputation approach, assuming that the imputation model fitted to the selected sample is “transportable” and applies to all other non-selected cases; 2) the PPMM-Reg approach, which as discussed earlier relaxes the transportability assumption; 3) the SMUB approach, which as a reminder does not require *any* microdata for non-selected cases; and 4) the doubly robust approach for estimation of means based on selected non-probability samples, which unlike the PPMM-Reg and SMUB approaches requires the full NSFG data set with microdata on both the *Z and* *A* variables for both selected and non-selected cases. For this second simulation, we could not evaluate the shadow variable approach, as none of the predictors considered had larger support than the outcome variable.

***Results.*** We first consider the ability of the MUB adjustment method to recover the regression coefficients of interest for the hypothetical NSFG population. Table 1 below presents the estimated coefficients based on the selected sample (those with a smartphone and less than high school education), the true coefficients for the full population, and the adjusted coefficients based on the MUB adjustment (including 95% credible intervals).

**Table 1.** Adjustment of estimated coefficients in prediction model based on the MUB approach.

|  |  |  |  |
| --- | --- | --- | --- |
| Coefficient | Selected Sample  (Estimate, SE) | Full Population | MUB Adjustment (Posterior Median, 95% Credible Interval) |
| Intercept | 1.06 (0.13) | 2.09 | 2.47 (1.95, 3.25) |
| Male | 1.34 (0.16) | 1.02 | 0.82 (0.30, 1.13) |
| Age 19-29 | 5.33 (0.20) | 5.64 | 5.53 (5.03, 6.08) |
| Age 30-49 | 5.75 (0.18) | 6.44 | 5.99 (5.54, 6.53) |

We see that the estimated coefficients based on the selected sample are substantially biased, especially the intercept and the coefficients for male and age 30-49. The MUB adjustment approach shifts the estimated coefficients in the correct direction and all 95% credible intervals cover the true population coefficients. We would therefore expect improved imputations among the non-selected cases based on the adjusted coefficients relative to assuming ignorability and imputing using the estimates from the selected sample.

**Table 2.** Results of NSFG-Based Simulation Study.

|  |  |  |  |
| --- | --- | --- | --- |
| Method  (Assumed Selection Mechanism) | Posterior Mean / Estimate  (Truth = 7.61) | 95% Credible / Confidence Interval | Relative Bias (%) |
| Naïve (SRS Estimation) | 4.41 | (4.23, 4.59) | -42.05% |
| Multiple Imputation (MAR) | 7.11 | (6.88, 7.33) | -6.57% |
| **PPMM-Reg Approach (MNAR)** | **7.51** | **(7.07, 8.04)** | **-1.31%** |
| SMUB Adjustment (MNAR) | 7.82 | (7.09, 8.59) | 2.76% |
| Doubly Robust (MAR) | 7.80 | (7.58, 8.03) | 2.50% |

Estimates of the mean number of months worked using different modeling approaches are shown in Table 2. First, we note the significant bias that would arise from attempting to estimate the population mean based on the selected sample only. All competing approaches do a reasonable job of shifting the mean in the direction of the population truth, with the PPMM-Reg producing the lowest relative bias in the estimate. In terms of efficiency, the doubly robust approach appears to be most efficient, but we reiterate that this approach is leveraging microdata on all *Z* and *A* variables for both the selected and non-selected samples, which would not be a common scenario in practice. Of the approaches that do not require microdata on *Z* and *A* for the non-selected cases, the PPMM-Reg appears to be more efficient than the competing SMUB adjustment.

**5. Case Study**

We now consider a case study using real survey data from the U.S. Census Bureau’s 2023 Household Pulse Survey (HPS), which is the selected non-probability sample (*n* = 62,505), and the 2023 National Health Interview Survey (NHIS), which is the selected probability sample (*n* = 29,552) assumed to arise from the same population. See U.S. Census Bureau (2025) for details about the HPS, and National Center for Health Statistics (2023) for details regarding the NHIS. The data sets were downloaded from https://www.census.gov/programs-surveys/household-pulse-survey/data/datasets.html and https://www.cdc.gov/nchs/nhis/documentation/2023-nhis.html, respectively.

In this application, we consider the outcome variable *Y* to be a measure of depression based on two items (adapted from the PHQ-2; Kroenke et al. 2003), the *Z* variables of interest to be a harmonized set of socio-demographic variables [including sex, education (less than high school, high school, some college, college degree, graduate degree), race/ethnicity (Hispanic, Non-Hispanic White, Non-Hispanic Black, Non-Hispanic Asian, and Other), age (18-25, 26-35, 36-45, 46-55, 56-65, 66-75, and 76+), and region of the U.S. (Northeast, South, Midwest, and West)], and the *A* variable to be a measure of anxiety based on two items (adapted from the GAD-2; Kroenke et al. 2007). The objective is to produce a national estimate of the mean on the two-item depression scale. We assume that depression is not available in the NHIS; the actual presence of the depression measure (PHQ-2) in the NHIS allows us to evaluate the performance of the alternative methods against the true (weighted) estimate based on the NHIS.

We evaluated eight estimators given these data: the naïve unweighted estimate of the mean based on the HPS alone (assuming simple random sampling), a weighted estimate based on the weights computed by the U.S. Census for the HPS (with a linearized variance estimate), the PPMM-Reg approach, the SMUB-based approach, mass imputation assuming transportability, IPW assuming ignorable selection, and the doubly robust method assuming transportability. We also implement the PPMM-Reg approach with the parameter set to 1 to evaluate the performance of the shadow variable approach in this context, under the assumption that *A* is in fact a shadow variable. We do not evaluate the shadow variable approach directly, as it is not entirely clear how to implement the approach when the probability sample arises from a complex sample design with survey weights involved. We see this as a good opportunity for future research in this area.

Table 3 presents estimates of the mean depression score and corresponding confidence / credible intervals for the mean, beginning with the “true” mean according to the weighted NHIS sample. The naïve estimate (unweighted estimate from the HPS) substantially overestimates the mean PHQ-2 score, and interestingly, the weights provided with the HPS move the estimate in the *wrong* *direction*. The methods assuming ignorable selection shift the estimate in the correct direction, but do not repair all of the bias due to selection. In contrast, the PPMM-Reg approach (using random Uniform(0,1) draws of the parameter) produces essentially the same estimate of the mean as using the PHQ score from the weighted NHIS data. The SMUB-based approach produces a similar estimate to the PPMM-Reg approach with just slightly more bias relative to the weighted NHIS mean.

**Table 3.** Case study results.

|  |  |
| --- | --- |
| **Estimator** | **Estimated Mean**  **(95% CI / CrI)** |
| Weighted NHIS Mean (Truth) | 0.56 |
| Naïve SRS Estimate, HPS | 1.17 (1.16-1.18) |
| Weighted Estimate, HPS | 1.31 (1.28-1.34) |
| Mass Imputation Assuming Transportability | 0.74 (0.72-0.76) |
| IPW Assuming Ignorable Selection | 0.74 (0.70-0.78) |
| Doubly Robust Estimation | 0.74 (\*\*\*) |
| PPMM-Reg Approach | 0.54 (0.31-0.73) |
| PPMM-Reg(1) Approach (Shadow Variable) | 0.29 (0.26-0.32) |
| SMUB Approach | 0.60 (0.44-0.74) |

\*\*\* We were unable to produce reasonable standard errors using the *nonprobsvy* package to implement the DR approach in this application. We only report the point estimate for comparison.

Finally, Table 3 shows that the PPMM-Reg approach with the parameter set to 1 results in an over-adjustment, as expected (given that it assumes extreme non-ignorability); it would not appear that *A* has the desirable properties of a shadow variable in this application.

**6. Discussion**

In this paper, we evaluate the performance of a proposed mass imputation approach that integrates probability and non-probability samples and adjusts the estimated coefficients in an imputation model based on the non-probability sample for a potential non-ignorable selection mechanism associated with that sample. Via two empirical simulation studies that consider a variety of selection mechanisms and a case study using real survey data, we find general support for the proposed approach vs. competing methods in the literature, including the standardized measure of unadjusted bias, mass imputation assuming transportability, doubly robust estimation methods, IPW methods, and shadow variable approaches.

We provide remarks on some of the more interesting results:

* The shadow variable methods can be thought of a special case of the proposed PPMM-Reg methodology, where the sensitivity parameter is set to 1 (representing maximum non--ignorability). If selection does not depend on the auxiliary proxy / shadow variable *A* when conditioning on the outcome of interest *Y* and the available covariates *Z* (i.e., *A* is truly a shadow variable), then both of these methods will perform well. However, this assumes that only a single shadow variable *A* with larger support than *Y* exists, with no tests available for determining whether *A* is a shadow variable. The PPMM-Reg approach can accommodate multiple *A* variables with varying support and has better performance (when allowing for a range of values of the sensitivity parameter between 0 and 1) if selection actually depends on *A* when conditioning on *Y* and *Z*.
* The SMUB method also tends to perform well in a variety of scenarios, but does not produce imputations of the variable of interest in the probability sample. This limits studies with larger analytic objectives that would use the imputed data in variety of ways.
* The doubly robust and mass imputation methods tend to perform well under ignorable selection mechanisms, and the doubly robust methods also perform well in the case of non-ignorable selection without *A* being a shadow variable, when the dependence of selection on *Y* is weaker. If a shadow variable exists, the performance of the doubly robust methods becomes worse, and the performance also worsens as dependence on *Y* increases.

In practice, we would recommend implementing the methods evaluated here and comparing the inferences that would result in light of the assumptions that the different methods are making. Our results provide general support for the use of the PPMM-Reg methodology, given that it produces imputations in the probability sample that would enable a variety of subsequent analyses. The PPMM-Reg approach has been implemented in R using code that is freely available in a GitHub repository (*site to be inserted*).

There are multiple possible directions for future work in this area. First, the PPMM-Reg approach assumes that the data are ultimately governed by a bivariate normal pattern mixture model for *Y* and the auxiliary proxy defined by the *A* variable(s). While emerging studies suggest that this approach is fairly robust to violations of this assumption (Gomez et al. 2023), developments are needed to improve imputation models for a broader class of *Y* variables (categorical, semi-continuous, count, etc.). West et al. (2021) describe adjustments that could be applied to the coefficients of a probit model used to impute a binary variable, but similar adjustment approaches are needed for other types of *Y* variables. We also assume that the model used to define the auxiliary proxy has been well-specified, and additional work should consider the implications of misspecification of this model when it is fitted to the non-probability sample. Finally, we call for additional applications of this approach in a variety of mass imputation problems, enabling assessments of the performance of the PPMM-Reg approach in a variety of contexts outside of those considered here.

**References**

Andridge, R. R., & Little, R. J. (2011). Proxy pattern-mixture analysis for survey nonresponse. *Journal of Official Statistics*, *27*(2), 153-180.

Andridge, R. R., West, B. T., Little, R. J., Boonstra, P. S., & Alvarado-Leiton, F. (2019). Indices of non-ignorable selection bias for proportions estimated from non-probability samples. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *68*(5), 1465-1483.

Chen, S., Yang, S., and Kim, J.K. (2022). Nonparametric Mass Imputation for Data Integration. *Journal of Survey Statistics and Methodology*, 10(1), 1-24.

Chen, Y., Li, P., & Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, *115*(532), 2011-2021.

Choi, S., Nandram, B., & Kim, D. (2021). Bayesian predictive inference of small area proportions under selection bias. *Survey Methodology*, *47*(1), 91-123.

Chrostowski, Ł., Chlebicki, P., & Beręsewicz, M. (2025). *nonprobsvy* – An R package for modern methods for non-probability surveys. arXiv preprint [arXiv:2504.04255](https://arxiv.org/abs/2504.04255).

Elliott, M. R., & Valliant, R. (2017). Inference for Nonprobability Samples. *Statistical Science*, *32*(2), 249.

Gomez, S., Pavlopoulos, D., Stoel, R., de Waal, T., and van Delden, A. (2023). The Sensitivity of Selection Bias Estimators: A Diagnostic based on a case study and simulation. *Paper presented at the European Survey Research Association Conference, July 17, 2023, Milan, Italy.*

Kim, J.K., Park, S., Chen, Y., and Wu, C. (2021). Combining Non-probability and Probability Survey Samples Through Mass Imputation. *Journal of the Royal Statistical Society (Series A)*, 184, 941-963.

Kroenke, K., Spitzer, R.L., Williams, J.B. (2003). The Patient Health Questionnaire-2: validity of a two-item depression screener. *Medical Care*, Nov;41(11):1284-92.

Kroenke, K., Spitzer, R.L., Williams, J.B., Monahan, P.O., Löwe B. (2007). Anxiety disorders in primary care: prevalence, impairment, comorbidity, and detection. Annals of Internal Medicine, 2007 Mar 6;146(5):317-25.

Little, R. J. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika*, *81*(3), 471-483.

Little, R. J., & Rubin, D. B. (2019). *Statistical Analysis with Missing Data, Third Edition*. John Wiley & Sons.

Little, R. J., West, B. T., Boonstra, P. S., & Hu, J. (2020). Measures of the degree of departure from ignorable sample selection. *Journal of survey statistics and methodology*, *8*(5), 932-964.

Ma, J., Sedransk, J., Nandram, B., and Chen, L. (2018). Bayesian predictive inference for finite population quantities under informative sampling. arXiv preprint arXiv:1804.03122 - 2018

arxiv.org.

Miao, W., Liu, L., Li, Y., Tchetgen Tchetgen, E. J., & Geng, Z. (2024). Identification and semiparametric efficiency theory of nonignorable missing data with a shadow variable. *ACM/JMS Journal of Data Science*, *1*(2), 1-23.

Nandram, B., Bhatta, D., Bhadra, D., & Shen, G. (2013). Bayesian predictive inference of a finite population proportion under selection bias. *Statistical Methodology*, *11*, 1-21.

National Center for Health Statistics (2023). National Health Interview Survey, 2023. Public-use data file and documentation. https://www.cdc.gov/nchs/nhis/documentation/2023-nhis.html. Date accessed: July 29, 2025.

Pfeffermann, D. (2017). Bayes-based non-bayesian inference on finite populations from non-representative samples: A unified approach. *Calcutta Statistical Association Bulletin*, *69*(1), 35-63.

Royall, R. M., & Pfeffermann, D. (1982). Balanced samples and robust Bayesian inference in finite population sampling. *Biometrika*, *69*(2), 401-409.

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

Shao J., & Wang L. (2016). Semiparametric inverse propensity weighting for nonignorable missing data. *Biometrika*, 103(1), 175–187.

Tang N., & Ju Y. (2018). Statistical inference for nonignorable missing data problems: A selective review. *Statistical Theory and Related Fields*, 2(2), 105–133.

U.S. Census Bureau (2025). Household Pulse Survey technical documentation. https://www.census.gov/programs-surveys/household-pulse-survey/technical-documentation.html. Date accessed July 29, 2025.

Wang S., Shao J., & Kim J. K. (2014). An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica*, 24, 1097–1116.

West, B. T., Little, R. J., Andridge, R. R., Boonstra, P. S., Ware, E. B., Pandit, A., & Alvarado-Leiton, F. (2021). Assessing selection bias in regression coefficients estimated from nonprobability samples with applications to genetics and demographic surveys. *The Annals of Applied Statistics*, *15*(3), 1556.

Wiśniowski, A., Sakshaug, J. W., Perez Ruiz, D. A., & Blom, A. G. (2020). Integrating probability and nonprobability samples for survey inference. *Journal of Survey Statistics and Methodology*, *8*(1), 120-147.

Xu, Z., Nandram, B., & Manandhar, B. (2020). Bayesian inference of a finite population mean under length-biased sampling. *Statistical Methods and Applications in Forestry and Environmental Sciences*, 79-103.

Zhao J., & Ma Y. (2022). A versatile estimation procedure without estimating the nonignorable missingness mechanism. *Journal of the American Statistical Association*, 117(540), 1916–1930.