Imputation Classes as a Framework for Inferences From Non-random Samples.

Vladislav Beresovsky[*]*

[*]National Center for Health Statistics, 3311 Toledo Rd, Hyattsville, MD 20782

Abstract

The recent tendency of growing cost and nonresponse of traditional randomized surveys and rapid proliferation of web surveys and administrative data calls for developing a standard framework for inferences from nonrandom data samples. Approaches relying either on a propensity score model or on a predictive model of an outcome variable are overly sensitive to model assumptions. This paper proposes to: (a) supplement an initial nonrandom sample with a reference random sample, having missing detail target variables but containing core covariates shared with the nonrandom sample; (b) define imputation classes using both propensity and prediction scores, and impute target variables from the nonrandom to the random sample; and (c) use a delete-a-group version of the adjusted jackknife variance estimator, proposed by Rao and Shao (1992) for imputed data. Since imputation classes are defined by both propensity and predictive models, the proposed framework exhibits double-robust property against misspecification of either model. Reference samples, complete with imputed data and jackknife replication weights, can be released to end-users as public use files, allowing for any kind of inferences. The proposed paradigm for inferences from nonrandom samples may legitimize their use in official statistics.

Key Words: nonrandom samples; propensity score; predictive model; imputation class; jackknife with missing data.

Introduction

Rapid proliferation of surveys with opt-in online panels offers an expedient and relatively inexpensive alternative to traditional surveys. The disadvantage is that opt-in samples are not selected at random, and so, may not be representative of the general population. The difficulties of producing population estimates from web survey data were reported by Chmura et al. (2013), DiSogra et al. (2011) and Dever et al. (2008).

The problem of bias correction of estimates from nonrandom samples has been previously addressed in relation to estimation with missing data and in observational studies. Rosenbaum and Rubin (1983) proposed to estimate the treatment effect in clinical trials conditional on propensity scores derived for both treated and control patients from modeling their probability to obtain a treatment. In the case of missing data, Kim and Kim (2007) proved that propensity score adjustment (PSA) is asymptotically unbiased and consistent if the response propensity model is correctly specified. This approach was discussed in the application to web samples by Valliant and Dever (2011) and Lee and Valliant (2009). Beresovsky (2016) gave a rigorous justification for using PSA in a context of nonrandom and reference samples.

Imputation of missing data using a prediction model for a target variable, fitted on respondents' data, was proposed by Little and Rubin (1987) and has been widely used in survey practice ever since. Rubin and Schenker (1986) demonstrated that proper accounting for

*vberesovsky@cdc.gov

The findings and conclusions in this paper are those of the authors and do not necessarily represent the views of the National Center for Health Statistics, Centers for Disease Control and Prevention.

additional variability of estimates associated with missing data requires multiple imputations by the means of the approximate Bayesian bootstrap.
Deville and Sarndal (1992) proposed a generalized calibration theory for the unified treatment of post-stratification, raking, and a generalized regression estimator; see also Sarndal (2007). Though it was initially proposed as a method to reduce the variance of the Horwitz-Thomson estimator in the case of complete data, Sarndal and Lundstrom (2005) and Kott (2006) demonstrated that it may be used to address missing data. Haziza and Lesage (2016) demonstrated that compared with one-step calibration, in some cases the two-step procedure, using PSA on the first step to model the nonresponse mechanism and generalized calibration on the second step, is more flexible and effective at eliminating potential bias due to model misspecification.
All the described methods use either a response or a prediction model for bias reduction. Bethlehem (1988), who considered post-stratification nonresponse adjustment for bias correction, realized that nonresponse bias adjustment becomes more robust when both models are employed. He concluded that "The stratification should be done in such a way that strata are homogeneous with respect to the target variable (thus decreasing the variance and bias) and with respect to the response probabilities (thus decreasing bias)." This idea was reiterated in a review paper by Brick (2013) and used in simulations by Leacy and Stuart (2014) to estimate treatment effect in observational studies.
This paper proposes the imputation classes framework, which allows to employ both response and prediction models for making inferences from a nonrandom sample. Very similar double robust imputation classes estimator was proposed for missing data problem by Haziza and Beaumont (2007), who used prediction and response models to justify the proposed estimator. This paper is different because it explicitly assumes imputation classes model, proposes estimator applicable for both missing data and nonrandom sample problems and proves unbiasedness of the adopted variance estimator. In Section 1, the inferential methodology, which is developed for the missing data problem, is systematically extended to any nonrandom sample problem, such as web samples. Expressions for bias and variance of point estimator under hot-deck imputation within designated classes explains its double robustness against either model misspecification and justifies application of the delete-a-group extension of the adjusted jackknife of Rao and Shao (1992) for variance estimation. In Section 2 these ideas are applied for estimating means, medians and their variances in simple simulations involving predetermined imputation classes. Section 3 describes a more sophisticated simulation study where imputation classes are defined using propensity and prediction models. Dramatic improvement in bias reduction is demonstrated when imputation classes are formed as the intersection of imputation classes defined by both models. In conclusion, this paper summarizes the benefits of employing imputation classes for inferences from nonrandom samples and outlines subjects for future research. *Note:* Due to limited size of a proceedings paper, some expressions are given without proofs. The author will be happy to supply them via email.

1. Inferences from nonrandom samples within imputation classes framework

1.1 From point estimation with missing data to estimation with nonrandom web samples. Bias and double robustness.

Consider estimation of a population mean from a simple random sample s_r of size n_r with some of the observations missing due to unit nonresponse. The proposed imputation classes estimator is based on the idea of dividing the sample by imputation classes A_ν and imputing missing values within each class as

$$\bar{y}_I = \frac{1}{n_r} \sum\nolimits_{\nu} \sum\nolimits_{i \in A_\nu} \left(\delta_{\nu i} y_{r\nu i} + (1 - \delta_{\nu i})\, y^*_{m\nu i} \right) \tag{1.1}$$

where y_{rvi}, y^*_{mvi} are responding and imputed units in class A_ν and $\delta_{\nu i}$ are response indicators. In principle, imputation classes may be defined by prediction and propensity models utilizing covariates X, with the goal to have residuals of both models to be randomly distributed with minimal variances within these classes

$$y_{r\nu i} = \mu_{r\nu} + \varepsilon_{r\nu i},\ \varepsilon_{r\nu i} \sim \left(0, \sigma^2_{r\nu}\right) \tag{1.2a}$$

$$y^*_{\nu m i} = \mu^*_{m\nu} + \varepsilon_{m\nu i},\ \varepsilon_{m\nu i} \sim \left(0, \sigma^2_{m\nu}\right) \tag{1.2b}$$

$$\delta_i \sim\ Bernoulli\left(p_{\nu i}\right),\ p_{\nu i} = p_\nu + \varepsilon_{p\nu i},\ \varepsilon_{\nu i} \sim \left(0, \sigma^2_{p\nu}\right) \tag{1.2c}$$

Under the imputation class model (1.2a-c) the bias of the estimator (1.1) conditional on the responders data y_{rvi} can be found as

$$\text{Bias}\left[\bar{y}_I | y_r\right] = \frac{1}{n_r} \sum\nolimits_{\nu} \sum\nolimits_{i \in \nu} E\left((1 - \delta_{\nu i}) y^*_{m\nu i}\right) - (1 - p_v)\, \mu^*_{m\nu} = -\frac{1}{n_r} \sum\nolimits_{\nu} n_{r\nu}\, \text{cov}\left(p_{\nu i}, y^*_{mi}\right) \tag{1.3}$$

where $n_{r\nu}$ is the size of an imputation class ν.

Bias due to correlation between stochastic residuals of the imputation and response models within imputation classes is similar to biases of post-stratification and calibration estimators obtained byBethlehem (1988), Sarndal and Lundstrom (2005) and Haziza and Lesage (2016). It can be eliminated, in principle, by optimal utilization of covariates X for proper selection of imputation classes, for which either one (or both) of the residuals $\varepsilon_{m\nu i}$, $\varepsilon_{p\nu i}$ in expressions (1.2b-c) are independent random variables. This is another way to say that estimator (1.1) is *double robust* against misspecification of either one of these models. However, expression for bias (1.3) replaces exact requirements of (1.2b-c) with less stringent quantifiable condition.

To extend the concept of the imputation classes estimator (1.1) to estimation from nonrandom samples such as those collected from web surveys, this paper follows Beresovsky (2016) and considers two samples: *reference* simple random sample s_r of size n_r with complete nonresponse and nonrandom sample s_w with all units responding. In the context of the combined sample $s = s_w \cup s_r$, units of s_w and s_r samples are treated as "respondents" and "nonrespondents". Using the same notation as above, the imputation classes estimator of the population mean is

$$\bar{y}^w_I = \frac{1}{n_r} \sum\nolimits_{\nu} \sum\nolimits_{i \in A_\nu} (1 - \delta_{\nu i})\, y^*_{m\nu i} \tag{1.4}$$

This estimator is very similar to expression (1.1), except that it includes only contributions from the "nonresponding" units of the reference sample s_r, which must be imputed within designated imputation classes from the "responding" nonrandom sample s_w. Since it has the same bias (1.3), everything said in case of nonresponse about double robustness and proper selection of imputation classes is applicable for estimator (1.4). Details of fitting a response propensity model using combined sample s are discussed by Beresovsky (2016).

1.2 Variance of point estimator under hot deck imputation within imputation classes

Following the approach of Rao and Shao (1992), the estimators of population mean with imputed data, either from random samples with missing data or from nonrandom samples are

$$\bar{y}_I = \frac{1}{n}\sum_{\nu}\left(r_\nu \bar{y}_{r\nu} + m_\nu \bar{y}^*_{m\nu}\right) = \frac{1}{n}\sum_{\nu} n_\nu \bar{y}_{I\nu} \tag{1.5a}$$

$$\bar{y}^w_I = \frac{1}{n_r}\sum_{\nu} m_\nu \bar{y}^*_{m\nu} \tag{1.5b}$$

Here $\bar{y}_{r\nu}$ and $\bar{y}^*_{m\nu}$ are the means of target variable for responding and imputed values for an imputation class ν. For nonresponse, n_ν, r_ν and $m_\nu = n_\nu - r_\nu$ are the size, and numbers of respondents and missing units for a class. In case of estimation from nonrandom samples, $m_\nu = n_\nu$ and r_ν are the numbers of units in the reference random and nonrandom samples. To proceed with variance estimation, let's assume that imputed data is either drawn from respondents using hot deck imputation or generated with parametric model (1.2b) within imputation classes. In both cases, expectation and variance of $\bar{y}^*_{m\nu}$ estimated conditionally on responding data $A_{r\nu}$ are

$$E\left(\bar{y}^*_{m\nu}|A_{r\nu}\right) = \mu_{r\nu},\ V\left(\bar{y}^*_{m\nu}|A_{r\nu}\right) = \frac{\sigma^2_{r\nu}}{m_\nu} \tag{1.6}$$

The variance of the estimators (1.5a-b), calculated using variance decomposition formula, ultimately depends on the variance of respondents $\sigma^2_{r\nu} = E\left[s^2_{r\nu}\right]$ and the numbers of respondents r_ν and nonrespondents m_ν within imputation classes

$$V\left(\bar{y}_I\right) = V\left[E\left(\bar{y}_I|A_r\right)\right] + E\left[V_r\left(\bar{y}_I|A_r\right)\right] \approx \frac{1}{n^2}\sum_{\nu} n_\nu \sigma^2_{r\nu}\left(\frac{n_\nu}{r_\nu} + \frac{m_\nu}{n_\nu}\right) \tag{1.7a}$$

$$V\left(\bar{y}^w_I\right) \approx \frac{1}{n^2_r}\sum_{\nu} m_\nu \sigma^2_{r\nu}\left(\frac{m_\nu}{r_\nu} + 1\right) \tag{1.7b}$$

Expressions in brackets in (1.7a-b) are greater than 1 and indicate variance increase due to imputation, compared with naive variances, treating all data as responders.

1.3 Adjusted jackknife variance estimation for simple random samples under hot deck imputation

Expression for the naive jackknife variance estimator of the imputation classes estimator (1.5a), ignoring the imputation of missing data, is given by

$$v_j\left(\bar{y}_I\right) = \frac{1}{n^2}\sum\nolimits_{\nu}\frac{n_\nu - 1}{n_\nu}\sum\nolimits_{j\in\nu}\left(y_{I\nu}\left(-j\right) - y_{I\nu}\right)^2 \tag{1.8}$$

where $y_{I\nu}\left(-j\right)$ are estimates resulting from removing one responding or missing unit from either $A_{r\nu}$ or $A_{m\nu}$

$$y_{I\nu}\left(-j\right) = \begin{cases} \left(n_\nu - 1\right)^{-1}\left\{n_\nu \bar{y}_{I\nu} - y_j\right\}, \left(j \in A_{r\nu}\right) \\ \left(n_\nu - 1\right)^{-1}\left\{n_\nu \bar{y}_{I\nu} - y^*_j\right\}, \left(j \in A_{m\nu}\right) \end{cases} \tag{1.9}$$

In the case of imputation from nonrandom samples, only missing units need to be deleted from the imputation classes $A_{m\nu}$ of the reference random sample (correspondingly, n_ν must be replaced with m_ν in (1.8))

$$y_{I\nu}(-j) = (m_\nu - 1)^{-1}\left\{m_\nu \bar{y}^*_{m\nu} - y^*_j\right\}, (j \in A_{m\nu}) \tag{1.10}$$

Burns (1990) and Rao and Shao (1992) have noted previuosly that this estimator underestimates the variance (1.7a-b) because it does not properly account for variability associated with imputation. Rao and Shao (1992) proposed an adjustment to compensate for this deficiency. In case of the imputation classes estimator with missing data, their adjustment results in adding an extra term to expression (1.9)

$$y^a_{I\nu}(-j) = \begin{cases} (n_\nu - 1)^{-1}\left\{n_\nu \bar{y}_{I\nu} - y_j - m_\nu(y_j - \bar{y}_{r\nu})/(r_\nu - 1)\right\}, (j \in A_{r\nu}) \\ (n_\nu - 1)^{-1}\left\{n_\nu \bar{y}_{I\nu} - y^*_j\right\}, (j \in A_{m\nu}) \end{cases} \tag{1.11}$$

Similar adjustment in case of imputation from nonrandom samples implies, that, in contrast to expression (1.10), sampled units must be deleted from both $A_{m\nu}$ and $A_{r\nu}$

$$y^\alpha_{I\nu}(-j) = \begin{cases} (m_\nu - 1)^{-1}\left\{-m_v(y_{\nu j} - \bar{y}_{r\nu})/(r_\nu - 1)\right\}, (j \in A_{r\nu}) \\ (m_\nu - 1)^{-1}\left\{m_\nu \bar{y}^*_{m\nu} - y^*_{\nu j}\right\}, (j \in A_{m\nu}) \end{cases} \tag{1.12}$$

In expression (1.12), $y^\alpha_{I\nu}(-j)$ fluctuates around 0 when $j \in A_{r\nu}$. Consequently, it must be assumed that $y_{I\nu} = 0$ for the terms of expression (1.8) corresponding to the deleted units from the donor nonrandom sample. However, the normalizing constant must still use the sizes of imputation classes m_ν of the recipient reference random sample.
It can be proved that expectations over the model (1.2a-b) of the adjusted jackknife variance estimators for both missing data and nonrandom sample problems, are equal to the variances (1.7a-b) of the imputation class estimators (1.5a-b). The proof requires equality of variances of responding and imputed data $\sigma^2_{m\nu} = \sigma^2_{r\nu}$ within imputation classes and large numbers of responding and imputed units ($r_\nu \gg 1$, $m_\nu \gg 1$).

1.4 Delete-a-group modification to Rao-Shao adjusted jackknife variance estimator for weighted samples

Application of the adjusted jackknife variance estimator is not limited to simple random samples and delete-a-unit scenarios. Rao and Shao (1992) proved consistency of such an estimator in case of stratified multistage sampling when missing individual units were imputed to first-stage units (or clusters) using weighted hot deck imputation. In this case first-stage units were removed from the strata from which they were sampled. Rao and Shao (1992) extended their analysis to hot deck imputation within designated imputation classes. Following their example, this paper will show that a delete-a-group version of their estimator may be easily applied for variance estimation of the imputation class estimators of population mean when sampled units are selected with different probabilities. The weighted imputation classes estimator of the population mean with imputed data is

$$\bar{y}_I = \frac{1}{N}\sum\nolimits_\nu \left(\sum_{i \in A_{r\nu}} w_i y_{ri} + \sum_{i \in A_{m\nu}} w_i y^*_{mi}\right) = \frac{1}{N}\sum\nolimits_\nu \left(\hat{S}_{r\nu} + \hat{S}^*_{m\nu}\right) = \frac{1}{N}\sum\nolimits_\nu y_{I\nu} \tag{1.13}$$

When all units of a reference random sample are imputed from a nonrandom sample, this expression requires only summation over the imputed units $A_{m\nu}$. Population size N is assumed known.
Allowing that both respondents and imputed units are randomly split by G groups within each imputation class, the adjusted jackknife formula for variance estimation becomes

$$v_j\left(\bar{y}_I\right)=\frac{1}{N^2}\sum_{g=1}^{G}\frac{G-1}{G}\left(\sum\nolimits_{\nu}y_{I\nu}^{a}\left(-g\right)-y_{I\nu}\right)^2 \tag{1.14}$$

According to the standard jackknife theory, $y_{I\nu}^{a}\left(-g\right)$ is an estimator of total if a group g is removed. Additional variability due to imputation is accounted by a term, proportional to the difference of means of responders within imputation classes, having the group g either included or excluded

$$y_{I\nu}^{\alpha}\left(-g\right)=\hat{S}_{r\nu(g)}+\hat{S}_{m\nu(g)}^{*}+\left(\frac{\hat{S}_{r\nu(g)}}{\hat{T}_{r\nu(g)}}-\frac{\hat{S}_{r\nu}}{\hat{T}_{r\nu}}\right)\hat{T}_{m\nu(g)} \tag{1.15}$$

The terms of this expression are obtained by deleting the group g from summations over the sampled units

$$\begin{aligned}
\hat{S}_{r\nu(g)}&=\frac{G}{G-1}\sum_{i\in A_{r\nu},i\notin g}w_{gi}y_{gi}=\frac{G}{G-1}\left(\hat{S}_{r\nu}-\hat{S}_{r\nu g}\right);\\
\hat{T}_{r\nu(g)}&=\frac{G}{G-1}\sum_{i\in A_{r\nu},i\notin g}w_{gi}=\frac{G}{G-1}\left(\hat{T}_{r\nu}-\hat{T}_{r\nu g}\right);\\
\hat{T}_{m\nu(g)}&=\frac{G}{G-1}\sum_{i\in A_{m\nu},i\notin g}w_{gi}=\frac{G}{G-1}\left(\hat{T}_{m\nu}-\hat{T}_{mg}\right);\\
\hat{S}_{m\nu(g)}^{*}&=\frac{G}{G-1}\sum_{i\in A_{m\nu},i\notin g}w_{gi}y_{gi}^{*}=\frac{G}{G-1}\left(\hat{S}_{m\nu}^{*}-\hat{S}_{m\nu g}^{*}\right);
\end{aligned} \tag{1.16}$$

It can be proved that the expectation of the variance estimator (1.14- 1.16) over the imputation class model (1.2a-b) results in the following variances of the estimates of the population mean with imputed data for random samples with unit nonresponse and nonrandom samples

$$E\left[v_j\right]=\tfrac{1}{N^2}\sum\nolimits_{\nu}\left\{\sigma_{r\nu}^2\tfrac{\hat{T}_{\nu}^2}{\hat{T}_{r\nu}}+\hat{T}_{m\nu}\sigma_{m\nu}^2\right\} \tag{1.17a}$$

$$E\left[v_j^{w}\right]=\tfrac{1}{N^2}\sum\nolimits_{\nu}\left\{\sigma_{r\nu}^2\tfrac{\hat{T}_{m\nu}^2}{\hat{T}_{r\nu}}+\hat{T}_{m\nu}\sigma_{m\nu}^2\right\} \tag{1.17b}$$

Here $\hat{T}_{\nu}=\hat{T}_{r\nu}+\hat{T}_{m\nu}=\sum_{i\in\nu}w_i$ are the weighted population sizes of imputation classes. The proof of (1.17a-b) assumes that population sizes of the deleted groups of the responding and imputed units are approximately equal to mean group sizes within imputation classes: $\hat{T}_{r\nu g}\approx\hat{T}_{r\nu}/G,\ \hat{T}_{m\nu g}\approx\hat{T}_{m\nu}/G$. The approximation holds if group sizes are not too small and if deleted groups are randomly selected, so sampling weights are uncorrelated with group assignment.
The expressions (1.17a-b) are exactly equal to variances (1.7a-b) in case of simple random sample if estimated population counts are substituted with the corresponding sample sizes: $N\to n,\hat{T}_{\nu}\to n_{\nu},\hat{T}_{r\nu}\to r_{\nu},\hat{T}_{m\nu}\to m_{\nu}$

2. Bias reduction and adjusted jackknife variance estimation in case of deterministic imputation classes

A simulations study was conducted to demonstrate application of the imputation classes estimator for inferences of population mean and median from nonrandom samples. The results presented here highlight reduction of bias depending on choice of imputation classes and application of adjusted jackknife for variance estimation.

Each unit of the simulated population of size $N = 10,000$ is characterized by random variables (U, X, Y). $U \sim \text{Unif}(0, 1)$ is uniform random variable, which is considered unobserved. Observed categorical variable $X \in (1, 2, 3, 4)$ corresponds to quartiles of the distribution of U. Target variable Y is normally distributed $Y(x) \sim N\left(\mu_x, \sigma_x^2\right)$ within strata defined by X. Means and variances of normal distributions were set to $\mu_x = (10, 12, 20, 22)$ and $\sigma_x^2 = (4, 9, 16, 25)$. The target variable mean averaged over the simulated populations is $\bar{Y} = 15.7$ and the median is $Y^m = 14.8$.

For each of the conducted $2,000$ simulations, two samples of size $n = 1,000$ were drawn from the population. One is a simple random sample s_r, representing *reference* sample with known sampling weights and unknown target variable, as discussed in Section 1.1. The second is a stratified random sample s_w with sample counts $n_x^w = (100, 400, 100, 400)$ within X strata. This sample is referred as *nonrandom*, because the correspondence between sampled and population units by strata is considered unavailable for a sampler, while the target variable Y is observed.

Imputation class estimation described in Section 1 may be applied in the following steps: (1) determine how to designate imputation classes spanning the recipient random s_r and donor nonrandom s_w samples; (2) within each class impute the target variable y from the donor to the recipient sample by either deterministic imputation, or random hot deck or parametric imputation; (3) calculate population characteristics using available weights for the reference sample and the imputed values of y; and (4) if delete-a-group adjusted jackknife is used for variance estimation, randomly designate G delete groups within imputation classes of both samples and estimate variance using formula (1.14).

Success of inference from nonrandom samples ultimately depends on assignment of the imputation classes, which is usually based on a prediction model for Y and/or a propensity model for the nonrandom sample inclusion indicator. The bias of estimates can be eliminated or reduced if either one of the models is correct (or approximately correct). This is demonstrated by comparing two estimators based on different sets of imputation classes.

The first set of imputation classes is constructed by considering only the prediction model for Y conditional on X. Since μ_x is clustered for $X = 1, 2$ and $X = 3, 4$, suppose that the prediction model was unable to differentiate between all four population strata. Instead, it came up with just two imputation classes $\nu_{y|x} = (\nu_{1y}, \nu_{2y})$, where $\nu_{1y} = X(1, 2)$ and $\nu_{2y} = X(3, 4)$. This set of imputation classes is referred as *ClassY*.

Another set of classes takes into account both prediction and propensity models, which reflects stratification of nonrandom sample n_x^w. Propensity-based imputation classes are $\nu_{p|x} = (\nu_{1p}, \nu_{2p})$, where $\nu_{1p} = X(1, 3)$ and $\nu_{2p} = X(2, 4)$. Cross-classification $\nu_{y|x} \times \nu_{p|x}$ results in imputation classes matching stratification by the observed covariate X and will be referred to as *ClassYP*.

Inferences were conducted for deterministic imputation of the predicted mean $\hat{\mu}_\nu$ for all units of an imputation class, hot deck imputation Y_ν^{hd*} and parametric random imputation Y_ν^{p*} assuming normality and using estimated means $\hat{\mu}_\nu$ and variances $\hat{\sigma}_\nu^2$ within imputation classes.

Variances of the estimates of means and medians were calculated directly over the simulations and estimated using delete-a-group adjusted jackknife variance estimation (1.14)

and (1.15) with $G = 10$ delete groups. They were compared with hypothetical variances of trivial SRS estimates from the random sample s_r, if the target variable Y would be observed. These results, as well as coverage of the finite population parameters by the estimated confidence intervals are presented in Table 1 for imputation classes *ClassY* and *ClassYP*.

Table 1: Inferences of the population mean $\widehat{\bar{Y}}$ and median $\widehat{Y}^m$ using imputation classes *ClassY* and *ClassYP*. Presented results include relative bias; ratio of variances of estimates from nonrandom and random samples s_w and s_r over the simulations, if Y would be known for s_r; ratio of the estimated variance to the variance of estimates; and coverage of the population parameter by the calculated confidence intervals.

Estimate	Mean $\widehat{\bar{Y}}$			Median $\widehat{Y}^m$		
Imputation	$\widehat{\mu}_\nu$	Y_ν^{hd*}	Y_ν^{p*}	$\widehat{\mu}_\nu$	Y_ν^{hd*}	Y_ν^{p*}
	ClassY					
Relative bias	0.036	0.036	0.036	0.086	0.038	0.034
$\text{Var}(s_w)/\text{Var}(s_r)$	1.05	1.40	1.50	224	1.33	1.16
$\widehat{\text{Var}}(s_w)/\text{Var}(s_w)$	1.09	1.13	1.05	4.72	1.20	1.12
Coverage	0.15	0.26	0.27	0.37	0.67	0.67
	ClassYP					
Relative bias	0.0	0.0	0.0	0.088	0.0	0.0
$\text{Var}(s_w)/\text{Var}(s_r)$	1.12	1.46	1.46	127	1.42	1.35
$\widehat{\text{Var}}(s_w)/\text{Var}(s_w)$	1.09	1.42	1.42	4.9	1.29	1.17
Coverage	0.93	0.93	0.93	0.37	0.92	0.90

Biases of estimates of both means and medians are ≈ 0.035 for *ClassY* and negligible for *ClassYP*. Because variances of the estimates are also small for the discussed simulations, even small bias results in incorrect inferences about the population parameters by *ClassY* estimators.

Unbiased estimates with imputation classes *ClassYP* may be attributed to both prediction and propensity models being exactly correct, see expression (1.3), which is an artifact of the setup of these simulations. Relevance of the obtained results for more general settings was demonstrated by modifying simulated final population, so even for the imputation classes *ClassYP* both models remain incorrectly specified. This is achieved by introducing explicit dependence on the unobserved variable U of both $Y(x,u) \sim N\left(\mu_x + 3u, \sigma_x^2\right)$ and non-random sample selection probability $\text{logit}(p(u)) = -5 + 5u$ within the original strata n_x^w. Nonzero bias of the estimates with imputation classes *ClassY* for the new population leads to insufficient coverage of ≈ 0.2 for the means and ≈ 0.45 for the medians. Estimates with imputation classes *ClassYP* were more robust to models misspecification, acquiring only small relative bias RB ≈ 0.01 with relatively minor coverage reduction of ≈ 0.9 compared with nominal value 0.95.

Imputation of the predicted mean $\widehat{\mu}_\nu$ within imputation classes provides for more efficient estimation of the population mean $\widehat{\bar{Y}}$ compared with hot deck Y_ν^{hd*} or parametric Y_ν^{p*} imputations, leading to variance reduction of 35-40%. However, the estimates of the population median $\widehat{Y}^m$ become unstable when an identical value of $\widehat{\mu}_\nu$ is imputed for all units of an imputation class. Hot deck and parametric imputations perform equally well for both

estimates of means and medians.
Though the sizes of the random recipient and nonrandom donor samples s_r and s_w are both equal to $n = 1,000$, variances of the imputation classes estimates are 16-50% larger than presumptive estimates from s_r, if the target variable Y would be directly observed rather than imputed. This agrees with the general expressions (1.7a-b) for the variance of estimates with imputed data. Average estimated variances of the estimators of means and medians with delete-a-group adjusted jackknife variance estimator were sufficiently close to variances of estimates over the simulations. Regarding imputation classes *ClassYP*, when the point estimator was also unbiased, the coverage of the population parameters by the estimated confidence intervals was sufficiently close to nominal.

3. Bias reduction with estimates using model-defined imputation classes

Another set of simulations was designed to investigate how effective imputation classes estimators could be for reducing bias of estimates from nonrandom samples, when imputation classes are inferred dynamically from results of prediction and propensity models.
The idea for the simulated population comes from the simulations of Kang and Schafer (2007). General finite population U_r of size $N = 15,000$ was generated with four random variables (U_1, U_2, U_3, U_4), which are identically normally distributed $N(0,1)$ and considered unobserved. For each population unit these covariates define the mean of the target variable Y and the probability p_w to belong to the subpopulation U_w, from which the nonrandom sample s_w with observed target variable can be drawn

$$Y \sim N\left(50 + 27.4U1 + 13.7U_2 + 13.7U_3 + 13.7U_4, \sigma_Y^2\right) \tag{3.1a}$$

$$\text{logit}\left(p_w\right) = -1 + U_1 - 0.5U_2 + 0.25U_3 + 0.1U_4 \tag{3.1b}$$

For these simulations the variance associated with the target variable was $\sigma_Y^2 = 100$.
As Kang and Schafer (2007) discussed, this paper presumes that the observed population covariates (X_1, X_2, X_3, X_4) are intractable nonlinear functions of the unobserved covariates. However, for these simulations we introduced additional variability associated with covariate measurement error ε_u

$$\widetilde{U}_{1,..,4} = U_{1,..,4} + \varepsilon_u,\ \varepsilon_u \sim N\left(0, \sigma_u^2\right) \tag{3.2a}$$

$$X_1 = \exp(\widetilde{U}_1/2) \tag{3.2b}$$

$$X_2 = \frac{\widetilde{U}_2}{1 + \exp(\widetilde{U}_1)} + 10 \tag{3.2c}$$

$$X_3 = \left(\widetilde{U}_1\widetilde{U}_3/25 + 0.6\right)^3 \tag{3.2d}$$

$$X_4 = \left(\widetilde{U}_2 + \widetilde{U}_4 + 20\right)^2 \tag{3.2e}$$

For each of the conducted 500 simulations, reference random sample s_r of size $n_r = 2,000$ was drawn without replacement from the general population U_r with probability p_r proportional to the measure of size, equal to the observed covariate X_{1i}. Design based estimates of population parameters with weights $w_{ri} = 1/p_{ri}$ would be unbiased, however the target variable Y is considered unavailable for the units of s_r.
A simple random sample s_w of size $n_w = 600$ is drawn from the subpopulation U_w. The target variable Y is considered available for s_w, but its distribution is expected to differ between the general population U_r and subpopulation U_w due to mutual dependence of the

probability p_w (3.1b) and the target variable Y (3.1a) on the unobserved covariates $U_{1..4}$. Consequently, estimation of parameters of the general population U_r requires model-based methods utilizing data of both random and nonrandom samples. Any model using the observed covariates $X_{1..4}$ will always be misspecified, since the true population models (3.1a-b) depend on the unobserved covariates $U_{1..4}$. The utility of the observed covariates $X_{1..4}$ for bias reduction is strongly affected by a single parameter- the random noise variance σ_u^2.

Parameters of the linear model LM.Y $E\left(Y_i^w\right) = X_{ij}^w \beta_j^w,\ j = 1, ..4$ were estimated with the nonrandom sample s_w data and used to predict the target variable $Y_i^{\text{LM.pred}} = X_{ij}^r \widehat{\beta}_j^w$ for the units of the random sample s_r. These predictions were used in the Horwitz-Tompson estimator of the population mean $\widehat{\bar{Y}}_{\text{LM.y}}^{\text{pred}} = \sum_{i \in s_r} Y_i^{\text{LM.pred}} w_{ri}/N$ utilizing available randomization weights w_{ri}.

Predicted scores $Y_i^{\text{LM.pred}}$ for both s_r and s_w were used to define imputation classes of the estimators (1.13). The target variable $Y_i^{\text{LM.imp5}}$ was imputed by unweighted hot deck from the nonrandom sample s_w to the random sample s_r within quintiles of these distributions. The regular Horwitz-Thompson estimator using the imputed values and randomization weights can be defined as $\widehat{\bar{Y}}_{\text{LM.y}}^{\text{imp5}} = \sum_{i \in s_r} Y_i^{\text{LM.imp5}} w_{ri}/N$.

Another group of estimates utilizes the propensity (3.1b) to belong to the subpopulation U_w. Estimating p_w is problematic because subpopulation U_w is unavailable. (Beresovsky (2016)) described the details of estimating propensity using random and nonrandom samples s_r and s_w instead of populations U_r and U_w. However, a propensity model utilizing the observed covariates $X_{1..4}$ is expected to be misspecified.

Propensity score $\widehat{p}_{wi}$ estimated for the units of the nonrandom sample can be used for estimating the target variable mean by the traditional propensity score adjusted (PSA) expansion estimator $\widehat{\bar{Y}}_{\text{LM.p}}^{\text{PSA}} = \sum_{i \in s_w} \left(Y_i/\widehat{p}_{wi}\right) / \sum_{i \in s_w} \left(1/\widehat{p}_{wi}\right)/n_w$. Since response propensity is expected to be relatively homogeneous within quintiles of the distribution of the predicted propensity score, they may be used to define imputation classes of the estimator $\widehat{\bar{Y}}_{\text{LM.p}}^{\text{imp5}}$. It uses the Horwitz-Thompson estimator with hot deck-imputed target variable Y similarly to the estimator $\widehat{\bar{Y}}_{\text{LM.y}}^{\text{imp5}}$.

Randomization weights w_r played dual role for the propensity model-based estimators. First, they were used by the logistic regression model to estimate $\widehat{p}_{wi}$. Second, imputation classes for the random sample s_r were defined by *weighted* quantiles of the distribution of $\widehat{p}_{wi}$, with weights equal to the *inverse* sampling weights $1/w_r$. Within these quantiles unweighted hot deck was used for imputation because units of the nonrandom sample s_w are not associated with sampling weights. Rao and Shao (1992) proposed to account for randomization weights by using weighted hot deck, under which probability of selecting a donor is proportional to its sample weight, however this leaves open the question of addressing the weights of recipients. Platek and Gray (1983) proposed to account for donor and recipient weights by using unweighted hot deck but imputing target variable modified by weights $y_i^{imp} = y_j^d \left(w_j^d/w_i^{imp}\right)$.

Though the author of this paper doesn't have a rigorous justification for accounting for the recipients' sampling weights in the implemented imputation procedure, simulation results indicated that imputation within classes defined by the weighted quantiles of the propensity score $\widehat{p}_{wi}$ were optimal for bias reduction for this simulation.

Imputation classes based on the quintiles and deciles of the predicted score $Y_i^{\text{LM.pred}}$ and weighted propensity score $\widehat{p}_{wi}$ were crossed to produce a new set of imputation classes. Some of the resulting classes had small, or even zero, number of donor units. To ensure stability of hot deck imputation, neighboring classes were collapsed until their aggregated

size exceeded 10 donor units using a procedure based on a distance between classes. The distance was defined as a dimensionless measure aggregating prediction and propensity scores averaged within each class.

Hot deck was used to impute the target variable from the nonrandom sample s_w to the random sample s_r within the combined and aggregated classes. Horwitz-Thompson estimators $\widehat{\widehat{Y}}^{\text{imp5}}_{\text{LM.yp}}$ and $\widehat{\widehat{Y}}^{\text{imp10}}_{\text{LM.yp}}$ used the imputed values and the randomization weights w_r for estimating the general population mean.

Linear models may not be the optimal choice for estimation in this case because the observed covariates $X_{1..4}$ are highly nonlinear functions of the unobserved covariates $\widetilde{U}_{1,..,4}$ (3.2a-e) associated with the true population characteristics. Therefore, this study used nonparametric recursive partitioning for classification and regression trees, implemented by the R package `rpart`. Recursive tree methods use covariates to partition sample data by nodes of relative homogeneity of a target variable. In this case, the terminal nodes of the trees resulting from modeling the target variable Y and response propensity p_w are ready-made for defining imputation classes for both the nonrandom and random samples. The Horwitz-Thompson estimator is then used to compute population mean with imputed Y and known sampling weights w_{ri}. Estimators $\widehat{\widehat{Y}}^{\text{pred}}_{\text{rpart.y}}$ and $\widehat{\widehat{Y}}^{\text{pred}}_{\text{rpart.p}}$ used predicted mean within classes defined by models of the target variable Y and response propensity p_w. Estimator $\widehat{\widehat{Y}}^{\text{imp}}_{\text{rpart.yp}}$ used hot deck imputation within classes defined by intersection of the terminal nodes produced by both prediction and propensity models.

Relative bias (RB), standard deviation (SD) and root mean square error (RMSE) of the estimates of mean of the general population U_r by different estimation methods are shown in Table 2. The variance of the covariate measurement error σ_u^2 (3.2a) indicates the degree of misspecification of a model using the observed covariates $X_{1..4}$.

Table 2: Relative bias, standard deviation and root mean square error of the estimates of mean of the general population U_r by linear and logistic models, and recursive tree R package `rpart`. Estimates are using predicted target variable Y, PSA and hot deck imputation within imputation classes. σ_u^2 is the variance of the covariate measurement error (3.2a-e).

Estimator	$\widehat{\widehat{Y}}^{\text{pred}}_{\text{LM.y}}$	$\widehat{\widehat{Y}}^{\text{imp5}}_{\text{LM.y}}$	$\widehat{\widehat{Y}}^{\text{PSA}}_{\text{LM.p}}$	$\widehat{\widehat{Y}}^{\text{imp5}}_{\text{LM.p}}$	$\widehat{\widehat{Y}}^{\text{imp5}}_{\text{LM.yp}}$	$\widehat{\widehat{Y}}^{\text{imp10}}_{\text{LM.yp}}$	$\widehat{\widehat{Y}}^{\text{pred}}_{\text{rpart.y}}$	$\widehat{\widehat{Y}}^{\text{pred}}_{\text{rpart.p}}$	$\widehat{\widehat{Y}}^{\text{imp}}_{\text{rpart.yp}}$
					$\sigma_u^2 = 0.04$				
RB	0.132	0.103	0.087	0.052	0.046	**0.015**	0.086	0.061	0.049
SD	1.31	1.63	1.33	1.99	1.70	1.60	1.66	2.51	2.21
RMSE	6.72	5.36	4.50	3.24	2.84	**2.51**	4.59	3.93	3.27
					$\sigma_u^2 = 0.64$				
RB	0.204	0.127	0.172	0.101	0.051	**0.029**	0.188	0.146	0.138
SD	1.46	1.92	1.43	2.61	2.28	2.78	1.76	2.02	1.87
RMSE	10.28	6.61	8.66	5.66	3.41	**3.13**	9.56	7.56	7.15
					$\sigma_u^2 = 4$				
RB	0.261	0.172	0.245	0.190	0.061	**0.037**	0.419	0.233	0.228
SD	1.56	2.39	1.54	4.83	4.19	4.45	2.65	1.88	2.31
RMSE	13.15	8.92	12.34	10.65	5.16	**4.81**	21.08	11.80	11.62

Analysis of estimates by different methods suggests the following conclusions.

- Estimators $\widehat{\widehat{Y}}^{\text{imp10}}_{\text{LM.yp}}$ and $\widehat{\widehat{Y}}^{\text{imp5}}_{\text{LM.yp}}$, using hot deck imputation within imputation classes

produced by *both* prediction and propensity linear models, have much smaller bias and RMSE compared with other estimates. Their robustness is particularly impressive for strongly misspecified models (large σ_u^2).

- Conventional estimator $\widehat{\widehat{Y}}_{\text{LM.y}}^{\text{pred}}$, using predicted $Y_i^{\text{LM.pred}}$, has a larger bias and RMSE, but smaller SD than $\widehat{\widehat{Y}}_{\text{LM.y}}^{\text{imp5}}$, which uses hot deck imputation within quintiles defined by the same prediction model. The same could be observed for the conventional PSA estimator $\widehat{\widehat{Y}}_{\text{LM.p}}^{\text{PSA}}$ and $\widehat{\widehat{Y}}_{\text{LM.p}}^{\text{imp5}}$, using hot deck imputation within quintiles based on the same propensity model. It seems that estimates using hot deck imputation within model-defined classes are more robust to model misspecification than estimates based on model predictions. This added robustness comes at the expense of larger variance.

- Hot deck estimator $\widehat{\widehat{Y}}_{\text{LM.yp}}^{\text{imp5}}$ within quintiles of both models is more robust (smaller bias and RMSE) than hot deck estimators $\widehat{\widehat{Y}}_{\text{LM.p}}^{\text{imp5}}$ and $\widehat{\widehat{Y}}_{\text{LM.y}}^{\text{imp5}}$ within quintiles of either one of the models. This leads to the conclusion that combining imputation classes improves robustness of estimates, but may introduce larger variance.

- Estimators based on recursive tree models outperform those based on linear models when observed covariates are strongly correlated with studied population characteristic and response propensity (small σ_u^2). The opposite is true for weak observed covariates, when estimator $\widehat{\widehat{Y}}_{\text{rpart.y}}^{\text{pred}}$ becomes completely unstable.

4. Two is better than one: - Robustness of estimators using hot deck within combined imputation classes defined by linear models

The imputation classes estimator provides a simple way to account simultaneously for predictive and propensity models by using imputation classes defined by both models. This can reduce bias of the imputation classes estimator (1.3) by making residuals of both models more homogeneous and less correlated within the combined classes.
Percent bias reduction by a given model-based estimator $\widehat{Y}_{\text{mod}}$ measures how much of the original bias of the direct estimator from a nonrandom sample $\widehat{Y}_{\text{dir}}$ is eliminated by using model adjustment

$$BR\left(\widehat{Y}_{\text{mod}}\right) = \frac{\widehat{Y}_{\text{dir}} - \widehat{Y}_{\text{mod}}}{\widehat{Y}_{\text{dir}} - Y_{\text{pop}}}\ 100\% \tag{4.1}$$

Figure 1 shows dependence of BR on the covariate measurement error (3.2a) for two conventional estimators of the mean and two proposed imputation classes estimators utilizing quintiles and deciles of both prediction and propensity linear models. The covariate error quantifies a degree of a model misspecification. Imputation classes estimators utilizing both models are more robust than conventional estimators to model misspecification based on any one model. Their advantage becomes particularly clear for strongly misspecified models. Table 2 shows that enhanced robustness comes with a price of larger variability of estimates, but this tradeoff may be worthwhile, if large nonrandom samples will be readily available from web surveys or administrative data and bias of estimates will become of greater concern.
Demonstrated remarkable robustness of imputation classes estimators motivates more research of using double robust imputation estimators from nonrandom samples data. Two issues require further exploration. First, clarity is needed for accounting for donors and

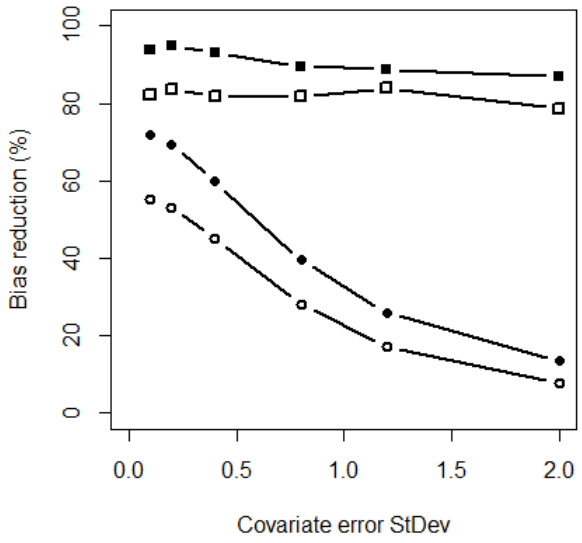

Figure 1: Percent bias reduction (BR) as a function of the standard deviation of the covariates measurement error $\sqrt{\sigma_u^2}$ for selected conventional and double robust estimators: $\circ - \widehat{\bar{Y}}^{\text{pred}}_{\text{LM.y}}$,$\bullet - \widehat{\bar{Y}}^{\text{PSA}}_{\text{LM.p}}$, $\square - \widehat{\bar{Y}}^{\text{imp5}}_{\text{LM.yp}}$, $\blacksquare - \widehat{\bar{Y}}^{\text{imp10}}_{\text{LM.yp}}$.

recipients weights in hot deck imputation between weighted samples. Results of these simulations show that using imputation classes defined by properly weighted quantiles on a *recipient* side made a big difference for bias reduction comparing to using unweighted quantiles. Weighted hot deck discussed by Rao and Shao (1992) does not take a recipient's weight into consideration. Platek and Gray (1983) accounted for these weights by imputing modified values proportional to a donor weight and inverse of a recipient weight. This may be acceptable for imputation of continuous variables, but not for binary variables.

The second issue, is how to account for correlation between a target variable and propensity to belong to a nonrandom sample within unified optimization approach, instead of modeling them separately and then crossing both models' quantiles? This method may be more efficient compared with combining imputation classes without losing much of its robustness.

References

Beresovsky, V. (2016). Using official surveys to reduce bias of estimates from nonrandom samples collected by web surveys. *Proceedings of the American Statistical Association, Survey Research Methods Section*, pages 1804–1819.

Bethlehem, J. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4(3):251–260.

Brick, J. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics*, 29(3):329–353.

Burns, R. (1990). Multiple and replicate item imputation in a complex sample survey. *Proc. Sixth Annual Res. Conf., Survey Research Methods Section*, pages 655–65.

Chmura, L., Rivers, D., Bailey, D., Pierce, C., and Bell, S. (2013). Modeling a probability sample? An evaluation of sample matching for an internet measurement panel. In *Presented at AAPOR 2013 Conference*.

Dever, J. A., Rafferty, A., and Valliant, R. (2008). Internet surveys: Can statistical adjustments eliminate coverage bias? *Survey Research Methods*, 2(2):47–62.

Deville, J. and Sarndal, C. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87:376–382.

DiSogra, C., Cobb, C., Chan, E., and Dennis, J. (2011). Calibrating non-probability internet samples with probability samples using early adopter characteristics. In *Proceedings of the American Statistical Association, Survey Research Methods Section*, pages 4501–4515.

Haziza, D. and Beaumont, J.-F. (2007). On the construction of imputation classes in surveys. *International Statistical Review*, 75(1):25–43.

Haziza, D. and Lesage, E. (2016). A discussion of weighting procedures for unit nonresponse. *Journal of Official Statistics*, 32(1):129–145.

Kang, J. and Schafer, J. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):523–539.

Kim, J. and Kim, J. (2007). Nonresponse weighting adjustment using estimated response probability. *The Canadian Journal of Statistics*, 35(4):501–514.

Kott, P. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Proceedings of the American Statistical Association, Survey Research Methods Section*, 32(2):133–142.

Leacy, F. and Stuart, E. (2014). On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: a simulation study. *Statistics in Medicine*, 33:3488–3508.

Lee, S. and Valliant, R. (2009). stimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods & Research*, 37(3):319–43.

Little, R. and Rubin, D. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

Platek, R. and Gray, G. (1983). *Imputation methodology: total survey error*. Academic Press.

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

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rubin, D. B. and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81:366–74.

Sarndal, C. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, 33:99–119.

Sarndal, C. and Lundstrom, S. (2005). *Statistical Analysis with Missing Data*. John Wiley & Sons, New York.

Valliant, R. and Dever, J. A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, 40(1):105–137.