A Modeling Approach to Compensate for Nonresponse and Selection Bias in Surveys?

Tien-Huan Lin[1], Ismael Flores Cervantes[1]
[1]Westat, 1600 Research Blvd, Rockville, MD 20850

Abstract
In surveys, errors such as selection bias, nonresponse, or noncoverage are all potential causes of biased estimates. This paper focuses on selection bias, which could be self-inflicted due to erroneous sample selection or could occur as not missing at random (NMAR) nonresponse. As examples, tobacco use surveys may be subject to selection bias since young males who are more prone to tobacco use are also less likely to participate; and surveys of domestic violence with an unbalanced sample of older females could induce biased results since the prevalence is highly correlated with age and gender. The common approach of mitigating bias using weighting adjustments justified by models for response propensity may increase the variance of weighted estimates. This paper examines empirically the bias and variance of estimators incorporating weighting adjustments that take into account the correlation between survey outcomes and response propensity using gradient boosting, a popular statistical learning method. Simulations are used to study the behavior of the estimators in three settings: 1) missing at random; 2) NMAR with partial model specified, and 3) NMAR with selection bias and partial model specified.

Key Words: selection bias, not missing at random, statistical learning, gradient boosting method

1. Introduction

It is common practice in survey research to attempt to mitigate bias due to unit nonresponse by making weighting adjustments to the base weights that account for the sampled units' unequal selection probabilities. This approach, however, is built upon the assumption that an individual's probability of responding does not depend on the unobserved data. This assumption is not always met in practice. For example, in a tobacco use survey, young males who are more prone to tobacco use may also be less likely to participate in the study. Following the terminology proposed by Rubin (1976) and Little and Rubin (2002), there are three assumptions of nonresponse: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). The simplest and strongest assumption is MCAR, where it is assumed that nonresponse is unrelated to any variables in the data. MCAR implies that respondents can be viewed as simple random sampling of the original sample; it is the most restrictive assumption and is rarely satisfied in practice. The more common assumption, in which most of the survey statistics are built upon, is the MAR. MAR assumes that if covariates are observed for all sampled units, respondents and nonrespondents, and if missingness occurs only in the outcome variable, the probability to respond depends only on the covariates. The final assumption, NMAR, has been gaining attention in recent years due to the decline in response rates. Under this assumption, the probability to respond depends on unobserved data after conditioning on observed data. Since the data necessary for adjusting for nonresponse is not available, a natural concern for estimates under this assumption would be bias.

Morral, Gore, and Schell (2014), inspired by Little and Vartivarian (2005), developed a novel nonresponse approach for a specific survey featuring statistical learning tools to compensate for nonresponse bias under a NMAR setting. Their approach consisted of two steps. The first step used data from respondents to create predictions of key outcome variables given all of the data available for respondents and nonrespondents. The second step took advantage of the predicted key outcome variables in utilizing them in response propensity modeling to create final survey weights. Morral, Gore, and Schell (2016) reported evidence in favor of their nonresponse approach as reducing bias with an acceptable increase in variance.

Fay and Riddles (2017) studied the feasibility of applying this two-step approach to broader use through simulation. Although their study ended without a general answer to the question of bias reduction with acceptable increases in variance, they reported a somewhat promising result of the two-step approach. Lin and Flores Cervantes (2018) compared the two-step approach to alternative methods that incorporated survey outcomes in nonresponse weighting adjustments to study the effect on variance using data from a survey with a complex sample design; however, the results were inconclusive.

Since the Lin and Flores Cervantes (2018) study used survey data, the true population parameter is unknown and, therefore, the bias of the estimators could not be evaluated. The purpose of this paper is to address this shortcoming with a simulation study. The survey outcome and response propensity assumptions in the simulation are entirely synthetic, allowing for the measurement of bias and variance of the estimators from different approaches.

The rest of this paper is organized as follows. In Section 2, we describe our simulation setup. Section 3 gives a more detailed illustration of the two-step or modeling approach and describes the estimates calculated in the simulations. Section 4 presents the results. We conclude in Section 5 with a discussion of our results.

2. Simulation Study Population and Design

Data from the 2012 National Health Interview Survey (NHIS) were treated as the population for the simulation study, and repeated samples were drawn with three response-generating mechanisms for this population. Table 1 lists the variables used from the 2012 NHIS to generate a synthetic survey outcome as well as the response mechanisms for different nonresponse assumptions.

Table 1: Variables Used from NHIS for Outcome and Response Models

Variable name	*Variable description*
EDUC1	Education
INCGRP3	Income
OCCUPN2	Work class
FM_KIDS	#kids in HH
AGE_P	Age
RACERPI2	Race/Ethnicity
ALCSTAT	Alcohol drinking status
CANEV	Ever told by a doctor you had cancer

Table 1: Variables Used from NHIS for Outcome and Response Models (continued)

CNKIND8	Esophagus cancer
CNKIND17	Mouth, tongue, lip cancer
CNKIND14	Lung cancer
CNKIND27	Throat cancer
CNKIND13	Liver cancer
CNKIND10	Kidney cancer
HRTEV	Ever had heart condition/disease
CHDEV	Ever had coronary heart disease
MIEV	Ever had heart attack
STREV	Ever had a stroke
AASMEV	Ever had asthma
RESPALYR	Had respiratory allergies, past 12 months
AOVRWTYR	Had problems being overweight, past 12 months
COPDEV	Ever had COPD
SEX	Sex
FM_TYPE	Family type
FWKLIMYN	Work limitation due to health problem (family member)
FSRUNOUT	Worried food would run out before got money to buy more
PLAWKNOW	Unable to work now due to health problem (individual)
LA1AR	Any limitation – all conditions
REGION	Region

For this study, only households with number of members less or equal to 12 were retained in the population. Households with 5 or more members were repeated to match the number of large households in the population. Missing values for variables listed in Table 1 were recoded to 0 and no records were dropped. After these changes, the population used for simulation consisted of 57,356 adults ages 18 and older.

The survey outcome variable used in the simulation was a synthetic variable created as a function of several variables listed in Table 1. This model allowed us to evaluate models fitted to predict survey outcome. The survey outcome variable (Y) was generated based on the following equation:

$$
\begin{aligned}
Y = \; & (2.2 - 0.1 \cdot (educ)) \cdot (income < \$35k) \cdot (work\ class) \\
& -0.04 \cdot (\#kids\ in\ household) \\
& +0.0045 \cdot \left(\frac{age \geq 46}{10}\right) + 0.06 \cdot (age \leq 30) \\
& +0.3 \cdot (black\ male\ age \leq 30) + 0.065 \cdot (male) - 0.12 \cdot (female) \\
& +0.03 \cdot (white) + 0.045 \cdot (AIAN\ female) \\
& +0.065 \cdot \big(0.1 \cdot (alcohol\ drinking\ status - 2.5)\big) \\
& +0.005 \cdot (has\ cancer) + 0.03 \cdot (has\ coronary\ heart\ disease) + 0.03 \cdot (ever\ had\ heart\ attack) \\
& +0.03 \cdot (has\ a\ heart\ condition, disease) + 0.03 \cdot (ever\ had\ a\ stroke) + 0.03 \cdot (has\ COPD) \\
& +0.08 \cdot (lung\ cancer) + 0.07 \cdot (esophagus\ cancer) + 0.06 \cdot (mouth, tongue, lip\ cancer) \\
& +0.06 \cdot (throat\ cancer) + 0.05 \cdot (liver\ cancer) + 0.04 \cdot (kidney\ cancer) \\
& +0.045 \cdot (has\ asthma) + 0.015 \cdot (had\ respiratory\ allergy) + 0.01 \cdot (had\ problems\ being\ overweight)
\end{aligned}
$$

In the simulation, repeated one-stage samples of households were drawn from this fixed population depending on three scenarios depending on the nonresponse assumption. The samples were drawn using a Bernoulli sample design with a probability of selection of 0.7 for MCAR. For MAR and NMAR, the sample was drawn using Poisson sample design, where the selection probability was proportional to the measure of size, which was the household size with a differential error by region.

Within each sampled household, all adults were selected. Nonresponse was introduced at the person level, with three different response-generating mechanisms based on the three nonresponse assumption. For MCAR, the response propensity r_{mcar} is independent from the covariates in the population, and it can be treated as an additional stage where respondents are a simple random sample of all persons in sampled households. In contrast, the response propensities for the r_{mar} model were generated based on the following equation:

$$
\begin{aligned}
r_{mar} = {} & 0.432 - \mathbf{0.42} \cdot (educ < HS) \cdot (income < \$35k) \\
& -\mathbf{0.1} \cdot (\#kids\ in\ household) \cdot (female) \\
& -0.0005 \cdot (age) + 0.3 \cdot (age/50) \\
& -0.15 \cdot (BLK\ male\ age \leq 30) - 0.1 \cdot (male \leq 30) + 0.1 \cdot (female) \\
& +0.08 \cdot (family\ type - 3) \\
& +0.05 \cdot (ever\ had\ cancer) \\
& +1.5 \cdot (lung\ cancer) + 1.45 \cdot (esophagus\ cancer) + 1.4 \cdot (mouth, tongue, lip\ cancer) \\
& +1.35 \cdot (throat\ cancer) + 1.3 \cdot (liver\ cancer) + 1.25 \cdot (kidney\ cancer) \\
& -0.1 \cdot (family\ member\ has\ work\ limitations\ due\ to\ health\ issues) \\
& +0.01 \cdot (worried\ food\ would\ run\ out\ before\ got\ money\ to\ buy\ more) \\
& +0.065 \cdot (unable\ to\ work\ now\ due\ to\ health\ issues) \\
& -0.02 \cdot (any\ limitations, all\ member, all\ conditions) \\
& -0.05 \cdot (midwest) - 0.045 \cdot (south)
\end{aligned}
$$

The NMAR response propensities r_{nmar} were essentially the same as the MAR propensities r_{mar} with two differences: 1) a selection bias was added to the coefficient of two important predictors, namely $(educ < HS) \cdot (income > \$35k)$ (i.e., coefficient was set to 0.84) and $(\#kids\ in\ household) \cdot (female)$ (i.e., coefficient was set to 0.6). This generated two selection bias patterns: low-income people with less education tended to have high estimates and were less likely to respond; and female with kids tended to have lower estimates and were less likely to response. 2) several variables used to generate r_{mar} were removed from the data used to simulate r_{nmar}; that is, these variables were not available for fitting the response models creating a scenario of unobserved data. The variables removed were education, income, number of kids in household, and family type.

For each response scenario, the sample selection was repeated 10,000 times and empirical estimates of bias and mean squared errors were used to evaluate the estimators. Note that unlike most simulation studies, the survey outcome (Y) was not built directly into the nonresponse mechanisms but was linked through covariates that appeared in both models.

3. Modeling Approach Using Statistical Learning Tools

Statistical learning theory is a framework for machine learning drawing from the fields of statistics and functional analysis (Hastie, Tibshirani, and Friedman, 2009). It deals with the problem of finding a predictive function based on data. Under this framework, Morral, Gore, and Schell (2014) and Fay and Riddles (2017) developed and implemented a two-step modeling approach with the goal to balance bias and variance, on a specific study

that featured a stratified simple random sample design. In the first step of this approach, separate models are fitted for each survey outcome using all data available for the respondents. The fitted models are then applied to both respondents and nonrespondents to generate predicted values for the survey outcomes of interest. In the second step, a response propensity model is fitted using the gradient boosting method using the predicted survey outcome(s) from the first step as independent variables. The fitted response propensity model is then used to compute a nonresponse adjustment factors to adjust the survey base weights. The statistical learning algorithm used by Morral, Gore, and Schell (2014) and Fay and Riddles (2017) to fit and predict for survey outcome(s) is the gradient boosting machine (Hastie, Tibshirani, and Friedman, 2009). Gradient boosting is a machine learning technique for regression and classification problems that produces a prediction model in the form of an ensemble of weak prediction models based on classification trees. The algorithm creates the trees sequentially, each one using information from the previously grown trees, allowing more and different shaped trees to slowly attach remaining residuals (James, Witten, Hastie, and Tibshirani, 2013).

In this simulation study, we follow the two-step modeling approach by fitting a model to a synthetic outcome Y (or y for the samples) and to predict for $\hat{y}$ using the gradient boosting method. We implement this with the R package `xgboost` (Chen, et al., 2018). The second step of the two-step approach is implemented in two different manners. The first method models our artificial response propensities (i.e., r_{mcar}, r_{mar}, $r_{nmar)}$ with a standard weighting class method (Lessler and Kalsbeek, 1992). Unlike the approach used by Morral, Gore, and Schell (2014) and Fay and Riddles (2017) that only uses $\hat{y}$(s) to model response propensity, the simulation uses all data available for respondents and nonrespondents in addition to the $\hat{y}$(s).[1] This is implemented with the R package `rpms` (Toth, 2018). For the remainder of this paper, we will refer to this procedure as the modeling approach with classification tree method, or "xgb + rpms". The second option predicts the response propensities with the gradient boosting method using all data available for respondents and nonrespondents in addition to the $\hat{y}$(s). This is, again, implemented with the R package `xgboost`. This procedure will be referred to as the repeated modeling method, or "xgb + xgb" for the remainder of this paper.

In addition to the two modeling approaches described above, the traditional estimator based on the weighting class method that adjusts for nonresponse by response propensity alone using all data available to respondents and nonrespondents is included in the simulation. The weighting class estimator is implemented with the R package `rpms` and is referred to as such. This calculation allows the comparison of the traditional (weighting class) vs. innovative (modeling) approaches. Finally, two baseline estimates are computed in the simulation: the Horvitz-Thompson (HT) estimate calculated using the full sample and the base-weighted estimate. The HT is useful as a benchmark since it is unbiased by construction. The second uses only respondents and assumes that respondents are a simple random subsample of the original sample, and the estimator is adjusted by the overall response propensity, which cancels out when computing the mean. This "naïve benchmark" is expected to yield the most biased estimate. The two baseline estimates along with estimates from the three adjustment methods (i.e., rpms, xgb + rpms, xgb + xgb) were

[1] The reason for including all the variables in the estimation process is that the gradient boosting is designed to handle many variables in contrast to a single variable as in Morral, Gore, and Schell (2014) and Fay and Riddles (2017).

computed in each of the 10,000 simulation runs, for the three different nonresponse assumptions, amounting to a total of 50,000 estimates for each nonresponse mechanism.

4. Results

In this section, we compare the estimates from each adjustment method for each nonresponse assumptions in terms of bias and mean square error.

Ultimately, any nonresponse adjustment is a balancing act between bias and variance. The evaluation tools used in this section are relative bias and relative root mean squared error, with relative bias defined as

$$Relative\ Bias: RB(\hat{Y}_E)\% = 100 \times B^{-1} \sum_{b=1}^{B} \frac{\hat{Y}_{E,b} - Y}{Y},$$

and relative root mean squared error (*RRMSE*) defined as

$$Relative\ Root\ Mean\ Squared\ Error: RRMSE = \sqrt{\frac{MSE(\hat{Y}_E)}{Y^2}},$$

where $MSE(\hat{Y}_E) = \frac{\sum_{b=1}^{B}(\hat{Y}_{E,b} - Y)^2}{B}$.

Table 2 shows the relative bias and *RRMSE* for the baseline estimates and for the nonresponse adjusted estimates for each of the three nonresponse mechanism assumptions. As expected, the empirical relative bias of the HT estimates is very small (less than ±0.4%). The *RRMSE* of the HT estimates represents the variance, and we will use it as the baseline of comparison for the nonresponse adjusted estimates.

The empirical relative bias for the base-weighted estimate and the three nonresponse adjusted estimates under MCAR are less than ±0.2% as expected since these estimates, like the HT estimate, should be unbiased. The *RRMSE* values are slightly elevated from that of the HT estimate (i.e., 9.96 vs. 12.15), suggesting that nonresponse adjustments would slightly increase variance. However, the increase is minimal.

For the MAR assumption, there is a modest increase in empirical relative bias ranging from 3.6 percent to 4.2 percent. The increase in empirical variance from the nonresponse adjustment is also marginal, averaging around 3 percent (i.e., the average of the *RRMSE* of the four estimates 13.20%-13.34% subtracted by the *RRMSE* of the HT estimate 9.94%). Overall, the simulation results under the MAR assumption are as expected since these estimators are unbiased.

The results for NMAR tell a different story. The empirical relative bias of the estimators is around 20 percent. This indicates that the three nonresponse adjustment methods marginally reduce the nonresponse bias, but none was successful in reducing the empirical bias to a level similar to that of the HT estimator. Furthermore, the differences in empirical RMSE of the estimators in the simulation results suggest that the adjustment methods have a differential effect on the bias and variance reduction. The largest reduction is for the traditional weighting class method, which removes close to 2 percent of bias (i.e., 21.7%-19.9%), followed by the modeling approach with classification tree removing close to 1 percent of bias (i.e., 21.7%-20.9%). The repeated modeling method is the least effective,

removing a mere 0.3 percent of bias (i.e., 21.7%-21.4%). Similar observations can be drawn from the *RRMSE* with the traditional weighting class method showing the lowest value (i.e., 25.75%) and the repeated modeling method showing the highest value (i.e., 26.62%).

Table 2: Relative Bias and Relative Root Mean Square Error

	Nonresponse assumption					
	MCAR		*MAR*		*NMAR*	
Estimates	*Relative bias (%)*	*Relative RMSE (%)*	*Relative bias (%)*	*Relative RMSE (%)*	*Relative bias (%)*	*Relative RMSE (%)*
Horvitz-Thompson	0.0	9.96	-0.4	9.94	-0.1	9.88
baseweighted	0.0	12.15	-4.0	13.20	21.7	26.82
rpms	0.1	12.15	-4.2	13.34	19.9	25.75
xgb + rpms	0.1	12.15	-4.0	13.24	20.9	26.39
xgb +xgb	-0.2	12.14	-3.6	13.29	21.4	26.62

5. Discussion

We started this research with the goal of extending our 2018 paper with a simulation experiment to study the effectiveness of bias reduction using a modeling approach for nonresponse adjustment. We also had hopes of confirming the results from the earlier works of Morral, Gore, and Schell (2016) and Fay and Riddles (2017), in which the modeling approach produced optimistic results in bias reduction. The simulation allowed us to detail the mitigation on nonresponse bias in different nonresponse assumptions, but the results lead us to different conclusions for the usefulness of a modeling approach under a not missing at random assumption. As shown in the previous section, all three nonresponse adjustment methods implemented in the simulation marginally reduced nonresponse bias, but none of methods could significantly reduce or remove the empirical nonresponse bias. Moreover, the method that yielded the most bias reduction and produced the lowest *RRMSE* was the traditional weighting class method, suggesting no real benefit in implementing a more sophisticated modeling approach when the response pattern is not missing at random. The result of the simulations are disappointing but not surprising since the existing literature reports that bias can only be removed when conditioning on correlated covariates. In the NMAR scenario, these variables are not available, and, therefore, bias cannot be removed.

Despite the disappointing results for the not missing at random case, the research still provides some insights for the missing completely at random and missing at random assumptions. The simulation results for the missing completely at random assumption is in accordance with the literature, showing that estimates under this assumption are unbiased regardless of adjustment method. It also demonstrates that although unnecessary, complicated nonresponse adjustment methods do not have a strong negative impact on variance. Under the missing at random assumption, a conclusion drawn from the results is that while an elaborated nonresponse adjustment method (i.e., modeling approach) does not induce negative effect on estimates, it presents limited benefits over a traditional weighting class approach and bears little practical value.

References

Chen, T., T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, M. Li, J. Xie, M. Lin, Y. Geng, and Y. Li. 2018. *XGBoost: Extreme gradient boosting, R package version 0.71.2*. Retrieved from https://cran.r-project.org/web/packages/xgboost/index.html.

Fay, R. E., and M. K. Riddles. 2017. One-versus two-step approaches to survey nonresponse adjustments. In *JSM Proceedings.* Baltimore, MD: American Statistical Association. 953-964.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The elements of statistical learning.* New York: Springer.

James, G., D. Witten, T. Hastie, and R, Tibshirani. 2013. *An introduction to statistical learning.* New York: Springer.

Lessler, J. T., and W. D. Kalsbeek. 1992. *Nonsampling errors in surveys* (1st Ed.). New York: John Wiley and Sons.

Lin, T.-H., and I. Flores Cervantes. 2018. Evaluating nonresponse weighting adjustments for the population-based HIV impact assessments surveys: On incorporating survey outcomes. *JSM Proceedings.* Vancouver, BC: American Statistical Society. 2515-2526

Little, R., and D. Rubin, D. 2002. *Statistical Analysis with Missing Data.* New York: Wiley.

Little, R. J., and S. Vartivarian. 2005. Does weighting for nonresponse increase the variance of survey means? *Statistics Canada, 31*(2). 161-168.

Morral, A. R., K. L. Gore, and T. E. Schell. 2014. *Sexual assault and sexual harassment in the U.S. Military: Volume 1. Design of the 2014 RAND Military Workplace Study.* Santa Monica, California: RAND Corporation. Retrieved from www.rand.org/t/RR870z1

Morral, A., K. Gore, and T. Schell. 2016. *Sexual assault and sexual harassment in the U.S. Military: Volume 4. Investigations of potential bias in estimates from the 2014 RAND Military Workplace Study.* Santa Monica, CA: RAND Corporation. Retrieved from https://www.rand.org/pubs/research_reports/RR870z6.html

Rubin, D. 1976. Inference and missing data. *Biometrika, 63.* 581-590.

Toth, D. 2018. *rpms: An R package for modeling survey data with regression trees.* Retrieved from R package version 0.4.0: https://cran.r-project.org/web/packages/rpms/vignettes/rpms_2018_01_22.pdf